Cognitive Endurance as Human Capital*

Christina Brown† Supreet Kaur Geeta Kingdon Heather Schofield

July 21, 2021

(Click here for latest version)

Abstract

We examine the possibility that schooling may build human capital not only by teaching academic content, but by expanding the mind's capacity for cognition itself. We hypothesize that one feature of formal schooling, engaging in effortful thinking for sustained periods, could increase cognitive endurance—the ability to maintain focus over time. To motivate this idea, we document that globally and in the US, the poor exhibit worse cognitive endurance than the rich across a variety of field behaviors; they also attend schools that are less likely to require them to engage in concentration. We test our hypothesis using a field experiment with 1,650 low-income Indian primary school students. We assign students to engage in cognitive activity for sustained periods during the school day, using either math content (mimicking good schooling) or non-academic content (providing a pure test of our mechanism). Each approach markedly improves cognitive endurance across disparate domains: academics, listening, IQ tests, and traditional psychology measures. Moreover, each treatment increases students' regular school performance in Hindi, English, and math by 0.08 standard deviations. This indicates that simply spending time in effortful thinking—without learning any subject content—substantively improves traditional measures of human capital. Our findings support a broader view of how schooling shapes general human capital, and suggest that worse environments may disadvantage poor children by hampering the development of core cognitive capacity.

*We thank David Autor, Stefano DellaVigna, Ernst Fehr, Xavier Gine, Caroline Hoxby, Lawrence Katz, Patrick Kline, David Laibson, Imran Rasul, and Sendhil Mullainathan for helpful comments and discussions. We gratefully acknowledge generous funding and support from USAID DIV, the Global Engagement Fund at the University of Pennsylvania, and The Weiss Family Program Fund for Research in Development Economics. We thank Pixatel for use of the imagine Math software and the Institute for Financial Management and Research (IFMR) for operational support. We thank Rolly Kapoor, Lubna Anantakrishnan, Simranjeet Dhir, Deepika Ghosh, Erik Hausen and Adrien Pawlik at the Behavioral Development Lab and Isadora Frankenthal, Medha Aurora, Joaquin Fuenzalida, Jed Silver, Letian Yin, Yige Wang for exceptional research assistance. All remaining errors are our own. We received IRB approval from the University of California, Berkeley and IFMR in India.

†Brown: University of California, Berkeley (christinabrown@berkeley.edu); Kaur: University of California, Berkeley and NBER (supreet@berkeley.edu); Kingdon: University College London (g.kingdon@ucl.ac.uk); Schofield: University of Pennsylvania (hschof@wharton.upenn.edu)

1 Introduction

A large body of work documents far-reaching, persistent benefits of schooling (CITES). While it's clear that schooling affects cognitive abilities, the pathways through which it does so are less well understood. Most studies focus on schooling's role in teaching academic content and skills, such as numeracy and literacy. However, education researchers have long hypothesized that its role in shaping cognition could go beyond learning content (e.g. Dewey 1934): schooling may alter the mind itself by expanding underlying cognitive capacity. This possibility would suggest a more expansive view of how education shapes general human capital, but evidence is limited. In this paper, we empirically examine this possibility using a field experiment with elementary school students.

Schooling is a bundle with many important facets. We isolate one specific facet: formal education engages students in effortful thinking for continuous stretches of time. From doing in-class exercises to reading a textbook, the act of learning often involves periods of sustained concentration. We investigate the potential for this facet to improve a core mental capacity: "cognitive endurance"—the ability to sustain focus in a task over time—also referred to as "sustained attention".[1] We investigate effects on this capacity both because of its importance for human cognition, and because of a large literature in psychology hypothesizing its potential malleability through "training" (e.g. Chun et al., 2011).

To motivate the empirical relevance of cognitive endurance, we begin with an illustrative example. Because concentrating continuously on a task will begin to create mental fatigue, this generates a key implication: performance will decline over time during any effortful cognitive task. In Figure 6, we examine performance declines in TIMSS, a math and science test administered in over 50 countries to fourth graders during the school day. Question order is randomized and students are given ample time, so that declines are not driven by changes in question difficulty or test completion.[2]

Figure 6 illustrates two stark patterns. First, consistent with cognitive fatigue, when the same question appears later in the test rather than earlier, students are considerably more likely to get it wrong. For example, among low socioeconomic status students, the rate of performance decline is 12% in the global sample and 6% in the US sample. Such performance declines have been documented in other tests such as PISA, as well as myriad field settings where the stakes for maintaining focus are high—including among paramedics, data entry workers, and judges while they are at work, or among voters at the ballot box (Brachet et al., 2012; Kaur et al., 2015; Danziger et al., 2011; Warm and Dember, 1998; Augenblick and Nicholson, 2015). In our field experiment, we provide more

[1] In cognitive psychology, sustained attention is considered one of the three key elements of attention (Chun et al., 2011). While sustained attention has a specific definition in psychology, in this paper, we use this term interchangeably with the more general phrases "cognitive endurance" and "capacity to focus". This is both because our goal is not to parse specific attentional mechanisms, and because they have similar implications for the field behaviors we examine.

[2] The test is explicitly designed to allow for sufficient time to complete it. Only 3.2% of questions are skipped and 4.5% of questions are not reached (Foy et al., 2011). Moreover, these patterns are similar if we restrict the sample on completed questions only. Note that we view this as motivational evidence. We provide a more carefully controlled test of decline effects in our experiment.

carefully controlled measurement of cognitive fatigue effects. While only suggestive, the ubiquity of such declines across domains supports the premise that the capacity for sustained focus matters for economic outcomes.

Second, there is systematic heterogeneity in cognitive fatigue effects: across all four panels, performance declines are more severe for more disadvantaged students. For example, in the US, Black and Hispanic students show XX% more decline than White students; this difference in decline accounts for 10% of the total White/Non-white test score gap (Panels A and B). We see similar patterns by wealth globally (Panels C and D). Below, we document similar systematic heterogeneity by wealth among adults in behaviors outside of schooling.

One potential interpretation of this heterogeneity is that cognitive endurance could be malleable—as has been hypothesized in the attentional "training" literature in psychology. Time use data from TIMSS teacher surveys suggest that schooling may be relevant for such training: both globally and in the US, the schools attended by richer students allot more time to independent focused practice. In other words, richer students spend more time in effortful thinking on their own during the school day.[3] In schools with such pedagogy, we also see fewer performance declines over time—even after controlling for wealth. While only correlations, these stylized patterns point to the potential for practice at school to affect cognitive endurance.

In this paper, we use a field experiment to examine whether cognitive endurance is indeed malleable. We select a school setting where the time spent in focused cognitive activity is limited: low-income primary schools in India. We randomly increase the amount of time in such activity for some students. We use this variation to examine whether simply spending more time in effortful thinking expands the underlying capacity for focus, with downstream effects on traditional measures of human capital.

We conduct our experiment with 1,650 low-income Indian students in grades 1-5. The schools in our sample exhibit many features common to low-income educational environments. Students rarely undertake much individual practice, and homework is either not regularly assigned or not completed. Aside from exams, it is rare for students to be required to sit and undertake a specific focused activity for 10-20 minutes at a time without distractions. Consequently, unlike those in richer schools, the students in our sample have very little opportunity to practice exerting focus in a sustained manner.

We construct two treatments that engage students in intellectually challenging content during the school day. In the first treatment arm (Math), students practice math problems. This mimics what good schooling does: focused activity within the context of academic learning. However, under our hypothesis, practicing any cognitively challenging task should improve sustained attention—regardless of whether students learn anything from it. Consequently, in our second treatment arm (Games), students play cognitively demanding games, such as mazes and tangrams.

[3]This may be due, at least in part, to the fact that facilitating such practice is more difficult in the crowded, more disruptive environments of lower-income schools.

There is absolutely no academic content, such as numbers or letters, present at any point in these games—providing a pure test of our mechanism. For these treatments to be effective, the content must be difficult so that concentration is effortful, but also sufficiently engaging to retain student participation for 20-minute periods. To achieve this balance, and to overcome the hurdle of heterogeneous student ability, we deliver each treatment on simple tablets—enabling students to receive content appropriate to their skill level.[4]

We compare each of these two treatment arms to a control group, which receives a status-quo math "study hall" period. As is standard in this setting, in the control group, students are assigned a small number of math problems copied from the chalkboard, and can spend the remainder of the study hall session as they'd like.[5] This results in little effective time spent in cognitive practice.

Students are randomized at the individual level to either the control group or one of the two cognitive practice treatment arms. Each treatment arm takes place during study hall or an elective period 1-3 times per week. This results in 20 minutes of effective practice time per session in the two treatment arms.[6] In total, students in each of the two treatments receive 6-15 hours of additional cognitive practice over the course of a 6 month period during the academic year, with differences in hours due to different starting times across different schools. We examine impacts at the end of the school year, as well as in a 3-month follow-up. As we discuss below, we supplement this design with additional variation to help disentangle mechanisms, such as attentional capacity versus motivation.

Each treatment improves students' capacity for sustained focus, using three distinct sets of measures. First, we test for changes in attentional capacity using traditional measures from the psychology literature: the Sustained Attention to Response Task (SART), which uses reaction times in a simple game to measure whether an individual was focused, and a symbol matching task. Relative to the control group, the Math and Games treatments each improves performance on these canonical measures, with an average effect of 0.08 standard deviations (p =xx). Second, we examine effects on classroom behavior, adapted from the Vanderbilt ADHD diagnostic teacher rating scale. This includes measures such as following instructions and physical signs of fidgeting and looking around during lecture, rated by classroom observers that are blind to treatment status. Each of the two treatment arms also improves performance on this index, with an average effect of 0.094 standard deviations (p =xx). Together, these results indicate that sustained cognitive activity—whether academic or non-academic in nature—improves basic measures of attentional capacity as typically measured by psychologists.

Third, and most directly tied to field behavior, we test for improved cognitive endurance using changes in performance declines. We implement paper-and-pencil tests in three distinct domains—

[4]For the Math arm, we use the imagineMath software, developed by Pixatel. For the Games arm, we use simple games with limited animation downloaded from the Android app store. In each arm, the tablet software provides no instruction, only the practice of problems or games.

[5]We also provide control group students time practicing using tablets in non-stimulating tasks, to remove some differences in the novelty effect of tablets. Note all our primary outcomes are measured using paper-and-pencil tasks.

[6]In the control group, students end up spending 5-10 minutes in cognitive activity per session, with many students spending less time.

listening, Raven's Matrices (IQ), and mathematics—in order to test for broad cognitive impact. For example, in the listening test, students listen to a series of short stories, each of which is followed by a series of factual questions that check whether the student was attending to the story (e.g. "What color was the cat?"). This not only captures an important input into learning in school, the content of this test is completely unrelated to the treatments: there is no sense in which they required students to practice listening. Finding effects of cognitive practice on these three disparate test domains would represent "far transfer"—consistent with changes in a generalizable cognitive capacity. On each test, we randomize question order, and allow ample time for test completion.[7]

In line with finite cognitive endurance, in each of the three test domains, control group students exhibit significant performance declines over time within each test.[8] On average, the probability of getting a question correct declines by xx% from the beginning to the end of each test (p=xx). This matches the patterns in Figure 6 above.

Consistent with our hypothesis, cognitive practice substantively mitigates these performance declines. On average, the Math and Games treatments reduce the amount of performance decline across the listening, Raven's Matrices, and math tests by 17% (p=0.041), 33% (p=0.031), and 14% (p=0.014), respectively. If these results are applied to data from TIMSS, it would cut the gap in performance declines between high and low-income countries by xx%, or between black and white students in the US by 38%. In addition, for each test, the treatments have little impact in the beginning of the test when students are still mentally fresh (e.g. in the first decile). Rather, treatment effects only emerge later in the test, when control students become more cognitively fatigued. This is consistent with the idea that the treatments improved cognitive endurance, but did not teach new content—which would reflect in changes in initial levels as well.[9]

The above three sets of results provide positive evidence for improvements in the capacity for focus. To understand the overall impact of cognitive practice, we examine effects on downstream schooling outcomes. Specifically, we measure effects on students' regular school performance in the three core academic subjects in this setting: Hindi, English, and Math. Note that neither treatment arm taught students Hindi or English—making these subjects wholly unrelated to the content of the cognitive practice.

Each of the two treatment arms significantly improves outcomes in each of these three test subjects. On average, the cognitive practice treatments raise students' performance in Hindi, English, and Math by 0.095 SD (p =), 0.086 SD (p =), and 0.073 SD (p =), respectively.[10] These results

[7] We directly verify that students complete the tests in the data. From students' perspective, these were required school tests, providing natural stakes.

[8] We include question fixed effects in the analysis to cleanly identify performance declines.

[9] There is one exception to this: As one would expect, we see some suggestive evidence that the Math treatment improves initial level effects in the math test, particularly for more difficult math questions. However, we see no evidence for level effects in listening or Ravens Matrices for either treatment. In addition, we do not see evidence that the Math and Games treatments had differential treatment effects on *declines*, perhaps due to power limitations.

[10] "SD" is standard deviations. For each academic subject, we cannot reject that the two treatments had the same impact on student performance.

indicate that simply spending time concentrating—*without learning any subject content*—improves traditional measures of human capital. This directly supports the possibility that features of schooling have the potential to alter underlying cognitive capacity.

In addition, note that these effect magnitudes are substantial, especially when compared to prominent interventions in the education literature—for example, reducing class sizes in the US (0.1 SD), tracking students by ability in Kenya (0.14 SD) or remedial education with an additional teacher in India (0.14 SD) (Krueger and Whitmore, 2001; Duflo et al., 2011; Banerjee et al., 2007). These interventions involve continuous exposure each day throughout the entire school year, and specifically target academic learning in the subjects tested. In contrast, our results arise from 6-15 hours of cognitive practice, without any academic learning (e.g. in the Games arm).

To examine persistence in treatment effects, 3-6 months after the end of treatment activities, we conduct follow-up declines tests in listening, Raven's Matrices, and math. These tests are conducted after students return to school after end-of-year vacations. We find that relative to the control group, students in each of the two treatment arms continue to show less performance decline over time, and we cannot reject that the change in treatment effects is zero. This provides evidence for some persistence, though of course does not speak to persistence over longer horizons.[11] Whether we should expect persistence over longer horizons is ambiguous. For example, if richer and poorer individuals receive different levels of attentional "practice" in their daily lives—schooling as children, or the workplace as adults (e.g. white collar vs. manual jobs)—this could perpetuate differences in cognitive endurance even if effects themselves are not long-lived. Examining these features of malleability offers interesting directions for further research.

Could channels other than sustained attention—such as improved motivation, confidence, or memory—help drive the effects we see? Regardless of the channel, improving cognitive endurance (i.e. performance declines over time) is arguably interesting, as are the large treatment effects on school performance we observe. To help better understand the mechanism, we draw on three supporting pieces of evidence. First, the fact that we do not see treatment effects in the beginning of tests, but only later on, is potentially inconsistent with these alternate channels—but especially consistent with cognitive endurance. For example, if treated students were more motivated to try harder, it is unclear why they should not do so early in the tests also, versus only later in the tests.[12] Second, we implement a more direct test for whether increased motivation reduces performance declines. For a subset of the declines tests, we randomize incentives so that some students have a chance to earn toys for higher test scores. As expected, this incentive sharply increases test performance, even at the beginning of the test—indicating high elasticity even when students are cognitively fresh. However, we see no evidence that being more motivated reduces performance declines over time. While only suggestive, these two sets of patterns are not consistent with a basic

[11] Our ability to collect data for further follow-up was halted by the Covid pandemic, which led schools to stop operating and some to shut-down completely.

[12] Mean control group performance in the first decile of the listening, Raven's, and math tests is 52%, leaving ample scope for treatment effects at the start of the test.

motivational channel as driving our main results. Third, we see strong improvements in traditional cognitive psychology measures of sustained attention, offering positive evidence for our hypothesized channel. We argue that, together, these findings point to the role of cognitive endurance in driving at least some of the overall effects we see.

We conclude by examining the broader relevance of cognitive endurance for economic behaviors—using performance declines as a suggestive marker. Using supplementary data, we replicate the patterns in Figure 6 among adults for domains outside of schooling: costly production errors among full-time piece rate data entry workers, and deterioration in decision-making among voters at the ballot box.[13] In each case, we document substantial performance declines over time—over the course of the work shift or further down the ballot. Moreover, these declines are considerably more severe for poorer individuals. While only suggestive, the consistency of these patterns in other domains provides impetus for more work examining the role of cognitive endurance in economic behaviors.

This paper makes several contributions. First, we demonstrate that a fundamental element of cognition—the ability to sustain and direct attention—is malleable in childhood.[14] These changes are not limited to the domain that is directly trained. Rather we find "far transfer" where training impacts a broad set of domains unrelated to the content practiced, indicating broad-based and generalizeable impacts. This constitutes, to our knowledge, the first evidence of far transfer effects in attentional training in any setting. Moreover, our methodological approach of detecting sustained attention using performance declines can be applied broadly, adding a new measurement tool for use by psychologists and economists.

Second, relatedly, we advance rapidly growing work in behavioral economics, which examines the implications of limited attention for economic behavior (e.g. Gabaix, 2019). This work takes as given that individuals have finite attentional capacity, and examines the subsequent implications for decision-making and behavior—for example, if individuals fail to attend to certain dimensions of the environment (e.g. Bordalo et al., 2016; Chetty et al., 2009; Hanna et al., 2014; Gagnon-Bartsch et al., 2018), or how involuntary allocation of attention can affect outcomes for the poor (e.g. Banerjee and Mullainathan, 2008; Mullainathan and Shafir, 2013; Mani et al., 2013; Kaur et al., 2021). We expand on this work by documenting that this capacity is worse for low income individuals across a range of settings (exams, worker productivity and voting behavior).

Third, we extend our understanding of the role of schooling in human capital accumulation

[13]The choice of these two examples was driven by data availability to fulfill two requirements: situations where declines over time are interpretable as cognitive fatigue effects due to the absence of obvious confounders, and heterogeneity in socioeconomic status that isn't likely to be subject to large selection effects. Data are from Kaur et al. (2015) and Augenblick and Nicholson (2015). It would be interesting in future research to examine cognitive fatigue in a broader set of behaviors.

[14]Laboratory studies document that some executive functions, such as working memory, are malleable (Klingberg et al., 2005; Jaeggi et al., 2008). However, they typically find little evidence of effects transfer beyond the specific domain that is directly trained (Bergman Nutley et al., 2011; Holmes et al., 2009; Klingberg et al., 2005; Diamond, 2013)—perhaps because sample sizes are typically quite small (e.g., 15-40 individuals per arm), limiting the ability to capture small effects. A notable exception is Berger et al. (2020), who find far transfer effects in training working memory.

(Acemoglu et al., 2012). Schooling's potential influence on basic cognition may provide an alternative explanation for the observed education, wage, and health gains observed from interventions which improve schooling quality (Chetty et al., 2009; Heckman et al., 2006; Alan and Ertac, 2018; Kautz et al., 2014). Further, differences in pedagogy and the quality of the schooling environment by income have the potential to widen disparities in such skills. More affluent students naturally obtain practice throughout their school day, inputs that many low-income students often fail to receive (Association For The Evaluation Of Educational Achievement, 2013).

2 Background: Cognitive Practice and Schooling Environments

Schooling engages students in effortful thinking for continuous stretches of time. As an example, in many schools, this feature is explicitly incorporated into pedagogy: students are required to sit and independently work on academic problems on their own. In classroom time use data from the TIMSS teacher survey, the average student spends xx% (yy%) of class time in independent practice globally (in the US). To varying degrees, other aspects of schooling—taking a test, reading a textbook, doing homework, possibly even listening to a lecture—may also engage students in effortful thinking for extended periods. In other words, school is about more than just learning content; the act of learning the content often involves periods of sustained concentration.

However, the degree to which students engage in sustained concentration varies across schools—and does so systematically by socioeconomic status. Specifically, poorer students appear to spend less time in activities that elicit sustained concentration. As an example, both globally and in the US, poorer students spend less time in focused independent practice during the school day (Figure xx). This amounts to 40% less practice among students in poorer countries compared to richer ones, or xx% less practice among more disadvantaged students in the US compared to more advantaged ones. In addition, the environmental conditions faced by poorer students—more crowded classrooms, more disruptions from fellow students, and less ability to focus on homework at home—may make it less likely that they can effectively engage in concentration, even when it is attempted (CITES).[15]

Work in cognitive psychology hypothesizes that practice—engaging in effortful thinking for continuous stretches of time—could "train" the mind's capacity for sustained attention (e.g. Chun et al., 2011). Potentially consistent with this, students who are exposed to more cognitive practice time in school exhibit much less steep performance declines over the length of the TIMSS exams (Appendix Table **??**, Col. 2). This correlation holds even controlling for income differences across students (Col. 3).[16] While these are simply correlations and therefore only suggestive, they provide

[15]These environmental conditions could also discourage teachers from attempting to engage a class in focused practice work. For example, in our setting of India, classrooms are extremely heterogeneous by ability, where half the students in a class may not be at grade level. Interviews with teachers indicate that when they assign independent practice, many children cannot even attempt the problems, and end up disrupting other students.

[16]We conduct this analysis within the global sample, where there is more variation and therefore greater power to examine these correlations.

motivational support for the possibility that exposure to periods of focused practice could affect cognitive endurance. The field experiment we design provides a clean test for this idea.

3 Experimental Design

The goal of our study is to construct a field experiment to test whether simply spending time in effortful thinking expands

cognitive endurance is malleable—and specifically whether it can be improved through cognitive practice—with downstream consequences for student

While the evidence from Section **??** provides suggestive correlational evidence that children's educational experiences may shape their long-term capacity for sustained attention, it is challenging to fully isolate causality with the TIMSS data alone. In order to test this idea explicitly, we design an intervention with students in six Indian, primary schools to vary how much of their class time is spent on activities in which students focus independently on a task for an extended period of time. We then test whether these interventions improve their capacity to sustain attention on a variety of unrelated tasks.

3.1 Intervention

Beginning in September, we randomize students into one of three different elective periods:

- *Control:* Students' elective period is the traditional study hall period where students work individually on solving several math problems which are written on the board, while a proctor grades other homework.[17] Engagement levels vary, but on average students spend about *10 minutes* on continuous, focused practice of the 30 minute period as measured by the proctor.

- *Treatments:* Both treatments were designed to boost the time spent in sustaining focus on a cognitively challenging activity. Specifically, in the treatment arms students' elective period involved working on cognitively challenging activities delivered via tablet applications. Students worked individually on their own tablet with immediate feedback on their progress, and the activities became progressively harder as students mastered them and easier if students were struggling. These activities were selected because they mirror the pedagogy of higher income environments (e.g. substantial individual practice) and because piloting showed they were particularly effective at promoting sustained engagement despite requiring cognitively challenging work in this environment. The dynamically adaptive nature of the content, the immediate feedback, and novelty of the technology resulted in students in treatment classes

[17] This activity was chosen because this practice is common among teachers overseeing study hall periods. However, the individual overseeing each classroom during this period was employed by the study and randomly rotated across experimental arms every few weeks.

spending on average 20 minutes of the 30 minute period focusing independently on the task.[18] Treated students are assigned to one of two treatment sub-arms which vary the content they receive on their tablet:

- *Math Treatment:* Students receive grade-appropriate math problems.
- *Games Treatment:* Students receive academic content-free puzzles which do not require any literacy or numeracy skills, such as mazes and tangrams.[19]

The math sub-treatment arm mimics the way in which students traditionally practice exerting directed attention over sustained periods in good schooling settings—academic practice, in our case by solving math problems. However, it also potentially boosts academic learning. In contrast, our hypothesized mechanism suggests that any sustained cognitive engagement should deliver attentional benefits. Consequently, we include an additional treatment arm, the games arm, which requires students to engage in cognitive activities that do not entail any academic learning or practice.

Students receive these 30-minute elective periods on average two times a week through January of the following year. Each experimental arm took place in a separate room. In total during the intervention period, the control group receives about 10 hours of total focused practice time during their elective periods as compared to an average of roughly 20 hours for both treatment groups.

3.2 Sample

We conducted this experiment with 1,650 students in six Indian primary schools in and around Lucknow, India. These schools serve students in low to middle-income households, with per capita incomes between $1.50 and $5 per person per day (a common range for low-cost private schools in India). All students in grades one through five of these schools (ages five to eleven) were enrolled into the program and randomized at the individual level, stratified by class section and baseline math test scores.[20] In this setting, as is common in many low-income classrooms (Bank, 2004), pedagogy is focused on rote memorization and recitation during the school day. Outside of school, students spend little time on homework or other cognitively challenging tasks. Consequently, students seldom have the opportunity to engage in focused cognitive activity for sustained periods of time either inside or outside the classroom.

[18] The treatments are intended as a proof of concept designed to fit the local environment. We hope that future research will shed further light on the features of the interventions required to promote such engagement across many environments.

[19] The specific games were chosen carefully by the study team to meet three criteria: 1) they should be dynamically adaptive to continue to challenge all students regardless of initial skill, 2) they should not be related to the outcomes of interest (e.g. no games with sound or listening were selected), 3) they should be challenging and require concerted effort, but still sufficiently engaging that students would work for an extended period. The final criteria relied heavily on piloting a variety of potential games and selecting those which appeared effective by visually judging the children's engagement.

[20] We also included income tercile in constructed strata in the subset of schools where parental income was available.

3.3 Data

School Exams Term exams are conducted for all students in each subject the student is currently taking, typically English, Hindi, math, science, social studies, and several elective subjects. These exams are part of the school systems' standard assessment of students and designed to assess mastery of the curriculum. Data from these exams was provided by the schools to the research team. Term exams are conducted after the fall term in December and at the end of the school year in March.

Student Baseline, Midline and Endline Tests Study administered tests (listening, ravens, math) and the traditional psychology measures were generally administered at four times: Baseline (September), Mid-line (December), Endline (February), and Follow-up (April).[21] Tests were conducted during students elective period. Classroom behaviors were measured at endline only by treatment blind observers. At baseline students provided information on household assets and activities outside of the school day.

Intervention Data During the treatment and control classes, we collected information on student attendance, time spent using the tablets (math and games arm), and number of math problems completed (math and control arm).

3.4 Intervention Fidelity

The randomization was smooth, with no more than the expected level of imbalance across all testing outcomes at baseline (Table A.1). In addition, attrition was low – 11% for school administered exams, 3% for experimental exams – and well balanced across experimental arms (Table A.2). Our research staff implemented the elective classes rather than having local teachers conduct the classes. This helped ensure the intervention followed the protocols described above and that the students received the correct class according to their randomization assignment. Based on random spot checks of the elective classes, we do not find any evidence of student non-compliance.

4 Effect on School Administered Exam Performance

In this section and the following section, we present the main results of the intervention on a variety of student outcomes. An important feature of basic cognition is its broad applicability; basic cognitive processes are used in nearly all activities. Consequently, in measuring outcomes, we take a broad approach and study changes across a variety of topical domains and through a variety of testing strategies. First, to test for impacts on sustained attention as an underlying mechanism, we use three approaches. First, we measure impacts on traditional psychology measures of sustained attention. Second, we examine treatment-blind observer measures of students' classroom behavior, adapted

[21] Certain tests were randomly not administered in all rounds due to logistical constraints on test administration. These logistical constraints impacted all arms of the study equally and did not result in any imbalances in measurement of outcomes.

from an ADHD teacher rating scale. Third, and most substantively, we utilize the decline approach developed in Section 2 to test whether treated students exhibit smaller performance declines over time. This third set of tests also builds in ancillary features that enable us to rule out confounds such as motivation or confidence. Finally, to better understand the magnitude of the effects and capture the net effects of attentional changes including feedback through learning in the classroom, we examine effects on students' regular school administered tests. Throughout all these measures, we test students in domains that were unrelated to the content they practiced as part of the treatment arms—enabling us to draw conclusions about whether our results capture a change in core cognitive capacity.

5 Effect on Sustained Attention

5.1 Measures of Sustained Attention from Psychology

Psychologists traditionally measure sustained attention using laboratory measures that capture whether individuals can sustain focus over time (Johnson et al., 2007; Oades, 2000). We used two such measures to test for positive evidence of attentional improvements. In addition, in order to detect changes in classroom behavior, we draw on the literature regarding diagnosis of Attention-Deficit/Hyperactivity Disorder (ADHD) and adapt elements of a commonly used ADHD diagnostic scale to the local environment.

(1) *Lab measure 1: Symbol matching.* Students are given a paper-based workbook, each page of which contains a grid of randomly ordered pictorial symbols. A specific set of 2-3 target symbols is displayed at the top of the sheet above the grid. Students are asked to go through the grid, crossing out any of the target symbols they encounter. Scores are a positive function of the number of symbols correctly crossed out and a negative function of the number of symbols incorrectly crossed out. We follow the convention in the psychology literature and measure mean performance on the task.

(2) *Lab measure 2: Sustained Attention to Response Task (SART).* Students look at a computer screen for ten minutes, during which time various shapes (i.e. stimuli) randomly appear and then quickly disappear from the screen. The student is tasked with simply pressing the space bar as quickly as possible each time a particular shape (i.e. a bell) appears to show that she has seen it (Peebles and Bothell, 2004). Overall performance is measured as a mixture of speed and accuracy common to the literature.

(3) *Classroom measure: Vanderbilt ADHD Diagnostic Teacher Rating Scale.* Students' behavior was observed in their classrooms by individuals who were blind to treatment status. Students were observed in: (1) their ability to follow directions, (2) their response to stimuli, and (3)

their physical signs of inattention.[22]

We find positive impacts of our treatment on an index of the two lab-based sustained attention, with gains of 0.08 SD ($p < 0.05$) (Table 1). These results are similar across treatment arms (column 4: 0.09 SD for Math and 0.07 for Games). In addition, the effects also appear to generalize to observable classroom behaviors. As seen in Table 2, treated students improve on an index of three measures of classroom attention adapted from a teacher rating scale used to measure ADHD by 0.09 SD, with the effects driven primarily by improved ability to attend to and follow instructions and improved responses to stimuli. Visual signs of distraction were also reduced, however the effect is not statistically significant. Similar to both the classroom tests and the sustained attention measures drawn from the psychology literature, these results are indistinguishable across treatment arms (column 5).

5.2 Measuring Sustained Attention using Performance Declines

Exam Design and Theoretical Predictions. To be able to test whether we had an effect on sustained attention *and* be able to eliminate other confounding explanations we design an additional series of tests designed to capture attention on the test itself.[23] These tests have three important features: i) ability to isolate performance at the beginning versus end of the test (capturing declines in sustained attention), ii) coverage of a variety of domains (indicating the skill is broadly generalizable), and iii) ability to isolate differences, if any, at the beginning of the exam (a key element in eliminating potential alternative channels).

First, to be able to isolate performance declines over the length of the exam, we need to be able to ensure that there are no other differences between the beginning of the exam and the end of the exam. To accomplish this, we randomize the order of the questions in each exam[24] and ensure that students have time to finish the exam, so effects cannot be driven by students running out of time. Our key prediction is that students with better sustained attention will show less decline in performance over time. If we have trained sustained attention, we expect to see a gap emerging over the length of the exam, where the control students lose focus more quickly and performance declines in the latter part of the exam.

[22]For the following instructions component students were asked to complete two activities – moving classroom supplies from one part of the classroom to another, and writing their roll number in a specific location on a paper and turning it in 5 minutes later. Failure to complete the tasks in line with the instructions indicates a failure to attend to the instructions. The response to auditory stimuli component recorded whether students are able to notice and respond to an auditory stimuli outside the classroom. The physical symptoms of inattention measured whether the student showed physical symptoms of inattention (e.g. fidgeting, looking out a window, pestering their seat-mate).

[23]All tests are conducted during the school day, either in program class time or during additional study hall periods. Test varied in length by grade, with a minimum of 15 minutes and a maximum of 30 minutes. Note that students interpreted these tests as being regular school tests.

[24]This means, for example, the same question item could occur as question 1, 10, etc. in a students' test packet. Test packets were randomized across students. The test packets were well randomized with the number of imbalances across experimental arms no more than would be expected by chance.

Second, since sustained attention is a general cognitive function, we should see effects of the intervention not just on tasks that are similar to the tasks completed in their elective classes but also on completely unrelated tasks. Effects should not only be present in tasks "similar" to those trained (e.g. math), but also in tasks unrelated to the tasks undertaken in the intervention (e.g. one's ability to listen and retain information). To determine whether such effects are present, we conduct three tests with varying distance to the tasks trained:

(1) *Listening:* This task measures students' ability to listen to a passage without losing focus, as is required in nearly all typical classroom settings. Using headphones, each student listened to a pre-recorded set of short simple stories. After each story, the student was asked questions about the content of the story, for example, "what color was the dolphin?" In order to avoid any concerns about literacy, answers were multiple choice and visual (e.g. in the above example, green, blue, black, and grey squares to denote the color of the dolphin).

(2) *Ravens Progressive Matrices:* This is a non-verbal multiple-choice test of reasoning in which the participant is asked to identify the element that completes a pattern in a figure (Raven, 1936, 2000). This test is often said to capture "fluid intelligence". Students took a shortened paper-and-pencil version of the test, adapted for appropriateness for each grade level.[25]

(3) *Math:* A standard paper-and-pencil test, which focuses on the content in the math curriculum for each student's given grade level. This test was chosen because of its direct policy relevance.

Third, to rule out a variety of potential confounds (motivation, literacy, etc), we compare how treatment versus control students perform at the beginning of the exam. For example, it is possible that the treatments could increase the confidence or motivation of treated students. However, most of these other confounds would effect students performance at the beginning of the test as well as the end (a level shifter), so testing for equality of performance in the first few questions of the exam will allow us rule out most confounds. Section 6 explains how we are able to rule out other confounds which are unable to be addressed by comparing performance at the beginning of the test.

While designing the tests with these key features helps to rule out learning other potential confounds, it comes at a cost. Specifically, these design features are likely to significantly limit the margins for treatment effects. For example, designing the tests to ensure that students finished, limits the scope for attention to influence scores through completion. In addition, testing skills which are not taught in the classroom (e.g. listening) eliminates potential positive impacts of increased attention in the classroom. These features suggest that the reductions in the rate of decline in performance is a lower bound, capturing simply attention on the test in settings with ample time. To more fully capture the magnitudes of impact of improved attention we rely on the school administered exams – are the standard metric of the total gain in human capital achieved by the

[25]While this exam typically proceeds from the easiest to most difficult questions, with the exception of a short set of easy practice questions which are not included in the analysis, the order is randomized in this case as well.

intervention – described in Section 5.3.

Performance at the Beginning versus the End of the Test. We begin by providing non-parametric local polynomial plots of the treatment effects in Figures 3, 4, and 5. Because initial performance is quite similar across tests and arms (Table 3, panel A, columns 3-5), to more clearly visualize declines initial levels are normalized to zero.[26] Each of these plots shows two consistent patterns. First, similar to the declines observed in the TIMSS data, students perform worse on a given question item if it occurs later in the test and students are cognitively fatigued. This is true despite very high test completion rates and the fact that question order is randomized and we residualize on question fixed effects to control for question difficulty. Second, consistent with the hypothesis, treated students' performance declines more slowly across the course of the exam, with improvements of 14% to 33%. Notably, students in both treatment arms experience slower declines in performance than control students in each of the tests, though the magnitudes of the effects vary. We discuss each plot in more detail below.

Listening: Initial performance is, again, not statistically different across Treated and Control students, but a gap emerges fairly rapidly and continues to grow over throughout the exam. In the final decile, treated students exhibit 17% less decline than control students, with notable differences for both treatments but somewhat larger effects among Math Practice students. Notably, this test is one which has fixed timing (e.g. one can not skip ahead), ruling out any confounds due to test-taking strategies. In addition, neither of the treatments involved any additional time listening to an instructor, suggesting that gains can not be due to additional training on the task.

Ravens: Performance on Raven's Matrices, often taken as an IQ test, shows a similar overall pattern. Overall, treated students decline 33% less in the second half of the exam. Notably, the cumulative impacts are substantial as both Math Practice and Games Practice students are able to better maintain attention from fairly early in the exam, with an increasing gap over time (Figure 4). These variations in the exact pattern of decline may be related to variations in the difficulty of the tasks and underscore the importance of the flexible functional form used to estimate these effects. Although differences are not statistically distinguishable, in contrast to the listening test, with Raven's matrices the Games Practice students appear to experience the more substantial reductions in the rate of decline.

Math: Finally, we see similar overall patterns in the math test. Similar to other low income countries in the TIMSS data, declines in performance over time are substantial among control students. Control students are roughly 15 percentage points more likely to answer a question correctly if it

[26]Plotting the raw data in the listening test, there are clear "reset" effects between passages. Hence, we examine declines within each passage which are substantial.

occurs at the beginning of the test rather than at the end of the test (Figure 5). The roughly 15 to 20 hours of treatment reduces this decline by 14% in the second half of the exam.

To test for the statistical significance of these results, we estimate:

$$Correct_{ils} = \beta_0 + \beta_1 CogPractice_s + \sum_{l=2}^{10} \lambda_l Location_{il} + \beta_2 CogPractice_s * 1[2 \leq Location_{il} \leq 5] + \beta_3 CogPractice_s * 1[6 \leq Location_{il \leq 10]} + \beta_4 Baseline_s + \chi_i + \epsilon_{ils} \quad (1)$$

Where $Correct_{ids}$ denotes whether a question item, i, in location (decile), l, for child, s, was answered correctly. β_1 captures the difference in performance at the beginning of the test. We predict $\beta_1 = 0$, with the possible exception of the math test as the groups received a differential amount of math practice. This prediction is important to ruling out potential confounds, as described further below. λ_2 - λ_{10} are location (decile) bins which flexibly capture declines in control group.[27] While performance on a wide variety of tests declines across time, the exact pattern of decline in performance varies significantly across tests. These variable patterns of decline motivate our non-parametric empirical approach. β_3 is primary coefficient of interest, indicating whether there is differential fatigue among treated students. We hypothesize that β_3 will be positive (the treatments will ameliorate the rate of decline). $Baseline_s$ controls for the child's baseline test scores. χ_i are question fixed effects, controlling for the difficulty of the test item. For inference, we cluster standard errors by student—the unit of randomization—throughout the analysis.

As shown by the coefficient on the treatment dummy in Table 3, panel A, column 1, there are no significant differences in initial levels of performance across any of the tests. Overall, the level difference between the Treatment and Control students is -0.0027, or -0.5%. Also consistent with our hypothesis, pooled across all three exams, the reductions in the rate of decline among treated students are both meaningful in terms of magnitude — roughly one-fifth of the total decline is ameliorated — and highly statistically significant. We also see suggestive evidence of effects earlier in the test, with a coefficient magnitude roughly two-thirds as large as in the second half of the test.

These effects are even more notable given the relatively limited training in this program and the diverse subject matter tested. Students spend fewer than 20 hours in this program, yet spend roughly 800 to 1,000 hours per year in instruction and practice at school. While the training effects may not be linearly additive over time, they do suggest that even small differences in the instructional quality could have a substantial impact on the ability to sustain attention over time. Further, the school administered tests suggest that gains in attention may aggregate further through direct learning in the classroom for those skills taught in the classroom.

Results are also similar for both the math and games sub-treatment arms. Each of the treat-

[27] To account for varied test lengths, we use question item as a proxy for elapsed time and normalize the length of all tests to 100%. In specifications where we pool across all test subjects (i.e. listening, math, ravens), we interact the decline bins with test subject to allow each subject to have its own decline rate. Note that all regressions also include fixed effects for the (randomly assigned) version of a given test taken by a student.

ments was designed to increase the time spent sustaining focus on a task. As such, we would also expect that both treatments should be effective at mitigating declines in performance in the later portion of the test. Consistent with this hypothesis, the results in Table 3, panel B, column 1 show very similar impacts of the two treatments, with coefficients on the effect in the second half of the test of 0.0127 and 0.0131 for Math and Games, respectively.

Allowing for Flexible Decline During Exam. One potential limitation in estimating Equation 1 is that the estimation (inflexibly) takes a stance on the rate of decline. However, it is not clear ex-ante when such declines should occur. For example, the rate of decline may be a function of the difficulty of the content, the length of the exam, or baseline attentional capacity of students. Declines (and therefore the potential for treatment effects) could set in at the third decile, or the seventh. Hence, we supplement our first specification with an additional higher-power specification which takes a two-step approach, using the baseline data to tell us when to expect declines. We first flexibly estimate the rate of decline in the *baseline* data according to Equation 2, and then use this variable as a proxy for the "predicted" decline at each point in the endline tests in estimating Equation 3. This approach draws on the intuition that there is no scope for a treatment effect unless there is a decline, and allows us to estimate the fraction of the decline mitigated by the treatment. The predicted decline variable is estimated for each school separately to account for variation in skill across schools. This approach relies on the assumption that baseline decline rates are predictive of endline decline rates over time—an assumption we can directly verify through coefficient β_2 in Equation 3 below.

$$PredictedDecline_l = \frac{1}{SI}\sum_{s=1}^{S}\sum_{i=1}^{I}[(Correct_{ils}|quintile = 1) - (Correct_{ils}|quintile = l)] \quad (2)$$

$$Correct_{ils} = \beta_0 + \beta_1 CogPractice_s + \beta_2 PredictedDecline_l + \beta_3 CogPractice_s * PredictedDecline_l + \beta_4 Baseline_s + \chi_i + \epsilon_{ils} \quad (3)$$

In this specification, β_3 captures the extent to which the treatments mitigate the rate of decline. As with Equation 1, we predict that β_3 will be positive. To account for the fact that $PredictedDecline_l$ is estimated from the data, we bootstrap standard errors.

Results using Equation 3 are quite similar, with coefficients of 0.098 and 0.087, respectively, for the Math Practice and Games Practice arms (Table 3, panel B, Column 2). We also see that effects are similar across all three exams: listening, ravens and math. We also examine the effects of the pooled interventions on each exam individually in Table 3 panel A columns 2-5. As in the pooled analysis, initial level differences are consistently small across each of the three tests. Notably, the treatment effects are statistically significant and relatively similar in magnitude across each of the

tests, ranging from 0.07 for Listening to 0.11 for Math.

We further disaggregate these results to examine the treatment effect of each arm on each test in columns 3 through 5 of panel B. The impact of the treatments are generally a similar order of magnitude and statistically indistinguishable from each other, though with less precise estimates as the results are more fully decomposed. As expected, there is a slightly larger initial level effect for the Games arm on the math test, as the Games arm did not receive any additional math practice.

5.3 School Administered Exams

To measure overall gains in human capital, we first examine students' test scores on the end of term exams administered by schools—the standard field measure for educational impacts. We look at the three core subjects taught and tested by all schools in our sample: Math, Hindi, and English. Note that seeing effects across these subjects would in itself be indicative of broad "transfer" effects, consistent with the idea that our treatment affected an underlying cognitive resource. This is because, for example, neither of our treatment arms provided any exposure to Hindi, and the Games arm arguably provided no academic training across subjects.

Analysis of the school administered tests rely primarily on a simple intention to treat analysis. Specifically, we estimate:

$$z\_score_s = \beta_0 + \beta_1 Treated_s + \beta_2 Baseline_s + \gamma_1' X_i + \epsilon_s \tag{4}$$

Where z_score_s denotes the normalized score for the student. $Treated_s$ represents a treatment indicator (or indicators for each treatment arm, when treatment arms are disaggregated). The regression also includes a baseline measure of the outcome when available – $Baseline_s$ – as well as a vector of relevant controls X_i as noted in each table. When more than one observation per student is available, standard errors are clustered by student.

Treated students show noticeable improvements on the exams administered by the schools. Pooling across the core subjects of English, Hindi, and Math we find a 0.08 SD improvement in performance (Table 4, panel A, column 1). This improvement is notable given the short duration of the intervention – an average of approximately 6-15 hours of additional focused practice for the treatment group integrated into schooling across four months. As seen in columns 2 through 4, these results are not simply driven by effects of the Math treatment arm on the math exam. Rather, performance improvements are consistent across each of the subjects despite the fact that neither treatment arm trained in English or Hindi.[28] Further, as seen in panel B, the point estimates between the two treatment arms are very similar in magnitude (0.08 SD for both arms) and are statistically indistinguishable, suggesting both treatments were effective. As shown in Panel B,

[28] A subset of questions for the Math arm did include English text (e.g. Add 1 and 4). However, the questions for the Control arm study hall were drawn from the same question bank. Notably neither the Math nor the Games arm involved any additional exposure to Hindi.

columns 2 through 4, the effect of each sub-treatment is also similar when disaggregating across tests.

Benchmarking Magnitudes. When scaled for time of the intervention, the effects are similar in magnitude to other computer assisted learning interventions in India (Muralidharan et al., 2019). The magnitude of the increase is similar to that found in Project Star, which substantially reduced classroom sizes for an entire year (Krueger and Whitmore, 2001). Similarly, the effects are also only slightly smaller than those found for tracking students by ability (0.14 SD) or a remedial education program with an additional teacher (0.14 SD) with continuous exposure to the interventions over an entire school year in each of these programs (Duflo et al., 2011; Banerjee et al., 2007).

5.4 Persistence of Effects

To enable a test for persistence in effects, we conducted a three-month follow-up, in which we again administered tests in math, listening, and Ravens Matrices. These tests were conducted after a one-month break between when students progress from one grade to the next; during this break, students do not attend school—providing a stronger test for persistence. The prior analysis pooled these follow-up tests with the main midline and endline tests for power. In Table 5, we separately estimate the treatment effects in the 3-month follow-up by adding an interaction term for the follow-up tests. We show these results for each of our two estimation strategies (Equations 1 and 3) in columns 1 and 2 respectively. In each specification, the interaction term is essentially zero and insignificant. In addition, in the higher-powered specification in column 2, treated students show significantly less decline than control students (p-value 0.0325, reported at the bottom of the table). We further decompose these results to examine the impact of each treatment independently and find similar and statistically indistinguishable effects across the two arms (columns 3 and 4).

The treatments produce broad and generalizable improvements in the ability to sustain focus over time. Treated students improve performance on a range of classroom test by 0.1 SD and experience 14% to 33% less decline on unrelated tasks using tests to cleanly isolate attention as a key channel driving these effects. The results are further supported by improvements in both cognitive psychology tasks designed to measure sustained focus as well as in observations of classroom behavior.

Taken together, these results suggest that basic cognitive function is malleable and that a simple school-based intervention can improve children's ability to sustain focus over time. These patterns also support our view that the treatments do not teach new content within the test domains, but rather improve students' ability to sustain cognitive effort over time. Further, the improvement across a wide range of domains (e.g. Hindi, English, Math, listening, IQ) as well as test types ranging from traditional measures of human capital accumulation such as school administered tests to standard cognitive psychology measures suggests that these effects are very widely applicable and are likely to be large when aggregated across the many contexts in which they apply.

6 Potential Confounds

The study administered tests were explicitly designed to isolate the mechanism of sustained attention while ruling out potential confounds. We use a variety of test design features, cross randomizations, and additional analyses detailed below to address these potential confounds.

6.1 Confidence, learning, and motivation

Although designed to target the ability to sustain focus, it is possible the treatments could improve the students' confidence, motivation or interest in school. In the Math sub-treatment, math skills may also be improved. Or, in either treatment, fine motor skills could improve. However, each of these mechanisms would result in a level-shift throughout the exam: they should improve performance in both the early parts of the test and the later parts of the test. In contrast, our mechanism predicts that the treatments will ameliorate declines rather than generate uniform improvements. Correspondingly, the prediction of reduced decline in performance relative to the Control over the course of the exam is a very specific to the mechanism of sustained attention. Since we do not find any treatment effects in the first decile of the test, we can eliminate any of these mechanisms which should be constant across the exam.

Learning. In addition to the test described above, the wide variety of tests used allow us to rule out learning as a potential confound. The Math treatment may serve to directly alter math skills or change the cognitive costs of completing math questions through a variety of mechanisms (e.g. solving math problems requires less effort with additional practice). However, the games treatment is devoid of such effects. In addition, neither treatment teaches listening or reasoning skills, ruling out learning as a channel for the listening and Ravens Matrices tests.

6.2 Persistence, resilience, or grit

Motivation may not be consistent across time. Rather, effects on motivation could come in the form of "trying harder when you are tired" (i.e., persistence) or be in response to facing challenges (i.e. "grit" or resilience) (Duckworth and Duckworth, 2016). The level shifter test described above would not fully rule out such dynamic versions of motivation. Hence, in order to address this concern we undertook two additional tests. First, to address persistence, we incentivize a sub-set of the tests via enticing prizes (e.g. toys, colored pencil sets, etc).[29] The incentives are indeed effective at generating increased effort on the tests – performance improves at the beginning of the exam. However, the additional incentives do not lead to a shallower rate of decline or have differential effects between the control and treatments, either in the initial level of performance or in the rate

[29] We ensured that the prizes were appealing throughout the distribution of performance by offering increasing prizes by place in the score distribution. Students could choose a specific prize among a set designated for their quartile of performance.

of decline (see Table 6).

Second, to consider whether the treatments influenced "grit" or resilience, we leverage the tests' random question ordering. Specifically, by chance, some students received a version of the test which began with easier questions and some received a version which began with more difficult questions. We test if the appearance of a difficult question — either early in the test, or generically throughout — lowers later performance. We find no such impacts across many different definitions of "difficult" questions. Specifically, only 1 of 16 specifications results in a statistically significant impact, suggesting the results are likely due to chance.

6.3 Test-taking strategies and exposure to technology

While not trained directly, it is possible the Math treatment may help students intuit better test taking strategies, such as skipping hard questions. However, this skill is not trained in the Games arm, where the games do not permit strategies of these types. Additionally, we address this concern by designing the tests to ensure sufficient time and high completion rates. Over 95% of students reach the final question on all exams (Table A.3). Test monitors were also instructed to look for such "skipping" behavior, but it was only very rarely noted given the young age of the students. In addition, a subset of our tests (e.g. listening) mechanically do not permit students to skip around or move faster through the tests. Finally, results are very similar if we restrict to attempted questions.

Similarly, because the training occurs on tablets, which are a novel technology for some of the students, it is possible that the treatment students will simply become more familiar with the technology. To rule out this potential confound, all of the primary outcome measures and all but one of the secondary outcome measures are paper-and-pencil-based.[30]

In short, the richness generated by the random ordering of questions, diverse tests, and multiple treatment arms helps us distinguish our proposed mechanism from confounding explanations.

7 Cognitive Endurance and Poverty

7.1 Attention in Economic Life

Attention is a core cognitive resource which underlies all activity. Hence, schooling is just one setting in which cognitive endurance is likely to impact performance. Acting as a constraint on processing power, attentional limits influence economic decision-making in myriad ways (Gabaix, 2019). Sustained attention, the ability to direct attention toward a task over time, is also particularly

[30]SART, which must be electronic to accurately measure reaction times, is computer-based. However, the task does not require any knowledge of technology – one simply presses the space bar when a stimulus appears on a screen – and we specifically administered the task on a computer with a large independent keyboard to make it as distinct as possible from the tablet-based interventions.

relevant to performance in a wide variety of professions. For example, Brachet et al. (2012) find that fatigue during long paramedic shifts "result in a 0.76 percent increase in 30-day mortality." Danziger et al. (2011) find that judges become significantly harsher in their judgements as their shifts progress, but leniency returns following a break. More broadly, Warm et al. (2018), documents the crucial role of attentional capacity in a wide variety of professions such as sentries, truck drivers, air traffic control operators, and industrial quality control. In each of these domains, error rates and attentional lapses increase over time with economically meaningful impacts on productivity and outcomes with long-run consequences such as criminal conviction and mortality.

7.2 Heterogeneity in Cognitive Endurance

Although it is well known that attention is central to economic decision-making, it is an open question whether there is systematic heterogeneity in this skill. As documented in Section **??**, the performance of children from less wealthy backgrounds declines much more rapidly than that of wealthier children. Next, we document that this pattern holds not only among children in school, but also among workers engaged in data entry and voters making decisions regarding ballot initiatives.

Worker Productivity. Performance in the work place is another important setting where sustained attention is crucial. Using data from Kaur et al. (2015), we examine hourly performance of full-time data entry workers over nine months. Workers' earnings were comprised of a piece rate for each *accurate* field entered, making errors costly. On average, error rates increase roughly 12% between 10am and 4pm.[31,32] In Figure 7(b), we show less educated workers experience a decline in accuracy that is twice as large as that of more educated workers. Overall, the magnitude of the decline among less educated workers amounts to 10% of the productivity gap between more and less educated workers in the sample.

Voting Behavior. Finally, Augenblick and Nicholson (2015) provide evidence of similar declines in attention in voting behavior. Using quasi-random variation in the order of ballot initiatives, the authors find that individuals are substantially more likely to vote the default option when items are further down-ballot. These effects are substantial enough to alter the outcome of 6% of the propositions in their data set. Using racial composition of a neighborhood as a proxy for socio-economic status, we find that individuals in less affluent neighborhoods are much more likely to rely on the default option when items are further down-ballot (Figure 7(c)).

[31] Analysis conducted using 10 am - 4 pm to avoid compositional effects of workers arriving and departing.

[32] To make up for the loss in productivity from the attentional decline, worker piece rates would need to increase by 2.4% by the end of the workday to achieve the same output as in the beginning of the day. Piece rate calculations are based on the effort elasticity of 0.33 estimated in Kaur et al. (2015).

Summary. Declines in performance over time are relatively universal and meaningfully influence important economic outcomes. Yet, these declines are not evenly distributed. Rather, this limited sustained attention is much worse for lower-income individuals, accounting for meaningful differences in overall performance in three disparate domains: education, worker productivity, and decision-making. The fact that these differences appear across three quite different contexts suggests that even if such differences are small in any one domain, they are likely to generate potentially sizable aggregate effects which further disadvantage the poor.

8 Conclusion

It is well known that schooling influences both "hard skills" such as literacy and numeracy and "soft skills" such as disciplinary behavior and socioemotional regulation. This paper suggests it may go further and shape a basic element of human capital essential to many aspects of life: attentional capacity.

Providing roughly 20 hours of focused practice to low-income students in India improves overall performance in a variety of subject areas – ranging from Hindi to Math – and reduces the rate of decline in performance across a wide range of unrelated tasks – such as listening retention and Raven's matrices. These gains were complimented by improvements in traditional cognitive psychology measures of sustained attention as well as changes in classroom behavior.[33]

While these results are encouraging, viewed from a different angle, this result also suggests differences in the quality of schooling or reliance on pedagogy that does not promote practice of this skill may widen long-run economic disparities. The foundational nature of this skill means that even small differences in attentional declines may have broad and economically meaningful consequences when aggregated. For example, we provide evidence of declines in attention in areas as diverse as schooling, errors at work, and voting. Further, we document the novel fact that the rate of decline in attention varies substantially with socio-economic status. Reducing these declines by the average change among treated students in the RCT could result in broad and economically meaningful improvements in these outcomes.

Yet, these are just a few of the many domains that would be affected. Further research to flesh out the scope and consequences of these declines in additional domains would allow us to more fully assess the downstream consequences of the improvements in the ability to sustain focus. For example, are high rates of traffic accidents in developing countries also potentially influenced by rapid attentional declines, for example during long shifts worked by truck and auto drivers?

[33]It is important to note that one challenge in promoting improvements in attentional capacity is that it is unlikely that a single activity can be used to train this skill across diverse populations. What holds a child's attention in a low-income school in India may not be what holds a child's attention in an affluent suburb in the United States. Given this challenge, we view this intervention as a proof of concept. While there is some intuition about what activities are likely to train this skill (e.g. active participation rather than rote memorization, dynamically adaptive tasks rather than static ones), much more needs to be done to understand what features of an activity generate sustained engagement over time.

These potentially broad impacts also suggest an additional potential micro-foundation for the persistence of achievement gaps between the rich and the poor. These gaps may create feedback loops (e.g. via jobs that do not reinforce cognitive focus, or enrollment in lower quality schools) that could contribute to the intergenerational transmission of poverty. While only suggestive, our findings provide impetus for further research testing for the role of cognitive training in schooling, and its potential impacts on cognition and productivity.

Further, these results raise the question of whether schooling has the potential to influence other elements of cognition. Overall, our study provides an initial proof of concept that one important component of basic cognition—attentional capacity—is malleable. More speculatively, our findings speak to the interpretation of the correlation between socio-economic status and cognitive performance (Banerjee and Mullainathan, 2008; Balart et al., 2018; Lawson et al., 2014; Hackman et al., 2015). To understand whether this gradient is a cause or a consequence of poverty it is crucial to determine whether basic cognitive skills are malleable and how they are formed. Examining the extent to which our cognitive capacity is shaped by the environment is an important direction for future research, with implications for understanding the the role of schooling in productivity as well as the underpinnings of economic mobility and inequality.

References

Acemoglu, Daron et al., "What does human capital do? A review of Goldin and Katz's the race between education and technology," *Journal of Economic Literature*, 2012, *50* (2), 426–63.

Alan, Sule and Seda Ertac, "Fostering Patience in the Classroom: Results from Randomized Educational Intervention," *Journal of Political Economy*, October 2018, *126* (5), 1865–1911.

Augenblick, Ned and Scott Nicholson, "Ballot position, choice fatigue, and voter behaviour," *The Review of Economic Studies*, 2015, *83* (2), 460–480.

Balart, Pau, Matthijs Oosterveen, and Dinand Webbink, "Test scores, noncognitive skills and economic growth," *Economics of Education Review*, April 2018, *63*, 134–153.

Banerjee, Abhijit V and Sendhil Mullainathan, "Limited Attention and Income Distribution," *American Economic Review*, April 2008, *98* (2), 489–493.

_ , Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying education: Evidence from two randomized experiments in India," *The Quarterly Journal of Economics*, 2007, *122* (3), 1235–1264.

Bank, World, "World Development Report," 2004.

Berger, Eva M, Ernst Fehr, Henning Hermes, Daniel Schunk, and Kirsten Winkel, "The Impact of Working Memory Training on Children's Cognitive and Noncognitive Skills," *Working Paper*, 2020.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Competition for Attention," *The Review of Economic Studies*, April 2016, *83* (2), 481–513.

Brachet, Tanguy, Guy David, and Andrea M Drechsler, "The effect of shift structure on performance," *American Economic Journal: Applied Economics*, 2012, *4* (2), 219–46.

Chetty, Raj, Adam Looney, and Kory Kroft, "Salience and taxation: Theory and evidence," *American economic review*, 2009, *99* (4), 1145–77.

Chun, Marvin M., Julie D. Golomb, and Nicholas B. Turk-Browne, "A Taxonomy of External and Internal Attention," *Annual Review of Psychology*, January 2011, *62* (1), 73–101.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso, "Extraneous factors in judicial decisions," *Proceedings of the National Academy of Sciences*, 2011, *108* (17), 6889–6892.

Diamond, Adele, "Executive Functions," *Annual Review of Psychology*, January 2013, *64* (1), 135–168.

Duckworth, Angela and Angela Duckworth, *Grit: The power of passion and perseverance*, Vol. 234, Scribner New York, NY, 2016.

Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya," *American Economic Review*, 2011, *101* (5), 1739–74.

Foy, Pierre, Michael O. Martin, Ina V.S. Mullis, and Gabrielle Stanco, "Reviewing the TIMSS and PIRLS 2011 Achievement Item Statistics," *Technical Report*, 2011.

Gabaix, Xavier, "Behavioral Inattention," in B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, eds., *Handbook of Behavioral Economics: Foundations and Applications*, Vol. 2, Amsterdam: Elsevier/North-Holland, 2019.

Gagnon-Bartsch, Tristan, Matthew Rabin, and Joshua Schwartzstein, *Channeled attention and stable errors*, Harvard Business School, 2018.

Hackman, Daniel A, Robert Gallop, Gary W Evans, and Martha J Farah, "Socioeconomic status and executive function: Developmental trajectories and mediation," *Developmental science*, 2015, *18* (5), 686–702.

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein, "Learning through noticing: Theory and evidence from a field experiment," *The Quarterly Journal of Economics*, 2014, *129* (3), 1311–1353.

Heckman, James J, Jora Stixrud, and Sergio Urzua, "The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior," *Journal of Labor economics*, 2006, *24* (3), 411–482.

Holmes, Joni, Susan E. Gathercole, and Darren L. Dunning, "Adaptive training leads to sustained enhancement of poor working memory in children," *Developmental Science*, July 2009, *12* (4), F9–F15.

Jaeggi, S. M., M. Buschkuehl, J. Jonides, and W. J. Perrig, "Improving fluid intelligence with training on working memory," *Proceedings of the National Academy of Sciences*, May 2008, *105* (19), 6829–6833.

Johnson, Katherine A., Simon P. Kelly, Mark A. Bellgrove, Edwina Barry, Marie Cox, Michael Gill, and Ian H. Robertson, "Response variability in Attention Deficit Hyperactivity Disorder: Evidence for neuropsychological heterogeneity," *Neuropsychologia*, January 2007, *45* (4), 630–638.

Kaur, Supreet, Michael Kremer, and Sendhil Mullainathan, "Self-control at work," *Journal of Political Economy*, 2015, *123* (6), 1227–1277.

_ , Sendhil Mullainathan, Suanna Oh, and Frank Schilbach, "Do Financial Concerns Make Workers Less Productive?," Technical Report, National Bureau of Economic Research 2021.

Kautz, Tim, James J Heckman, Ron Diris, Bas Ter Weel, and Lex Borghans, "Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success," Technical Report, National Bureau of Economic Research 2014.

Klingberg, Torkel, Mats Johnson, Christopher G Gillberg, and Helena Westerberg, "Computerized Training of Working Memory in Children With ADHD-A Randomized, Controlled Trial," *J. AM. ACAD. CHILD ADOLESC. PSYCHIATRY*, 2005, p. 11.

Krueger, Alan B and Diane M Whitmore, "The effect of attending a small class in the early grades on college-test taking and middle school test results: Evidence from Project STAR," *The Economic Journal*, 2001, *111* (468), 1–28.

Lawson, Gwendolyn M, Cayce J Hook, Daniel A Hackman, Martha J Farah, James A Griffin, Lisa S Freund, and Peggy McCardle, "Socioeconomic status and neurocognitive development: Executive function," *Executive Function in Preschool Children: Integrating Measurement, Neurodevelopment, and Translational Research. American Psychological Association*, 2014, pp. 1–28.

Mani, A., S. Mullainathan, E. Shafir, and J. Zhao, "Poverty Impedes Cognitive Function," *Science*, August 2013, *341* (6149), 976–980.

Mullainathan, Sendhil and Eldar Shafir, *Scarcity: Why having too little means so much*, Macmillan, 2013.

Muralidharan, Karthik, Abhijeet Singh, and Alejandro J Ganimian, "Disrupting education? Experimental evidence on technology-aided instruction in India," *American Economic Review*, 2019, *109* (4), 1426–60.

Nutley, Sissela Bergman, Stina Söderqvist, Sara Bryde, Lisa B. Thorell, Keith Humphreys, and Torkel Klingberg, "Gains in fluid intelligence after training non-verbal reasoning in 4-year-old children: a controlled, randomized study: Fluid intelligence gains after training non-verbal reasoning," *Developmental Science*, May 2011, *14* (3), 591–601.

Oades, Robert D, "Differential measures of 'sustained attention' in children with attention-deficit/hyperactivity or tic disorders: relations to monoamine metabolism," *Psychiatry Research*, March 2000, *93* (2), 165–178.

Peebles, David and Daniel Bothell, "Modelling Performance in the Sustained Attention to Response Task.," in "ICCM" 2004, pp. 231–236.

Raven, John, "The Raven's Progressive Matrices: Change and Stability over Culture and Time," *Cognitive Psychology*, August 2000, *41* (1), 1–48.

Raven, John C., "Raven. Mental Tests Used in Genetic Studies: The Performances of Related Individuals in Tests Mainly Educative and Mainly Reproductive," *Unpublished master's thesis, University of London*, 1936.

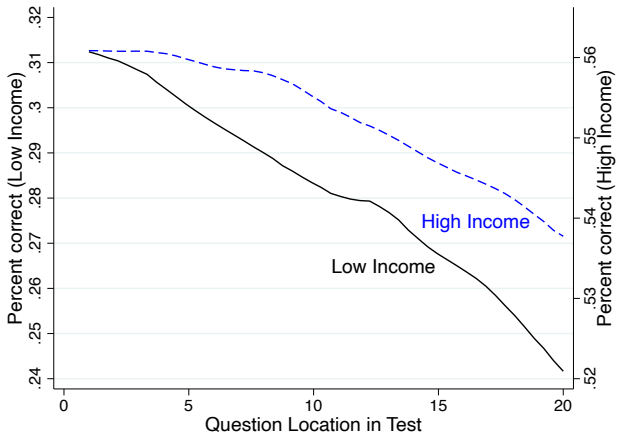
Association For The Evaluation Of Educational Achievement, "TIMSS 2011 Assessment," Technical Report, TIMSS PIRLS International Study Center, Lynch School of Education, Boston College, Cambridge, MA 2013.

Warm, J. S. and W. N Dember, "Tests of vigilance taxonomy," in "Viewing psychology as a whole: The integrative science of William N. Dember," American Psychological Association, 1998, p. 87–112.

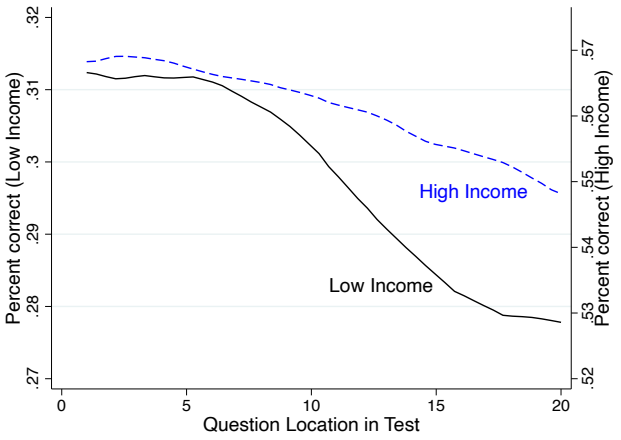
Warm, Joel S, Gerald Matthews, and Victor S Finomore Jr, "Vigilance, workload, and stress," in "Performance under stress," CRC Press, 2018, pp. 131–158.

9 Figures

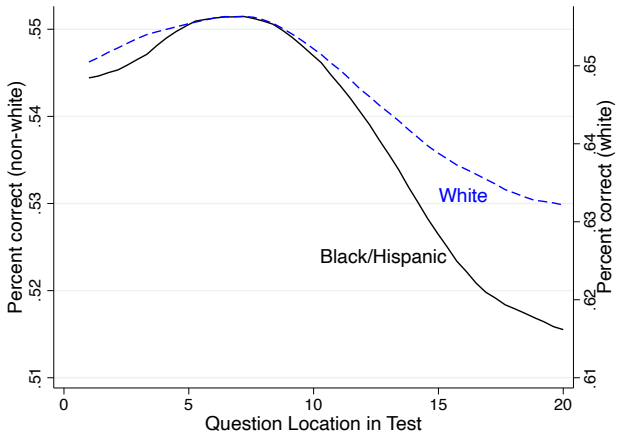
Figure 1: Declines in Achievement Tests



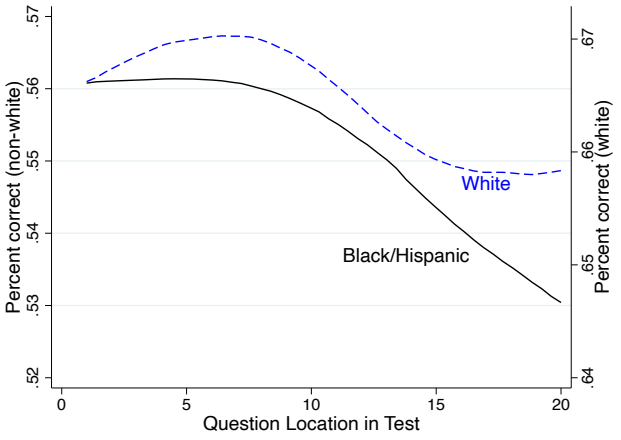
(a) Mathematics, global sample



(b) Science, global sample



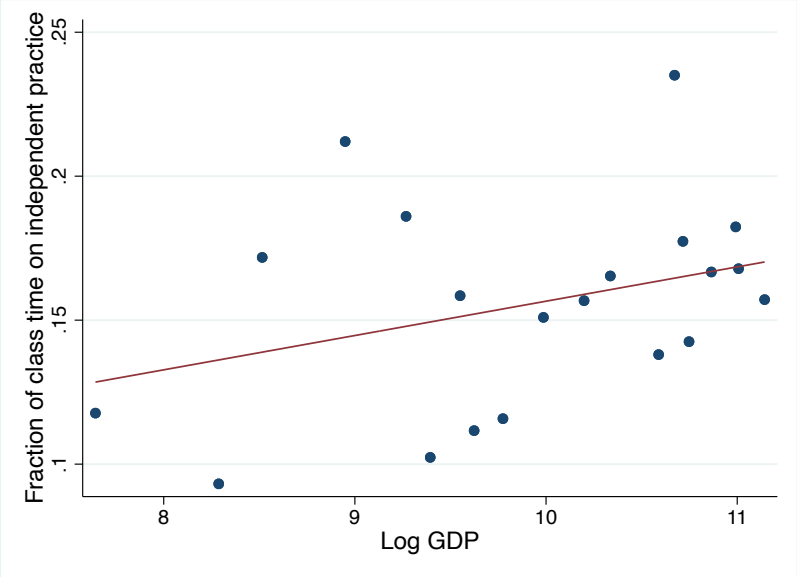
(c) Mathematics, US sample



(d) Science, US sample

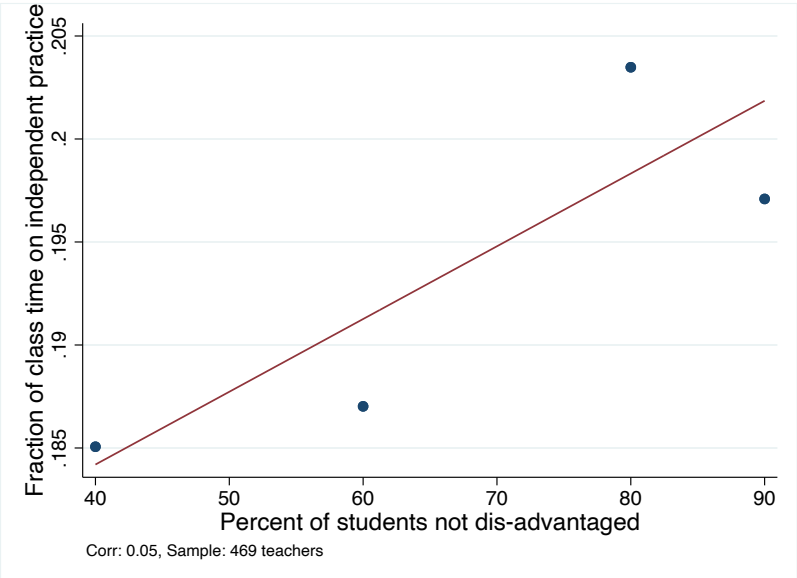
Notes: TIMSS data, authors' calculations. Question order is block randomized. In the global sample (subfigures (a) and (b)), high (low) SES countries are proxied by the top (bottom) decile of GDP/capita. In the US sample (subfigures (c) and (d)), high and low SES students are proxied by race (white and non-white, respectively).

Figure 2: Pedagogical differences across schools



(a) Global sample

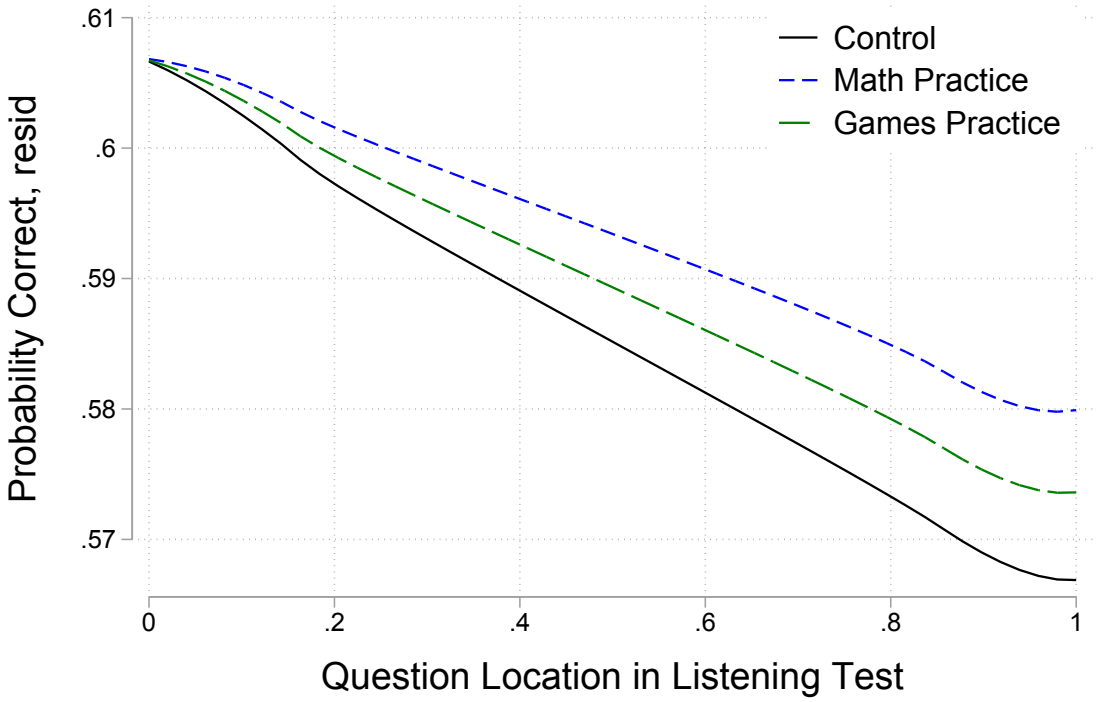
Notes: Higher income students spend 40% more time "practicing material on their own." Time use in the classroom is provided by teachers for a variety of activities on a four point scale of "never" (coded as zero) to "every or almost every lesson" (coded as 0.75). The fraction of time spent on practicing on their own is calculated by taking time practicing and dividing by the sum of total time across all activities on this scale.



(b) US sample

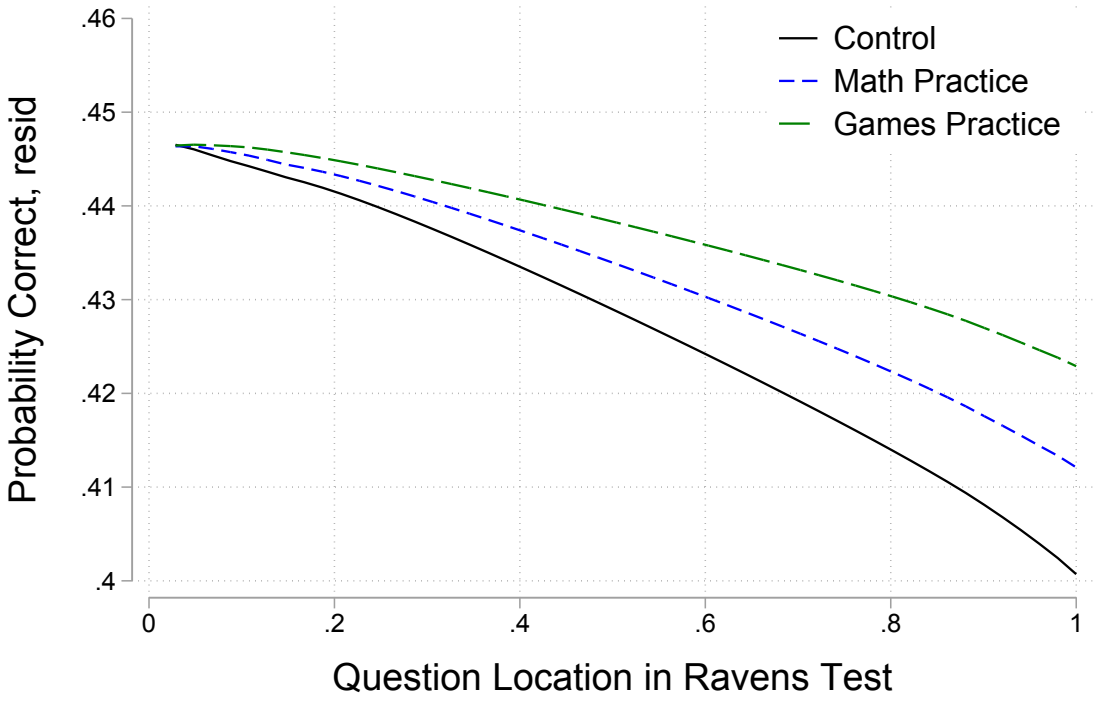
High-income students in the US are also more likely to spend time in independent practice than low income students. The percent of students who are disadvantaged is rated by the principal (n = 469) and falls into one of four categories. Time use in the classroom is provided by teachers and calculated as described in Figure 3(a).

Figure 3: Training Slows Decline in Listening Comprehension Test



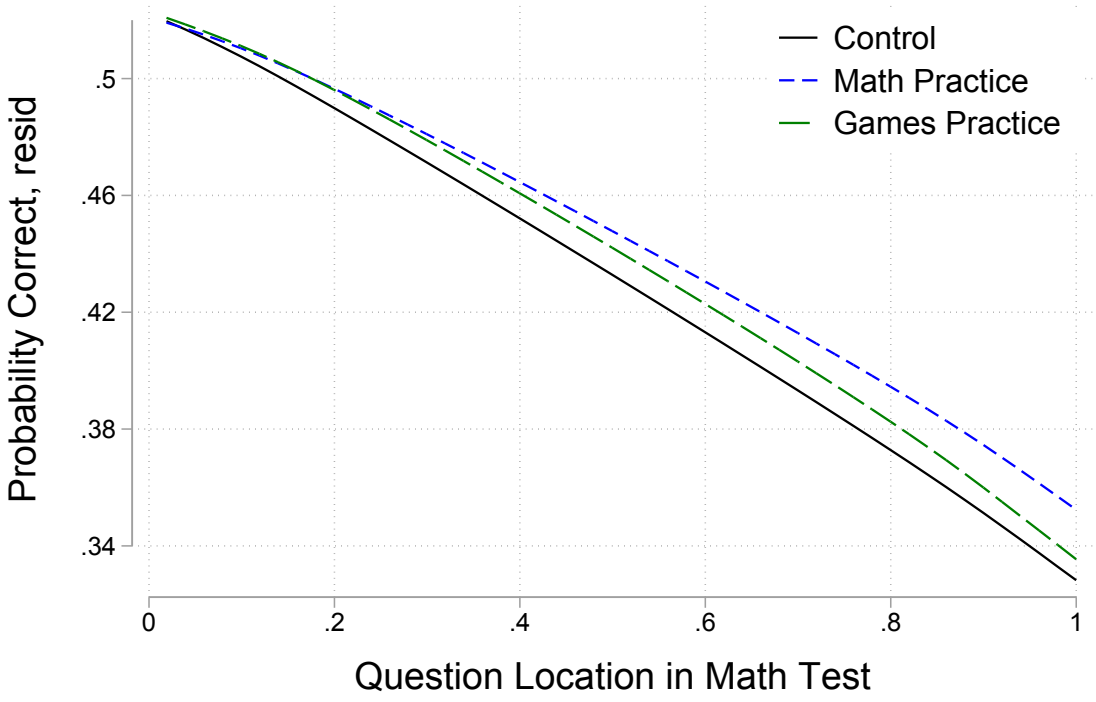
Treated students exhibit **17% less decline** in the second half of the exam (Table 3).

Figure 4: Training Slows Decline in Ravens Matrices (IQ) Test



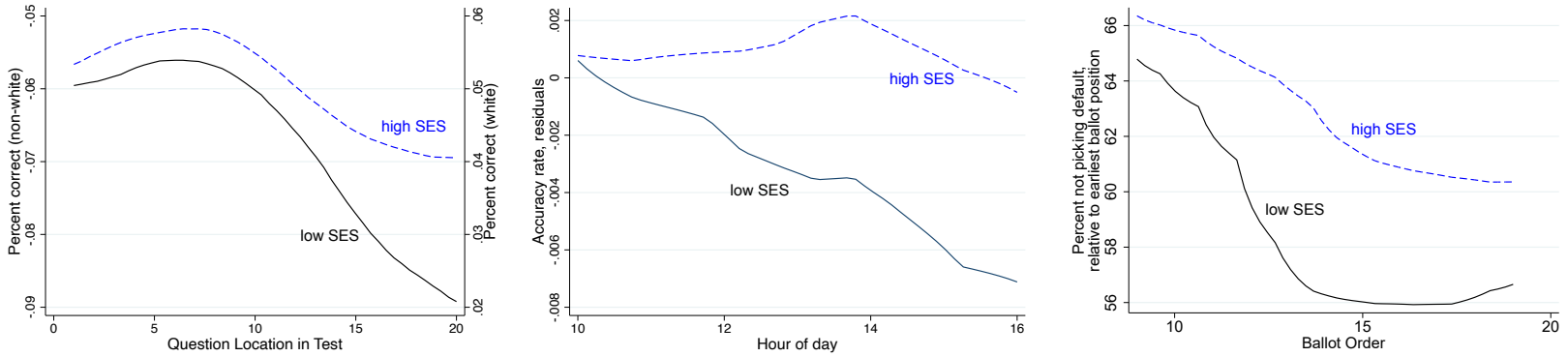
Treated students exhibit **33% less decline** in the second half of the exam (Table 3).

Figure 5: Training Slows Decline on Math Test



Treated students exhibit **14% less decline** (0.1 SD) in the second half of the exam (Table 3).

Figure 6: Heterogeneity in Cognitive Endurance



(a) Education (b) Worker Accuracy (c) Cognitive Effort in the Voting Booth

Notes: This figure plots the declines in performance across time in three disparate domains: education, data-entry work, and voting.

- *Figure (a) Education: TIMSS data, authors' calculations. Question order is block randomized. The figure pools data from both the science and math exams conducted in the US sample. High and low SES students are proxied by race (white and non-white, respectively).*

- *Figure (b) Worker Accuracy: Data from Kaur et al. (2015). Accuracy rate is the proportion of fields entered with no errors. Data are residualized accounting for worker fixed effects. High SES is defined as 1 if the worker has above high school education (corresponding to the median split of the sample). The sample is 8,382 worker-hours of data entry (90 workers). The sample is restricted to paydays (when attendance is high to mitigate selection concerns) and workers who were present from 10am-4pm on a given day (so that the composition of workers is constant within a worker-day during these hours). Patterns are similar without these restrictions.*

- *Figure (c) Cognitive Effort in the Voting Booth: Data from Augenblick and Nicholson (2015) and the United States census. Item order in the voting data is quasi-random. Voters become more likely to rely on defaults as an initiative is further down-ballot and these differences are larger for less affluent neighborhoods as proxied by a lower fraction of non-Hispanic white individuals in the precinct using a median split.*

10 Tables

Table 1: Psychology Literature Sustained Attention Tests

	Dependent Variable: Z-score Test Subject			
	Pooled (1)	SART (2)	COS (3)	Pooled (4)
Cognitive Practice	0.0814**	0.1080**	0.0650	
	(0.0371)	(0.0543)	(0.0449)	
Math Practice				0.0883**
				(0.0429)
Games Practice				0.0747*
				(0.0434)
Students	1636	1380	1628	1636
Observations	9704	3897	5807	9704

*Notes: Outcomes are measured as (True positive z-score - False positive z-score), winsorized at the 99th percentile. Clustered standard errors in parentheses. Controls for baseline test scores and the students' section. Observations are students. * p<0.10, ** p<0.05, *** p<0.01.*

Table 2: Classroom Observation Measures of Attentiveness

	Dependent Variable: Z-score				
	Pooled (1)	Task Completion (2)	Response to stimuli (3)	Physical signs (4)	Pooled (5)
Cognitive Practice	0.0940***	0.0971*	0.1363**	0.0452	
	(0.0340)	(0.0572)	(0.0623)	(0.0582)	
Math Practice					0.1174***
					(0.0393)
Games Practice					0.0703*
					(0.0394)
Observations	1206	1198	1197	1196	1206

*Notes: The Pooled measure is a simple average of the z-scores for the individual outcomes within the scale. The Physical Signs measure was reversed so that a larger number corresponds to a better outcome, as is true of the other two outcome measures. Classroom observers were blind to students' treatment status. Clustered standard errors in parentheses. * p<0.10, ** p<0.05, *** p<0.01.*

Table 3: Treatment Effects on Declines

	Dependent Variable: 1[Question correct] Test Subject				
	All	All	Math	Listening	Ravens
	(1)	(2)	(3)	(4)	(5)
Panel A: Pooled Treatment Arms					
Treat x Deciles 6-10	0.0129***				
	(0.0047)				
Treat x Deciles 2-5	0.0084*				
	(0.0049)				
Treat x Predicted decline		0.0925***	0.1055**	0.0674**	0.0977**
		(0.0289)	(0.0432)	(0.0335)	(0.0429)
Treat	-0.0027	-0.0050	-0.0088	-0.0013	-0.0050
	(0.0060)	(0.0093)	(0.0109)	(0.0088)	(0.0115)
Panel B: Disaggregated Treatment Arms					
Math arm x Deciles 6-10	0.0127**				
	(0.0055)				
Games arm x Deciles 6-10	0.0131**				
	(0.0054)				
Math arm x Predicted decline		0.0984***	0.0988**	0.0891**	0.1124**
		(0.0357)	(0.0498)	(0.0431)	(0.0549)
Games arm x Predicted decline		0.0869***	0.1132**	0.0455	0.0822
		(0.0297)	(0.0494)	(0.0440)	(0.0566)
Math arm	-0.0003	-0.0053	-0.0017	-0.0051	-0.0100
	(0.0068)	(0.0108)	(0.0129)	(0.0107)	(0.0132)
Games arm	-0.0051	-0.0048	-0.0162	0.0025	0.0001
	(0.0069)	(0.0109)	(0.0118)	(0.0099)	(0.0134)
p-value: Math decline = Games decline		0.7273	0.7707	0.2895	0.5824
Control Decline	0.12	0.12	0.18	0.06	0.03
Students	1632	1632	1632	1632	1632
Observations	329349	329349	200234	66932	62183

*Notes: Panel A estimates treatment effects for both treatments pooled relative to the control group. Panel B shows effects for the Math treatment and Games treatment arms (each relative to the control group) separately. Col. (1) corresponds to the specification in Equation 1. Col (2) corresponds to the specification in Equation 3. Treat is a binary indicator that equals 1 if the student was assigned to a treatment (either the Math or Games arm). "Predicted decline" is the amount of average decline in each quintile of the test location, relative to the first quintile of the test, within each given school. "Deciles 6-10" and "Deciles 2-5" are each binary indicators that equal one if the question appears in the second half of the test or in deciles 2-5 of the test, respectively. The omitted category are the questions in decile 1 (i.e. the beginning) of the test. Cols. (1) and (2) estimate treatment effects for all three tests pooled. Cols. (3), (4), and (5) show effects for the Math, Listening, and Ravens tests separately, respectively. The Coefficients in Cols. (3)-(5) are estimated from a single regression on all the data. "Control Decline" captures the average score in the first quintile of the test minus the fifth quintile of the test for students in the control group, controlling for question fixed effects. Question item order was randomized across students. All regressions contain question and test version fixed effects, and baseline controls. Standard errors clustered by student, and bootstrapped in columns (2)-(5). The dependent variable mean is 0.47 in the control group. * p<0.10, ** p<0.05, *** p<0.01.*

Table 4: Performance on School Exams

	Dependent Variable: 1[Question correct]			
	Test Subject			
	All	Math	English	Hindi
	(1)	(2)	(3)	(4)
Panel A: Pooled Treatment Arms				
Cognitive Practice	0.0818**	0.0725**	0.0859**	0.0947**
	(0.0319)	(0.0331)	(0.0382)	(0.0380)
Panel B: Disaggregated Treatment Arms				
Math Practice	0.0835**	0.0778**	0.0906**	0.0932**
	(0.0370)	(0.0386)	(0.0442)	(0.0437)
Games Practice	0.0801**	0.0671*	0.0811*	0.0962**
	(0.0366)	(0.0375)	(0.0441)	(0.0439)
Students	1487	1487	1487	1487
Observations	12928	4317	4295	4316

*Notes: This table reports treatment effects on exams administered by schools to students (pooling mid year and end of year exams). Observations are at the student-test level. The dependent variable is the student's z-score on the test. Treat denotes receiving any treatment, Math and Games denote the Math practice or Games practice treatments, respectively. Cols. (1)-(4) regress z-score on a dummy for pooled treatment, in Panel A, and on dummies for Math and Games Treatments, in Panel B. All regressions include class section (strata) fixed effects and baseline controls. Standard errors clustered by student. * p<0.10, ** p<0.05, *** p<0.01.*

Table 5: Persistence of Treatment Effects

	Dependent Variable: 1[Question correct]			
	Definition of Treat Variable			
	Treat (pooled)	Treat (pooled)	Math Arm	Games Arm
	(1)	(2)	(3)	(4)
Treat x Deciles 6-10	0.0143***			
	(0.0051)			
Treat x Deciles 6-10 x Follow-up	-0.0044			
	(0.0113)			
Treat x Predicted decline		0.0933***	0.1104***	0.0764**
		(0.0304)	(0.0341)	(0.0338)
Treat x Predicted decline x Follow-up		-0.0012	-0.0182	0.0155
		(0.0434)	(0.0512)	(0.0504)
F-test p-value: Sum of 2 coefficients = 0	0.3249	0.0325	0.0683	0.0628
Students	1632	1632	1085	1080
Observations	329349	329349	219341	217223

*Notes: Follow-up is a binary indicator that equals one if the test is a follow-up test, administered 3 months after the end of the intervention. Deciles 6-10 is a binary indicator that equals one if the question appears in the second half of the test. Predicted decline is the amount of average decline in each quintile of the test location, relative to the first quintile of the test, within each given school. Question item order was randomized across students. All regressions contain baseline controls, question fixed effects, and test version fixed effects. Standard errors are corrected to allow for clustering by student, and bootstrapped in columns (2)-(4). * p<0.10, ** p<0.05, *** p<0.01.*

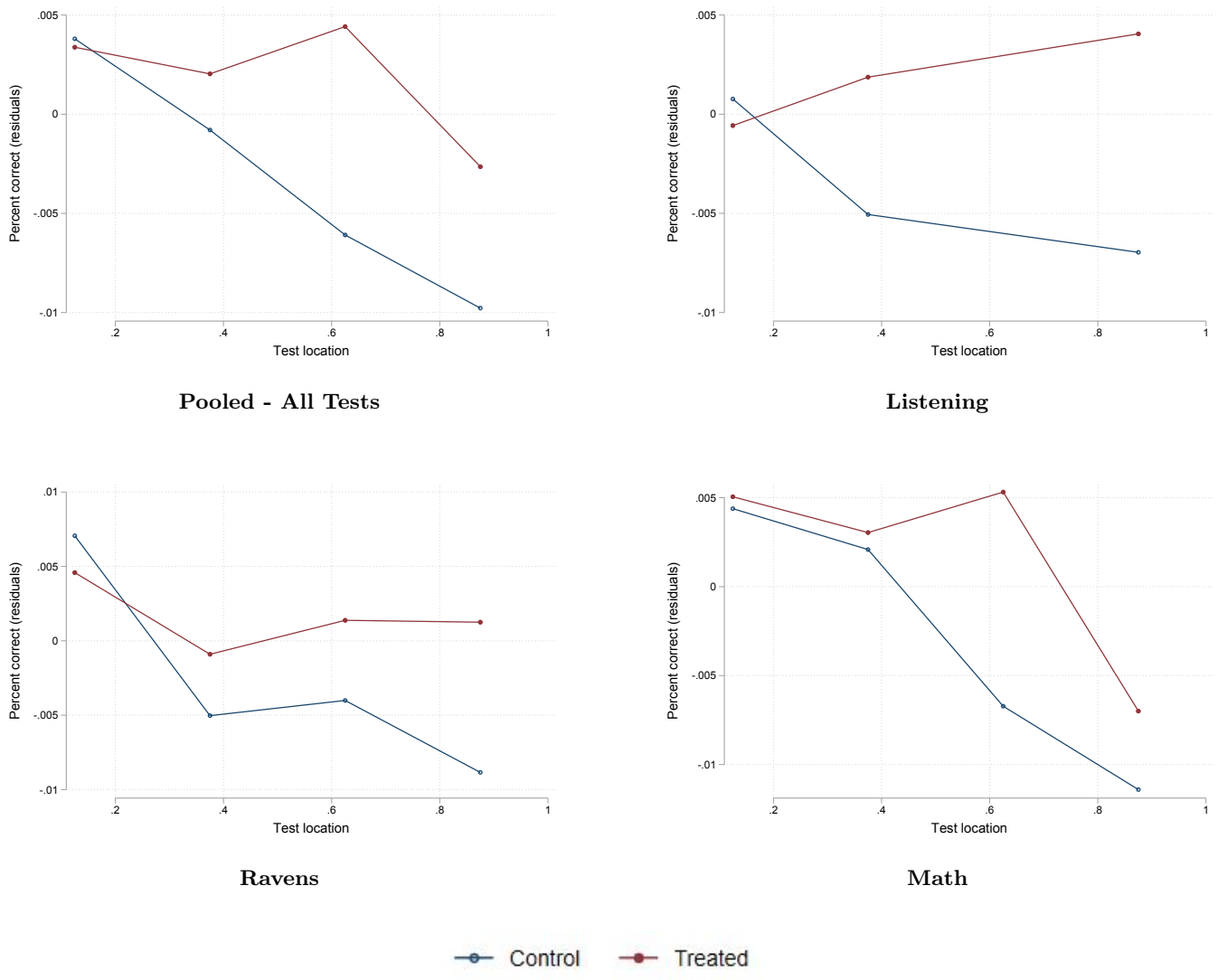
Table 6: Effect on Incentives on Test Performance

	(1)	(2)	(3)
Incentive	0.0916***	0.111**	0.168***
	(0.0330)	(0.0443)	(0.0532)
Treat		0.0257	0.0244
		(0.0214)	(0.0354)
Treat*Incentive		-0.0293	-0.0466
		(0.0440)	(0.0595)
Treat*Decile 2-5			-0.0194
			(0.0386)
Treat*Decile 6-10			0.0118
			(0.0250)
Decile 2-5*Incentive			-0.170***
			(0.0643)
Decile 6-10*Incentive			-0.0583*
			(0.0303)
Treat*Decile 2-5*Incentive			0.0680
			(0.0806)
Treat*Decile 6-10*Incentive			0.0152
			(0.0387)
R^2	0.224	0.225	0.226
Dependent variable mean	0.465	0.465	0.465
Number of students	704	704	704
Number of observations	11515	11515	11515

*Notes: This table reports the effect of offering students an incentive for their performance on the test. "Incentive" is a binary indicator that equals 1 if the student was provided a toy if they reached a certain score on the exam. "Treat" is a binary indicator that equals 1 if the student was assigned to a treatment (either the Math or Games arm). "Deciles 6-10" and "Deciles 2-5" are each binary indicators that equal one if the question appears in the second half of the test or in deciles 2-5 of the test, respectively. The omitted category are the questions in decile 1 (i.e. the beginning) of the test. Question item order was randomized across students. All regressions contain question and test version fixed effects, and baseline controls. Standard errors clustered by student. The dependent variable mean is 0.47 in the control group. * p<0.10, ** p<0.05, *** p<0.01.*

A Supplementary Tables and Figures

Figure A.1: Exam Performance by Treatment Status



Notes: This figure presents binscatters of the fraction of students answering each question correctly over time on each exam. The figures are residualized using the controls in Equation 1.

- Panel (a) shows performance pooled across all three exams.
- Panel (b) shows performance on the Listening test.
- Panel (c) shows performance on the Ravens Matrices test.
- Panel (d) shows performance on the Math test.

Table A.1: Baseline Balance

	(1) Control		(2) Pooled Treatments		(3) p-value	(4) Math Arm		(5) Games Arm		(6) p-value	(7) p-value	(8) p-value
	N	Mean/SE	N	Mean/SE	(1)=(2)	N	Mean/SE	N	Mean/SE	(1)=(4)	(1)=(5)	(4)=(5)
Baseline Listening (mean)	500	0.560	1029	0.548	0.58	521	0.555	508	0.541	0.85	0.44	0.55
		(0.385)		(0.386)			(0.386)		(0.386)			
Baseline Math (mean)	494	0.398	997	0.412	0.23	503	0.413	494	0.411	0.26	0.35	0.86
		(0.219)		(0.214)			(0.210)		(0.219)			
Baseline Ravens Matrices (mean)	492	0.366	1006	0.370	0.80	512	0.380	494	0.359	0.41	0.69	0.22
		(0.263)		(0.267)			(0.270)		(0.265)			
Baseline Listening (decline)	488	-0.001	1003	-0.019	0.47	507	-0.013	496	-0.025	0.67	0.41	0.67
		(0.433)		(0.457)			(0.445)		(0.469)			
Baseline Math (decline)	494	-0.071	997	-0.064	0.74	503	-0.058	494	-0.070	0.60	0.96	0.64
		(0.369)		(0.381)			(0.377)		(0.385)			
Baseline Ravens Matrices (decline)	461	-0.071	953	-0.028	0.26	490	-0.031	463	-0.024	0.27	0.32	0.89
		(0.534)		(0.722)			(0.571)		(0.854)			
Grade	528	2.737	1075	2.695	0.59	544	2.665	531	2.725	0.42	0.90	0.50
		(1.468)		(1.454)			(1.451)		(1.457)			
Baseline Stratification	519	4.355	1057	4.384	0.86	532	4.402	525	4.366	0.81	0.96	0.85
		(3.240)		(3.230)			(3.223)		(3.240)			
Income	262	23704.794	528	19176.610	0.08	263	18137.669	265	20207.709	0.11	0.33	0.10
		(55318.885)		(14575.256)			(11936.286)		(16749.132)			
Gender	385	0.384	774	0.349	0.24	385	0.343	389	0.355	0.23	0.39	0.73
		(0.487)		(0.477)			(0.475)		(0.479)			

*Notes: This table presents summary statistics for student baseline covariates by treatment group. Columns (1), (2), (4) and (5) present the sample size and mean for each covariate by treatment status. Columns (3) and (6)-(8) present p-values for the test of equality of means. * p<0.10, ** p<0.05, *** p<0.01.*

Table A.2: Attrition

	Pooled Treatments			Separated Treatments				
	Control	Treatment	p-value 1 = 2	Math Arm	Games Arm	p-value 1 = 4	p-value 1 = 5	p-value 4 = 5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: School Administered Exams								
Pooled	0.8870	0.8901	0.83	0.8929	0.8873	0.72	0.98	0.74
	(0.012)	(0.008)		(0.011)	(0.012)			
Math	0.8870	0.8901	0.83	0.8929	0.8873	0.72	0.98	0.74
	(0.012)	(0.008)		(0.011)	(0.012)			
Hindi	0.8870	0.8901	0.83	0.8929	0.8873	0.72	0.98	0.74
	(0.012)	(0.008)		(0.011)	(0.012)			
English	0.8870	0.8901	0.83	0.8929	0.8873	0.72	0.98	0.74
	(0.012)	(0.008)		(0.011)	(0.012)			
Panel B: Psychology and Classroom Measures								
Pooled	1.000	1.000	1.00	1.000	1.000	1.00	1.00	1.00
	(0.000)	(0.000)		(0.000)	(0.000)			
COS	0.9969	0.9962	0.80	0.9939	0.9985	0.43	0.56	0.18
	(0.002)	(0.002)		(0.003)	(0.002)			
SART	0.9089	0.9377	0.05	0.9289	0.9469	0.26	0.02	0.25
	(0.013)	(0.008)		(0.012)	(0.010)			
Panel C: Experimental Exams: Listening, Ravens Matrices, and Math								
Pooled	0.9708	0.9777	0.32	0.9778	0.9776	0.39	0.41	0.98
	(0.006)	(0.004)		(0.005)	(0.005)			
Math	0.9615	0.9744	0.09	0.9752	0.9736	0.13	0.18	0.85
	(0.007)	(0.004)		(0.006)	(0.006)			
Listening	0.9668	0.9692	0.76	0.9687	0.9697	0.84	0.75	0.91
	(0.007)	(0.004)		(0.006)	(0.006)			
Ravens	0.9602	0.9698	0.23	0.9687	0.9710	0.37	0.25	0.79
	(0.007)	(0.004)		(0.006)	(0.006)			

*Notes: This table presents the extent of attrition by treatment and test. The outcome is whether we observe at least one (non baseline) test each year. Panel A provides data for the school administered end of term exams. Panel B is for the psychological and classroom measures of attention (COS and SART). Panel C presents the results for the listening, ravens and math tests. Columns (1), (2), (4) and (5) present the percent of students for whom we have the respective exam. Columns (3) and (6)-(8) test for whether attrition is differential by treatment. * p<0.10, ** p<0.05, *** p<0.01.*

Table A.3: Test completion

	Math	Listening	Ravens
% attempted	0.773	0.996	0.992
% skipped	0.151	0.001	0.005
% of students completing last question item	0.793	0.996	0.983
Avg last question completed location	0.933	0.997	0.993

Notes: This table presents information about how much of the given test students completed. % attempted is the percentage of individual question items students did not leave blank. % skipped is the percent of questions in which students left a question blank but answered at least one subsequent question. % of students completing last question captures the percent of students who provided an answer on the last question of the exam, proxying for "finishing" the exam. Avg. last question completed location captures the average location of the last question item a student completed on the test as a percent of the total test. The listening and ravens tests are multiple choice tests and the math exam is free response.

Table A.5: Effect on Grit

	(1)	(2)	(3)	(4)	(5)
Treat*Decile 2-5	0.00704	0.00655	0.00641	0.00687	0.00663
	(0.00447)	(0.00447)	(0.00448)	(0.00446)	(0.00448)
Treat*Decile 6-10	0.0104**	0.00986**	0.00972**	0.0102**	0.00996**
	(0.00437)	(0.00436)	(0.00437)	(0.00437)	(0.00438)
Post hard question		-0.00240	-0.00200	0.000781	0.00161
		(0.00409)	(0.00392)	(0.00467)	(0.00414)
Treat*Post hard question		-0.00240	-0.00200	0.000781	0.00161
		0.00754	0.00618	0.00770	0.00706
		(0.00469)	(0.00447)	(0.00540)	(0.00477)
R^2	0.389	0.389	0.389	0.389	0.389
Hard question difficulty		Top 20%	Top 20%	Top 10%	Top 10%
Number of questions after		1	2	1	2
Number of students	1662	1662	1662	1662	1662
Number of observations	395094	395094	395094	395094	395094

*Notes: This table reports students performance after a very difficult question as a test of students "grit". "Treat" is a binary indicator that equals 1 if the student was assigned to a treatment (either the Math or Games arm). "Deciles 6-10" and "Deciles 2-5" are each binary indicators that equal one if the question appears in the second half of the test or in deciles 2-5 of the test, respectively. The omitted category are the questions in decile 1 (i.e. the beginning) of the test. "Post hard question" is a dummy for if the question comes after a difficult question. Question item order was randomized across students. All regressions contain question and test version fixed effects, and baseline controls. Standard errors clustered by student. The dependent variable mean is 0.47 in the control group. * p<0.10, ** p<0.05, *** p<0.01.*

Table A.7: Treatment Effect on Difficult Questions at Beginning of Test

	(1)	(2)	(3)	(4)
Treat	-0.00134	-0.00207	-0.00478	-0.00517
	(0.00913)	(0.00970)	(0.00716)	(0.00698)
Treat*Hard	-0.00114	0.000370	0.0181	0.0237**
	(0.0116)	(0.0109)	(0.0112)	(0.0107)
Question difficulty (above)	50%	30%	20%	10%
R^2	0.510	0.510	0.510	0.510
Test of treat + hard_treat (p-value)	0.752	0.755	0.127	0.0310
Number of students	1626	1626	1626	1626
Number of observations	30541	30541	30541	30541

*Notes: This table reports students performance after on difficult questions that appear in the first decile of the test. "Treat" is a binary indicator that equals 1 if the student was assigned to a treatment (either the Math or Games arm). "Hard" is a binary indicator for whether it is an above average difficulty question. All regressions contain question and test version fixed effects, and baseline controls. Standard errors clustered by student. The dependent variable mean is 0.47 in the control group. * p<0.10, ** p<0.05, *** p<0.01.*

Table A.9: Treatment Effect on Difficult Questions at Beginning of Test

Test	Length (minutes)	Baseline	Midline/Endline Date
Math	30	Yes	Dec 2017; Feb, Apr and Dec 2018; Feb and Apr 2019
Listening	12-15	Yes	Dec 2017; Feb, Apr and Dec 2018; Feb and Apr 2019
Ravens	15-20	Yes	Dec 2017; Apr and Dec 2018; Feb and Apr 2019
SART	8	No	Dec 2017; Apr and Dec 2018; Feb and Apr 2019
Symbol matching	15	Yes	Dec 2017; Feb, Apr and Dec 2018; Feb and Apr 2019

Notes: This table reports the length and frequency of tests administered by the research team.