WORKING PAPER · NO. 2022-73

Cognitive Endurance as Human Capital

Christina L. Brown, Supreet Kaur, Geeta Kingdon, and Heather Schofield

JUNE 2022

Becker Friedman Institute
FOR ECONOMICS AT UCHICAGO

5757 S. University Ave.
Chicago, IL 60637
Main: 773.702.5599
bfi.uchicago.edu

COGNITIVE ENDURANCE AS HUMAN CAPITAL

Christina L. Brown
Supreet Kaur
Geeta Kingdon
Heather Schofield

June 2022

We thank Ned Augenblick, David Autor, Stefano DellaVigna, David Deming, Claire Duquennois, Ernst Fehr, Caroline Hoxby, Xavier Gine, Lawrence Katz, Patrick Kline, Matthew Kraft, David Laibson, Sendhil Mullainathan, Imran Rasul, Jesse Rothstein, Andrei Shleifer, and Christopher Walters for helpful comments and discussions. We thank Pixatel for the use of the imagine Math software and the Institute for Financial Management and Research (IFMR) for operational support. We acknowledge generous funding 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 Jalnidh Kaur, Rolly Kapoor, Lubna Anantakrishnan, Simranjeet Dhir, Deepika Ghosh, Vatsala Raghuvanshi, Mudassir Shamsi, Alosias A., Erik Hausen, and Adrien Pawlik at the Behavioral Development Lab for their support in implementing this study, Trapti Dwivedi, Anshu Gupta, Jayshree Krishnan, Sharmila Singh, and Vinod Yadav for their support in coordinating with partner schools, and Isadora Frankenthal, Medha Aurora, Joaquin Fuenzalida, Jed Silver, Letian Yin, and 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; AEA RCT registry #0002673.

Cognitive Endurance as Human Capital
Christina L. Brown, Supreet Kaur, Geeta Kingdon, and Heather Schofield
June 2022
JEL No. D90,I24,I25,O12

ABSTRACT

Schooling may build human capital not only by teaching academic skills, but by expanding the capacity for cognition itself. We focus specifically on cognitive endurance: the ability to sustain effortful mental activity over a continuous stretch of time. As motivation, we document that globally and in the US, the poor exhibit cognitive fatigue more quickly than the rich across field settings; they also attend schools that offer fewer opportunities to practice thinking for continuous stretches. Using a field experiment with 1,600 Indian primary school students, we randomly increase the amount of time students spend in sustained cognitive activity during the school day —using either math problems (mimicking good schooling) or non-academic games (providing a pure test of our mechanism). Each approach markedly improves cognitive endurance: students show 22% less decline in performance over time when engaged in intellectual activities—listening comprehension, academic problems, or IQ tests. They also exhibit increased attentiveness in the classroom and score higher on psychological measures of sustained attention. Moreover, each treatment improves students' school performance by 0.09 standard deviations. This indicates that the experience of effortful thinking itself—even when devoid of any subject content—increases the ability to accumulate traditional human capital. Finally, we complement these results with quasi-experimental variation indicating that an additional year of schooling improves cognitive endurance, but only in higher-quality schools. Our findings suggest that schooling disparities may further disadvantage poor children by hampering the development of a core mental capacity.

Christina L. Brown
University of Chicago
11323 Hiawatha Lane
Indian Head Park, IL60525
christinalbrown@gmail.com

Supreet Kaur
Department of Economics
University of California, Berkeley
Evans Hall
Berkeley, CA 94720
and NBER
supreet@berkeley.edu

Geeta Kingdon
UCL Institute of Education
20 Bedford Way
London, WC1H 0AL
United Kingdom
g.kingdon@ucl.ac.uk

Heather Schofield
University of Pennsylvania
Perelman School of Medicine and
Wharton Business School
Blockley Hall 11th floor
423 Guardian Drive
Philadelphia, PA 19104
heather.schofield@gmail.com

1 Introduction

A large body of work documents the far-reaching, persistent benefits of schooling (Mincer, 1984; Schultz, 1988; Becker, 2009; Goldin and Katz, 2010). While it's clear that schooling affects cognitive abilities, the pathways through which it does so are less well understood (Deming, 2021). Researchers have long recognized its role could go beyond just teaching academic content or skills (e.g. Dewey, 1938; Bowles and Gintis, 1976; Morrison et al., 2019; Heckman et al., 2010; Conti et al., 2010). Schooling may expand our underlying capacity for cognition itself—for example, through our ability to engage in effortful thinking. This constitutes a more expansive view of how education shapes general human capital.

In this paper, we focus on one specific feature of formal education: schooling 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. Using a field experiment with elementary school students, we test whether such intellectual practice can, in and of itself, expand a particular mental ability: cognitive endurance.

We use the term "cognitive endurance" to refer to the ability to sustain performance over time during a cognitively effortful task. The psychology literature on sustained attention emphasizes the importance of this capacity: productive activity often involves sustaining mental effort, for example, over many minutes during a school test or hours during a work shift (Chun et al., 2011). This literature also hypothesizes that cognitive endurance could be improved through practice—raising the possibility that schooling could play a role in its development—but evidence remains limited (e.g., Rapport et al., 2013).

To motivate these ideas, we begin by examining a key prediction of limited cognitive endurance: when a person is engaged in an intellectually effortful task, their performance will decline over time. Such cognitive fatigue effects have been documented in myriad field settings, from paramedics at work to students taking academic tests (Brachet et al., 2012; Balart et al., 2018). We explore this prediction in PISA and TIMSS, two prominent global academic achievement tests. We examine how likely students are to get a randomly placed question correct when it appears earlier in the test (when they are still fresh) versus later in the test (when there is more scope for cognitive fatigue). Across academic subjects and geographic samples, we find a pattern consistent with prior studies: students are more likely to get a given question wrong if they encounter it later in the test. Moreover, we augment prior work by documenting that the rate of performance decline is considerably more severe for students from more disadvantaged backgrounds. Children in poor countries show three times the rate of decline as those in rich countries in TIMSS; similarly, in the US, Black and Hispanic students show 72% steeper decline over time than White students. In addition, this decline heterogeneity correlates with pedagogical differences across schools. Both globally and in the US, during the school day, more advantaged students spend more time in effortful thinking on their own—via independent, focused practice. These suggestive patterns motivate our approach to our experimental design.

We design a field experiment to increase the amount of time students spend engaged in periods of effortful thought. We conduct our study in a setting where time in focused cognitive activity is limited: low-income primary schools in India. Our sample is comprised of 1,636 students across six schools in grades 1-5. We randomize students to receive either continuous stretches of cognitive practice, or a control class period. We use two sub-treatments to deliver two distinct types of content. In the first sub-treatment (Math), students practice math problems. This mimics what good schooling does: intellectually effortful activity within the context of academic work. However, under our hypothesis, practicing any cognitively challenging task should improve endurance, regardless of whether students learn anything from it. Consequently, in our second sub-treatment (Games), students play cognitively demanding games, such as mazes and tangrams. The games contain no direct academic content, such as numbers or words, providing a pure test of our mechanism. We deliver each sub-treatment on simple tablet applications—enabling students to receive content appropriate to their skill level and promoting engagement—but without any animations or advanced graphics, so that students must deliberately exert cognitive focus.[1]

The control group receives a status-quo math "study hall" period. As is standard in this setting, control students copy a small number of math problems from the chalkboard and spend the remainder of the period as they would like. 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 sub-treatments. The experiment is implemented during a study hall or elective period 1-3 times per week between August and January, with practice sessions typically lasting 20 minutes at a time. In total, treated students receive 10-20 hours of additional cognitive practice.

We use this design to examine two sets of outcomes. First, we measure effects on school performance in academic subjects unrelated to the treatment content. This provides a positive test for changes in core mental ability, via aggregate impacts on a meaningful field measure—but does not shed light on specific mechanisms. Second, we develop tests to investigate changes in a particular mental capacity: cognitive endurance. To help clarify mechanisms, we also assess alternate channels, such as confidence, motivation, or working memory.

We examine students' academic performance in school, in the core subjects of Hindi, English, and math. Since neither treatment provides any practice in Hindi or English, this offers a test for generalized impacts. On average, cognitive practice improves student grades by 0.099 standard deviations (SD) in Hindi ($p = 0.012$), 0.092 SD in English ($p = 0.024$), and 0.085 SD in math ($p = 0.025$). Moreover, in each academic subject, the impacts are similar in magnitude and statistically significant across both the Math and the Games sub-treatments. This indicates that simply spending time in cognitive activity—*without learning any subject content*—improves the capacity to do well in school. Moreover, such thinking need not even be academic in nature: even students who receive

[1]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, with no writing or numerical content. In each arm, the tablet software provides no instruction, only the practice of problems or games. Note that all our primary outcomes are collected via paper and pencil tests.

the Games sub-treatment do substantially better in their classes.

To test for effects on cognitive endurance, we examine the rate of decline in performance over time in three distinct domains: listening comprehension, fluid intelligence (Ravens Matrices), and mathematics. We design the tests to minimize the scope for students to have "learned" relevant content during the treatment sessions—with the exception that this is, of course, not fully possible with the math test. For example, in the listening test, students listen to a series of short stories, each of which is followed by factual questions that check whether the student was attending to the story (e.g. "What color was the cat?"). This is completely unrelated to the treatments: there is no sense in which they enabled students to practice listening. In each domain, students take a 12-30 minute test with randomized question order and ample time to finish.

In each of the three domains, control students exhibit significant cognitive fatigue: the probability of getting a given question correct declines by 12% from the beginning to the end of the tests on average. In line with our predictions, the treatments improve cognitive endurance in each domain. On average, receiving cognitive practice mitigates performance decline in the second half of the test by 21.9% (p=0.006), with similar average effects across the Math arm (21.9%, p=0.018) and Games arm (22.0%, p=0.015). The improvements in cognitive endurance persist 3-5 months after the end of the intervention—after students return from the end of year vacation.

In addition, consistent with a lack of direct learning, the treatments have little impact at the beginning of the tests (i.e. the first quintile).[2] Rather, treatment effects only emerge later in the test, when there is more scope for cognitive fatigue. This helps distinguish our effects from confounding mechanisms that would raise performance throughout the test—for example, increased confidence, a desire to try harder, or improved working memory. If such channels were driving the results, students should also perform at least somewhat better early in the tests as well.

We supplement these findings on cognitive endurance with two additional tests from the psychology literature. First, we examine traditional laboratory measures of attentional capacity. This includes the canonical measure of the ability to sustain attention during a task, the Sustained Attention to Response Task (SART), which captures focus via accuracy and reaction times to stimuli. Cognitive practice improves performance on these traditional measures by 0.054 SD in the Math arm ($p = 0.023$) and 0.043 SD in the Games arm ($p = 0.071$).

Second, we find evidence for increased attentiveness in the classroom, adapting a diagnostic teacher rating scale—rated by observers that are blind to treatment status. We see improvements in this index of 0.110 SD in the Math arm ($p = 0.003$) and 0.058 SD in the Games arm ($p = 0.117$). These results suggest that cognitive practice bolsters students' ability to maintain attention in class.

Are these effects driven by cognitive improvements, or do they reflect general perseverance or motivation? Because psychologists consider such channels to be inherently related, we do not draw a strong line between them (Chun et al., 2011; Mischel, 2014; Zelazo et al., 2016). However, to

[2]This is particularly true in the listening and Ravens Matrices tests. For the math test, as expected, there is some evidence for level effects, but the Games and Math arms each has positive treatment effects on decline rates.

probe the potential role of motivation, we cross-randomize the chance to earn toys for higher test scores for a subset of the declines tests. This sharply increases performance at the beginning of the test. In contrast, the incentives do not reduce the severity of performance declines—suggesting that an internal drive to do better is not sufficient to mitigate observed fatigue effects. Similarly, to examine perseverance, we exploit discouragement effects in the declines tests: when control students encounter a question they are likely to get wrong, they do worse on the next (randomly ordered) question. We find no evidence that treatment mitigates these discouragement effects. These tests may not capture all dimensions of motivation, perseverance, or grit; but this, along with the positive evidence on sustained attention, suggests a likely role for cognitive improvements. More broadly, while our design allows us to evaluate some prominent alternate channels, it does not preclude the possibility that our treatments may have benefits through other pathways. Rather, we view our findings as offering positive evidence for one specific benefit: improvements in cognitive endurance.

Finally, we complement our experimental evidence—which relies on an outside intervention—by examining whether the natural experience of schooling develops cognitive endurance. We use quasi-random variation in years of schooling, due to birthday cut-offs for school enrollment, to construct a suggestive test for the effect of an additional year of schooling on cognitive endurance. Using data on primary school students from Brown and Andrabi (2021), we document large performance declines in academic tests, and find that an additional year of school mitigates these declines. These effects are three times as large as that of our more limited experimental intervention. Moreover, they are concentrated in better quality schools and those that assign more independent practice in class. In contrast, among the worst quality schools, an additional year of school produces no discernible improvement in cognitive endurance.

We conclude with two examples indicating the broader relevance of cognitive endurance among adults. We document substantial performance declines among full-time data entry workers and among voters at the ballot box, with more severe declines among more disadvantaged populations. While only suggestive, these patterns provide impetus for additional work on cognitive endurance.

Our paper contributes to two sets of literatures. First, we advance the literature on cognitive fatigue effects, including decision fatigue. A large body of work documents performance declines in numerous field settings—for example, whether a paramedic saves a life, if a proposition becomes law, or how well a student does on a standardized test (e.g. Endo and Kogi, 1975; Levav et al., 2010; Danziger et al., 2011; Brachet et al., 2012; Augenblick and Nicholson, 2015; Sievertsen et al., 2016; Meuter and Lacherez, 2016; Warm et al., 2018; Balart et al., 2018; Borghans and Schils, 2015; Hirshleifer et al., 2019; Zamarro et al., 2019; Akyol et al., 2021).[3] We augment this work by documenting that performance declines are more severe among disadvantaged groups across a variety of settings, and this contributes to overall performance gaps between groups.[4] This suggests, for

[3]Balart et al. (2018), Borghans and Schils (2015), Zamarro et al. (2019), and Akyol et al. (2021) document declines in observational test data such as PISA. By replicating declines in our experiment—e.g., the listening test, where running out of time or test-taking strategies cannot drive results—we validate and bolster previous findings.

[4]While many studies in the education literature examine performance declines, particularly in PISA, there has

example, that test scores may not only reflect content knowledge, and longer tests may especially disadvantage lower-income populations. Moreover, we provide the first evidence that cognitive endurance is not fixed and can be improved through practice—advancing work in both economics and psychology.[5]

Second, this study furthers our understanding of how schooling builds general human capital. Research in economics, education, and psychology argues that schooling builds skills—both cognitive and non-cognitive—that go beyond academic learning, and these skills are consequential for socioeconomic gaps in performance (for reviews, see Bowles et al., 2001; Cunha et al., 2006; Zelazo et al., 2016; Morrison et al., 2019).[6] This argument is typically based, for example, on looking at the impacts of an additional year of schooling on wide-ranging, later-life outcomes. We make several contributions to this literature. First, we highlight a new skill that can be developed through schooling, and which we argue belongs in our conception of general human capital: cognitive endurance. Second, while existing studies document broad benefits of schooling, there has been less work unpacking the education black box: what specific features of education are relevant, and how do they engender particular skills? We empirically demonstrate one such pathway, from sustained effortful thinking to cognitive endurance. Third, our results suggest that worse schools are less likely to inculcate this capacity. This offers an additional channel through which educational disparities could handicap more disadvantaged children, widening achievement gaps. Finally, we document that just the practice of thinking itself equips students to broadly perform better in school—a novel finding with direct policy implications, irrespective of mechanism. Of course, schooling likely confers other important cognitive and non-cognitive abilities; tracing the pathways for these constitutes an interesting direction for additional work.[7]

---

been limited work on such heterogeneity. A notable exception is Borgonovi and Biecek (2016), who explore decline differences along various dimensions, including socio-economic status (SES) and gender. In addition, Borghans and Schils (2015) document that performance declines in PISA predict later life outcomes, such as employment status and health.

[5]There is a related psychology literature on sustained attention, defined as the ability to sustain cognitive thought towards a goal and measured through lab tasks, such as the SART game. Attempts to "train" sustained attention have not generally looked for or found "far transfer"—improvements outside of the task or game that was practiced—likely due to small sample sizes (e.g. 15-35 individuals per arm). We advance this literature by demonstrating far transfer. In addition, this literature has focused nearly exclusively on training populations with attention deficits, whereas we examine the broader idea of malleability in general populations. See Chun et al. (2011) for an excellent review of the psychology literature on attention and Rapport et al. (2013) for a meta-analysis of programs training attention. In addition, there is a broader literature on training other cognitive processes such as fluid intelligence and working memory. Evidence for "far transfer" of these skills is mixed (e.g. Sala et al., 2019; Diamond and Lee, 2011; Simons et al., 2016), with some recent notable successes (e.g. Berger et al., 2020). Like Berger et al. (2020), we augment this broader literature by identifying causal impacts in a naturalistic setting with a large sample size, with far transfer effects on policy relevant outcomes such as academic grades.

[6]Related work highlights the importance of non-academic skills—such as higher-order non-cognitive skills or cognitive skills—for productivity, underscoring that human capital is multi-faceted (e.g. Heckman et al., 2006; Almlund et al., 2011; Chetty et al., 2011; Heckman and Kautz, 2012; Borghans et al., 2014; Chen et al., 2017; Deming, 2017).

[7]Related studies use interventions in schools to improve skills like mindset, patience, grit, working memory, and intuitive math (Bettinger et al., 2018; Alan and Ertac, 2018; Alan et al., 2019; Berger et al., 2020; Dillon et al., 2017).

2 Background and Motivation

2.1 Definition of Cognitive Endurance

We define "cognitive endurance" as *the ability to sustain performance over time during an activity that requires effortful thinking.* Because individuals have a limited capacity to sustain such thinking for long periods, doing so leads to mental fatigue (Boksem et al., 2005). This suggests a key empirical implication: when a person is engaged in a task that requires intellectual resources, performance on that task will decline over time. Note that this definition does not inherently take a stance on the specific cognitive or psychological mechanism that produces performance declines. We probe potential mechanisms within the context of our field experiment below.

The cognitive psychology literature posits that the ability to sustain cognitive effort over time could potentially be "trained"—through the practice of effortful thinking for continuous stretches (Chun et al., 2011). The process of receiving an education inherently includes such practice: working independently on academic work, taking a test, reading a textbook, or concentrating on a lecture. Moreover, in the education literature, engaging in independent, focused academic work during the school day—which requires students to practice thinking effortfully on their own—is considered a pedagogical feature of better quality schools (Araujo et al., 2016). While speculative, these ideas raise the possibility that the experience of (good) schooling could play a role in the development of cognitive endurance.

2.2 Motivation: Performance Declines among Students

To motivate the empirical relevance of the above ideas, we begin with a set of suggestive examples using data from two prominent global academic achievement tests: TIMSS and PISA. TIMSS is administered to fourth graders in over 50 countries during the school day, and PISA to 15-year olds in over 60 countries. The tests cover multiple academic subjects, with a duration of roughly 30 minutes per subject, and are designed to give ample time for completion. For example, in TIMSS, only 3.2% of questions are skipped, and 4.5% of questions are not reached (Foy et al., 2011). In each exam, question order is block-randomized across students—within each subject in TIMSS and across four academic subject blocks in PISA.

In Figure I, we plot performance declines over time: the proportion of students in each sample who answer a given question correctly (y-axes) against the question's location in the test (x-axis). The TIMSS (PISA) plots control for question (question block) fixed effects. We graph declines in the global sample as well as US sample, separately for each test subject and by proxies for socioeconomic status (using separate y-axes).

In each of the ten plots, we document 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, in the global sample, the rate of performance decline is 9% in

TIMSS among poor countries, and 25% in PISA among lower socioeconomic status (SES) individuals. These patterns are similar when restricting to attempted questions only (Appendix Figure A.1).

Second, more disadvantaged students consistently display significantly more severe performance declines. For example, globally, students in poor countries exhibit 240% more decline than those in rich countries (Panels C-D). We see similar patterns by race and SES within the US (Panels A-B and E-G). For example, Black and Hispanic students exhibit 72% more decline than White students in TIMSS (Panels A-B).[8]

Many factors may contribute to systematic differences in cognitive endurance. Teacher time use data in TIMSS suggests one potential source of this correlation: differences in schooling environments. In schools serving more disadvantaged populations, students spend less time in focused independent academic work during the school day—affording less opportunity for effortful thinking on their own (Appendix Figure A.2, Panels A-B). Relative to their richer counterparts, students in poor countries spend 40% less time in independent practice, and in the US, lower SES students spend 10% less time in such practice.[9] In addition, students who receive more focused independent practice in school exhibit less steep performance declines in the TIMSS exams, even when controlling for income (Appendix Table A.1, Cols. 2-3).

The above findings are simply suggestive patterns and correlations. They are intended only to offer motivational support for the empirical relevance of cognitive endurance, and for the possibility that it may be malleable. In our experiment, we construct a more interpretable measure of cognitive endurance. We design an intervention to improve this capacity by increasing the amount of time students spend in cognitively challenging activity on their own. This design choice is informed by both the psychology literature and the motivational patterns above.

3 Experimental Design

3.1 Context

We select a school setting where the time spent in focused cognitive activity is limited: low-income primary schools in India. In this setting, as is common in many developing countries, teachers predominately use rote memorization and recitation during the school day (World Bank, 2004). Classrooms are crowded, with frequent disruptions from environmental noise and other students. Students within a class also vary widely by achievement level: half the students in a classroom may be below grade level (e.g. ASER, 2019; Muralidharan et al., 2019). Consequently,

[8]However, more vs. less advantaged students perform very differently at the start of the tests (see the two y-axes in each plot), complicating the interpretation of performance decline differences in these motivating examples. Appendix Figure A.1 displays the plots on a single y-axis. In addition, note that these are crude proxies for relative advantage, so there is measurement error in how they map to socioeconomic status.

[9]Moreover, environmental conditions in lower-income schools may make it more difficult to engage in focused concentration (Burke et al., 2011; Kraft and Monti-Nussbaum, 2021; Figlio, 2007). For example, lower SES students attend schools with considerably more disruptions during class (Appendix Figure A.2, Panels C-D).

when teachers do assign independent practice—typically by writing 2-5 problems on the chalkboard and asking students to complete them in their notebooks—many students cannot even attempt the problems and end up disrupting other students. Outside of school, students spend little time on homework or other cognitively challenging tasks. Consequently, they seldom have the opportunity to engage in focused cognitive activity for sustained periods either inside or outside the classroom.

We conduct our experiment in six Indian private primary schools in the region of Lucknow, India. The schools cover a mix of urban and rural areas and serve students from largely low-income households. Our sample is comprised of 1,636 students in grades 1-5. Appendix Figure A.3 provides example pictures of the classroom environment in these schools.

3.2 Treatments

We design an intervention to increase the amount of time students spend in effortful thinking for sustained periods. We accomplish this by having students solve cognitively challenging problems on their own for 20-minute sessions during the school day. To test our hypothesized mechanism, we use two different approaches for this cognitive practice—academic and non-academic—and compare this with a control arm:

1) **Treatment: Cognitive Practice.** Students solve cognitively challenging problems. There are two variants of the treatment which vary the type of problems students work on:
    a) ***Math***: Students solve math problems.
    b) ***Games***: Students play cognitively demanding games that are free of academic content, such as mazes and tangrams.
2) **Control: Study Hall.** Students attend a status-quo study hall period, with limited cognitive practice.

For the Cognitive Practice sub-treatments to be effective, they each require activity that is cognitively demanding so that concentration is taxing and requires the deliberate exertion of attention, but is also sufficiently engaging to retain student participation for a continuous stretch of time (e.g. a 20-minute session). Moreover, the activity must be feasible in classrooms with starkly different achievement levels across students (i.e. with many students behind grade level). To achieve this balance, we deliver the treatment content using simple tablet apps.

Math Practice. The Math sub-treatment mimics what good schooling does: focused cognitive practice within the context of academics. Students solve a series of math problems on their own in each session. We use the imagine Math software, developed by Pixatel, a social enterprise. This software displays math practice problems via a simple interface, with no graphics, animations, or other visual features (see Appendix Figure A.4a). One problem appears on the screen at a time;

students are asked to solve the problem and then select the correct answer on the tablet.[10] Depending on the student's performance, subsequent problems become easier or more challenging. The software focuses exclusively on practice problems and does not provide any instruction. In addition, because the software is a general one, the problems do not necessarily correspond to those in students' regular class textbooks or tests. Consequently, we use the software as a tool to engage students in cognitive practice using material that is related to academics, without being tailored to specifically increase achievement in their regular school classes.

Games Practice. Under our hypothesis, practicing any intellectually demanding task should improve cognitive endurance—regardless of whether it is academic in nature. This motivates the design of the Games sub-treatment, which does not entail any direct academic practice. Students play intellectually challenging games, chosen so that they contain absolutely no academic content, such as numbers or letters—providing a more pure test of our mechanism.

We use simple games with limited animation downloaded from the Android app store. These should not be viewed as "fun" video games, but rather puzzles and problems, such as tangrams or maze puzzles, delivered through a bare-bones tablet interface (see Appendix Figure A.4b). The specific games were chosen to meet three criteria: i) they should involve no words or numbers, and be unrelated to test outcomes we would measure later (e.g. no games with sound or text were selected); ii) they should be dynamically adaptive to continue to challenge all students regardless of initial skill (so students do not get bored over time); iii) 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. Appendix Figure A.5 shows pictures of example classes implementing the Cognitive Practice sub-treatments.

Control. The control group receives a status-quo math "study hall" period. We follow the standard practice in this setting: the teacher writes a small number of math problems (i.e. 5 problems) on the chalkboard and then sits down to do her own work (e.g. grading or administrative work). Students can decide whether to attempt the questions and spend the remainder of the study hall session as they'd like. This leads to little effective time spent in cognitive practice: many students do not have the skill to attempt grade-level content and may not be engaged enough to do so.

In practice, treated students concentrate for about 20 minutes per program class period, compared to 0-10 minutes for students in the control group (reflecting wide heterogeneity across students). This results in 10-20 hours of cognitive practice in the treatment arms on average (see Section 3.3 for details).

Note that we do not view tablet-based training as necessary for our approach to be effective. Rather, in our particular context, piloting indicated that this was an effective way to retain engagement, while solving the practical challenge of heterogeneous ability. Consequently, we view this as

[10]Students were also provided paper and a pencil during this class in case they needed to work out the answer on scratch paper.

simply a convenient way to achieve our goal of engaging students in independent effortful thinking in this context. Note that this does not necessarily imply that playing any video games would achieve similar results: many such games are so engrossing that exerting attention is not very effortful—for example, attention may be captured by the appearance of constant stimuli via animations. In contrast, our treatment arms require students to deliberately exert effortful focus.

3.3 Implementation and Protocols

Students in grades 1-5 in the study schools were enrolled in the experiment. Each student was randomized into one of the three experimental arms for the duration of the study. Randomization was at the individual level, stratified by class section (i.e. classroom) and baseline math test scores.

The intervention was implemented during students' regular study hall or other elective periods, avoiding any crowd out of traditional academic teaching. At the start of each designated period, students in the class were split up and went to one of three classrooms based on their assigned treatment status. They returned to their normal classroom at the end of the elective period. In most schools, elective periods were about 30 minutes in length, so with transition times, the effective intervention time was roughly 20 minutes per session. Each school dedicated 1-3 elective periods per week from August to January for the intervention. Appendix Figure A.6 shows the timeline of the experiment.

The number of sessions varied across weeks based on other activities such as festivals, planned assemblies, or exams. They also varied across schools based on the intervention start date and the number of free elective periods available. We used staggered enrollment to recruit schools into the study. All six schools were onboarded with the intervention in place by September 2018. We completed endline testing across schools in the spring of 2019.[11]

At each school, we placed three study staff members who implemented the intervention, including overseeing activities during practice sessions. These staff members had the background one may expect of a teacher's assistant and were recruited with assistance from the schools; from the perspective of students, they resembled normal teachers. The study staff were randomly rotated across treatment arms each month to prevent any collinearity with treatment status. Having our own research staff undertake the implementation helped ensure study protocols were followed.

Across treatment arms, the amount of instruction provided during the practice sessions was minimal, and never covered any specific subject content (e.g. how to solve a particular math problem). When the experiment launched in each school, all students received a primer to learn how to start the tablets and bring up the relevant software for their respective group: the imagine Math software, the folder where games were stored, or the notepad typing application for the control group. This

[11] We began piloting in 2017. We enrolled three schools late in the academic year of 2017-18, with the remaining enrolled in the 2018-19 academic year. Since the former set of schools only had about a month of full practice time in the initial year, we continued their participation into the subsequent academic year, keeping treatment status for each student the same (and enrolling a new set of first graders to enter the study at that time).

took about 10 minutes per group. In addition, most students were familiar with computers since in 4 out of 6 schools, students attended a weekly computer lab elective class, and we chose software that was extremely simple to use. In the Games sub-treatment, students rotated across about four distinct games over the year. Each time a new game was introduced, the program staff would explain how to play that specific game, which typically took about 5 minutes given the simple nature of the games, with another 5-10 minutes of follow-up questions from students as they began using the game on the first day. Aside from this, there was no instruction across any of the experimental arms. At the start of each Math or Games treatment session, the program staff would announce which module in the imagine Math software or games folder to open, and only provided assistance if the tablet was malfunctioning or the student couldn't find the module, which was rare. If students asked for help with content in the Math or control arms, program staff were instructed to tell students to give their best guess.[12] Across all arms, students were free to stop working at any time, and put their heads down if they no longer wanted to engage in the content. After getting students settled, program staff typically did administrative paperwork tasks at a desk while students practiced—as is customary among teachers overseeing study hall periods in our setting. Supervisor spot checks verified compliance with these protocols.

To avoid feelings of unfairness, students in the control group were also allowed to use the tablets early in the year to practice typing and other simple activities, selected to avoid stretches of cognitive focus. Across all three groups, because the tablet activities were not very exciting (e.g. no animations or graphics), the novelty of the tablets wore off fairly quickly during the intervention. As a result of this and the exposure to tablets among all experimental arms, qualitative conversations suggest that students did not experience notable fairness concerns. In response to any parent inquiries, schools planned to explain the intervention as a pilot program on alternate education approaches—with different approaches being tried by lottery during the study year—with the plan that access would be equalized across all students the following year. In practice, however, we did not hear of any incidents of parent complaints or inquiries. Because the randomization assignment was overseen by research staff and children had assigned seats and tablets labeled with their name, it was not feasible for students to switch across treatment.

Finally, we endeavored to keep the regular school teachers blind to students' treatment status. Treatment assignment rosters were never shared with teachers. At the start of each practice session period, program staff collected all students from their classroom, walked them to the three rooms where the experiment activities were conducted, and returned them to their classroom after the end of the period. All students left and returned to their class section at the same time, and program classes were held in a different location in the school (usually a different floor). Because teachers typically remained in their own classroom or the teacher's lounge during the practice sessions, they would not have directly observed which students were in which group. In addition, students

[12] While this was necessary for the integrity of our design, this obviously handicapped the ability of the Math or control arms to produce improvements in math learning.

were assigned "labels" that were easy to remember but unrelated to the activities (i.e. the "green group" or "blue group")—helping the study staff organize the students, with the ancillary benefit of obscuring their assignment. This helps reduce concerns that teachers could systematically have treated students differentially in some way based on their knowledge of treatment status. Of course, it is certainly possible that some teachers may have learned the treatment status of some individual students.

3.4 Outcomes and Data

Our outcome measures are briefly summarized here and explained in more detail in the results sections. We examine impacts on two main outcomes: (i) Students' academic school performance; and (ii) Cognitive endurance as measured through performance declines in three domains: listening comprehension, Raven's Matrices (IQ), and math tests. We supplement this with additional measures of attention from the psychology literature: traditional laboratory measures of sustained attention (e.g. the Sustained Attention to Response Task) and measures of classroom attentiveness. All these outcomes were pre-registered.[13]

In addition, we undertake additional tests to assess two categories of alternate mechanisms. First, we explore whether the cognitive endurance effects should be understood as reflecting cognitive improvements, or could instead be driven by changes in motivation or perseverance. For example, to examine the potential role of motivation, we cross-randomize incentives (i.e. toys) for test performance on a subset of the declines tests. We use this to test whether increased motivation to perform better (due to the external incentives) mitigates performance declines. Second, in Section 6.4 below, we use our design and results to evaluate potential confounds that operate outside of cognitive endurance—such as confidence, desire to do well in school, or alternate cognitive mechanisms such as working memory.

School grades (outcome i) were provided by all schools for both mid-year (December) and at the end of the year (March) grades and were not cumulative. The research team had no engagement with students' regular academic classes or assessments. For a subset of schools, we also have grades from the year before the intervention, which we use as baseline measures. Unfortunately, we were not able to collect follow-up administrative data from the schools for the subsequent year due to disruptions from the Covid-19 pandemic, which led schools to close for an extended period, and one school has since shut down.

[13]The study is registered on the AEA RCT registry (#0002673). We pre-registered the performance declines tests (in the subjects of listening, Ravens Matrices, and math), and the psychology measures of SART and Symbol Matching. For the declines test, we pre-registered the fact that we would measure effects on both declines (i.e. slopes) as well as average performance, but did not include a pre-analysis plan with specific regression specifications. In addition, we pre-registered that we would examine school performance and the classroom observations, with the caveat that our ability to look at these would be subject to agreements from the schools to collect this data, which had not yet been obtained at the time of the pre-registry. The delay in finalizing these details delayed the registry of the study. We ultimately registered the study in January 2019—after the trial began, but before the February 2019 endline tests were conducted, and well before the post-treatment data had been entered and compiled across schools.

The performance decline tests (outcome ii) were administered during the school day at four times: Baseline (August), Midline (December), Endline (February), and Follow-up 3-5 months after the end of the intervention (April-June). Note that because the intervention only ran until January, the midline and endline tests are close together in time. Students in each class section cohort were tested together, so that each test batch had students from across the treatment arms. The tests were unannounced, so that students did not know they would be taking a given test the next day. Staff conducted make-up tests throughout the following month for students who were absent on the day of the exam. Certain tests were randomly not administered in some rounds due to time constraints.

Appendix Table A.2 provides summary statistics and tests of baseline balance. We find no significant differences in demographic characteristics or baseline performance across experimental arms. Of 44 tests, none are statistically significant.[14] In addition, attrition was relatively low: averaging 3% for the declines tests and 10% for school grades (due to either students leaving the school before end-of-year exams or missing data in the school records). Appendix Table A.4 verifies no differential attrition by treatment status: among the 33 p-values in Cols. 6-8, only one is significant.

4 Results I: School Performance

If Cognitive Practice meaningfully improves general mental capacity, this should be reflected in how well students do in school. In our setting, the core academic subjects are Hindi, English, and math; these are the only subjects taken by all students in our sample. We focus especially on performance in Hindi and English, since these are wholly unrelated to the content practiced in the treatments.[15]

Results are presented in Table I. Overall, receiving Cognitive Practice improves school performance by 0.0897 SD (Panel A, Col. 1, $p = 0.010$). These sizable gains are present even in the non-math subjects, with impacts of 0.099 SD in Hindi (Col. 3, $p = 0.012$) and 0.092 SD in English (Col. 4, $p = 0.024$).

Panel B disaggregates these results by sub-treatment arm. Each of the Math and Games arms has significant effects on each of the three core academic subjects, ranging from 0.080 SD to 0.102 SD. Within each subject, the magnitude of effects across the two arms are similar and statistically indistinguishable from each other. However, given the size of the confidence intervals, there is a possibility that one sub-treatment could have larger impacts than the other. For example, for math grades, the coefficient on Math Practice is 12% larger than that on Games Practice, but we cannot

[14] We have baseline grades for only two out of six schools, with data collection from schools disrupted by the COVID pandemic. Appendix Table A.3 tests for baseline balance in this limited sub-sample. Comparing Cognitive Practice to the control group (Cols. 1-3), baseline grades are balanced for each test subject (Hindi, English, and math). When decomposing by treatment arm (Cols. 4-7), students in the Math (Games) arm tend to have lower (higher) baseline scores than the control group, so that potential imbalance goes in opposite directions for each sub-treatment arm.

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

reject that they are the same.[16]

In addition, we find evidence for impacts in both the mid-year grades assessed in December (0.074SD, $p = 0.046$) as well as end-of year grades (assessed in February) (0.094 SD, $p = 0.009$) (Appendix Table A.5). While the estimated effect sizes grow over time, they are statistically indistinguishable ($p = 0.421$). Finally, we do not detect heterogeneity in treatment effects by covariates such as student grade, baseline average score; the interactions are typically small and insignificant (Appendix Table A.6, Panel A).

The magnitude of the effects in Table I is substantial, especially when compared to prominent interventions in the education literature. For example, Project Star reduced class sizes in the US for an entire year, and increased academic gains by 0.12 SD among kindergarten through grade 3 students (Krueger and Whitmore, 2001).[17] Tracking students by ability in Kenya or remedial education with an additional teacher in India each had impacts of about 0.14 SD (Duflo et al., 2011; Banerjee et al., 2007). Each of these three 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 10-20 hours of cognitive practice, without direct academic learning.

As a whole, the results in Table I indicate that simply spending time in effortful thinking—*without learning any subject content*—improves the ability to do well in school. Moreover, such thinking need not even be academic in nature: even the students who receive Games Practice do substantially better in their academic classes.

These findings offer positive evidence for the idea that a particular feature of formal education could improve general mental ability, as proxied by improvements in academic grades. Measuring the overall impact on school performance has the benefit that it is a meaningful field outcome. However, it does not shed light on what specific mental changes are induced by cognitive practice—including, for example, whether they are cognitive or non-cognitive in nature. In the next section, we complement these findings with a tighter mechanism test: we examine the impact of cognitive practice on one particular mental capacity, cognitive endurance.

5 Results II: Cognitive Endurance

5.1 Measuring Performance Declines

We test for improvements in cognitive endurance by examining whether the treatment mitigates the severity of performance declines during cognitively challenging activity. We construct tests in three diverse domains:

[16] In addition, as discussed above, the Math sub-treatment practice problems were not tailored to the content taught and tested in students' regular math class in school. This mitigates the scope for the Math sub-treatment to have a differentially larger impact on math grades relative to the Games sub-treatment.

[17] Results in Krueger and Whitmore (2001) were presented in percentile point changes. We assume a normal distribution to adjust to a standard deviation metric.

(1) *Listening:* This task measures attentiveness in listening. Using headphones, each student listens to a pre-recorded set of short, simple stories. After each story, the student is asked a series of simple factual questions about the content of the story, for example, "What color was the dolphin?" Each question is slowly dictated along with each possible answer choice twice. Students must continually attend to the voice on the audio, so that they do not miss the content of the question or the description of the four possible answer choices per question—requiring concentrated focus. After a fixed allotment of time to answer the question, the audio asks students to turn to the next page in their answer packet, and proceeds to ask the next question.[18] To avoid concerns about literacy, no text is written anywhere on students' answer sheets; instead the multiple-choice answer options are represented by simple pictures.

(2) *Raven's 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 viewed as capturing "fluid intelligence" and is commonly used as an IQ test. Students take a shortened paper-and-pencil version of this test, adapted to be appropriate for each grade level.[19]

(3) *Math:* A standard test of math problems. These include a mix of both remedial and more grade-appropriate questions and do not correspond to the specific content or format of the questions in the Math sub-treatment software. Consequently, this is not a diagnostic test to measure learning gains but rather a more general test requiring students to solve math problems.

To enable clean identification of performance declines, we design the tests to include three features. First, we minimize the scope for treated students to have directly learned content that would help them perform better on the tests. Specifically, note that for each of these domains, the content is wholly unrelated to at least one of the two sub-treatment arms. For example, neither arm could have taught students listening comprehension. The Math arm content does not involve any spatial reasoning, and so could not have taught students how to solve Raven's questions. Similarly, the Games arm involved no numbers or math content.[20] Minimizing the scope for direct learning helps prevent level effects at the beginning of the tests; if treated and control students do roughly equally well at the start of the tests, this simplifies the interpretation of differences in subsequent performance declines.

[18] When examining endurance effects, we examine performance across questions within a given story passage. Students receive a mental reset at the start of the next story, similar to the positive effects of brief breaks found in repeated vigilance tasks (Mijović et al., 2015; Finkbeiner et al., 2016).

[19] 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.

[20] However, since the Control and Math arms both practiced math problems, the students in the Games arm could experience negative level effects on the math test. In addition, it is possible that the Games sub-treatment, through some games like tangrams, could have provided training that is relevant for the Raven's Matrices test.

Second, for each test, we randomize the order in which questions appear across students. Specifically, for each test, we randomize question order across different test packet versions. We then randomize the test packet version across students.

Third, we ensure that students have sufficient time to finish the tests without time pressure. Moreover, the fixed timing of the listening test precludes the possibility that students move at their own pace and ensures all students reach the end of the exam. Consistent with this, nearly all students respond to questions near or at the very end of the exam: the last question with a response was on average 97%, 92%, and 98% of the way through the exam for the listening, math, and Raven's Matrices tests, respectively, and these rates are balanced across groups (Appendix Tables A.7 and A.8). Because students may not attempt questions they do not know how to answer, such as in the math test which included only free-response questions, this is likely an under-estimate of completion rates.[21] Consequently, declines over time are not confounded by differential non-completion. We also examine robustness to restricting the data to only attempted questions.

Note that while the above design features enable clean identification of performance declines, they necessarily handicap the potential magnitude of treatment effects on overall performance. For example, if treated students are less cognitively fatigued as they proceed through a test, they would be more likely to finish (and therefore perform well on) tests that are longer or where time is limited—a realistic feature of typical tests in school environments. Similarly, by picking content where there is no scope for learning—for example, by testing listening rather than Hindi knowledge from school—we shut down a myriad of ways in which cognitive endurance could increase performance, such as the ability to work through a long textbook chapter. Such forces would lead to positive level effects even at the start of the tests, making it more difficult to interpret performance declines. Consequently, we use this measure to develop a clean mechanism test, and rely on complementary measures (e.g. classroom attentiveness or school performance) to examine other dimensions of benefits and the aggregate impacts of Cognitive Practice.

Each set of tests is adapted to grade level. This includes differences in difficulty and length by grade, with a minimum of 12 minutes and a maximum of 30 minutes per test. Appendix Table A.9 summarizes the length of each test. Students take only one test per day, so they are cognitively fresh at the start of each test. All tests are conducted during the school day and are presented to students as regular school tests. Grading is done by research staff that are blind to treatment status. An example page from each of the assessments is shown in Appendix Figure A.7. Appendix Figure A.8 and Appendix Table A.10 verify the balance in test version and question difficulty by treatment status.

[21] For example, conditional on a final question on the math test being left blank, the fraction of times it is left blank when it appears earlier in the test (for other students) is on average 23%.

5.2 Performance Decline Patterns

We begin by visually examining performance declines over time. Figure II plots performance on each test over time—separately for the control and treatment groups. Panels A-C pool the Cognitive Practice arms together for ease of interpretation, while Panels D-F show each sub-treatment arm separately. In each panel, the x-axis is the percent location of the test, grouped in deciles (where 0.1 is the first decile of the test and 1 is the last decile), and the y-axis is the proportion of students who answer the question correctly. Each graph displays binscatter plots, with a fitted quadratic curve.[22] The data is residualized to remove question and test packet version fixed effects. Consequently, the plots can be interpreted as showing changes in average performance when the same question appears earlier versus later in the test.

The dashed grey line displays control group performance in each plot. Across tests, students are 12% less likely to get a question correct if it appears in the fifth quintile rather than the first quintile. In each domain, the control group declines by 18 p.p., 6 p.p. and 3 p.p. in the math, listening, and Raven's tests, respectively. Coincidentally, the average rate of decline on our experimental tests is similar to the decline rate on the TIMSS and PISA tests for low-income countries.

Consistent with our hypothesis, Cognitive Practice mitigates performance declines: on average, treated students decline less quickly over time than control students. Because test completion is high, these patterns are similar or stronger when restricting the data to only attempted questions (Appendix Figure A.9).

5.3 Empirical Estimation

To more formally examine treatment effects, we begin by estimating:

$$\begin{aligned} Correct_{ils} = \ & \beta_0 + \beta_1 CogPractice_s + \sum_{l=2}^{10} \lambda_l Decile_l + \beta_2 CogPractice_s * 1[2 \leq Decile_l \leq 5] \\ & + \beta_3 CogPractice_s * 1[6 \leq Decile_l \leq 10] + \beta_4 Baseline_s + \chi_{il} + \epsilon_{ijs} \end{aligned} \tag{1}$$

$Correct_{ils}$ is a binary variable that captures whether student s correctly answered question item i in location l. $CogPractice_s$ is a dummy that equals one if the student is assigned to one of the Cognitive Practice sub-treatments and zero if the student is in the control group. The λ_l are location (decile) fixed effects, which flexibly capture declines over time in the control group. χ_{il} is a vector of question fixed effects. We also control for the student's average baseline score, class section (strata) fixed effects, fixed effects for the version of the test packet taken by each student, and a linear control for the average fraction of students in the student's school who got question i correct (computed restricting to the control group only and excluding student s's own observation), which captures question difficulty.[23] In addition, we can run the above regression to separately estimate effects for

[22] Recall that the listening test typically had 3 question exercises per passage.

[23] When the baseline score is missing for a student, we code it as zero and include a dummy indicating the missing

each sub-treatment, replacing the $CogPractice_s$ dummy with two separate dummies for the Math and Games sub-treatments. Because tests for each subject were administered multiple times (e.g. December, February, and April), we pool across testing rounds to present average impacts unless otherwise stated. For inference, we cluster standard errors by student, the unit of randomization, in all analyses throughout the paper.

β_1 captures the treatment effect in the first decile of the test—i.e. the level effect at the start of the test when students are still cognitively fresh. β_2 captures treatment effects for questions in deciles 2-5 of the test. The primary coefficient of interest is β_3, which captures the treatment effect in the second half of the test, i.e. deciles 6-10. We predict that β_3 will be positive: cognitive practice will mitigate the rate of decline toward the end of the test when cognitive fatigue has set in.

While helpful in its simplicity, one potential limitation of the approach in Equation 1 is that it implicitly takes a stance on when treatment effects on declines should arise (i.e. the second half of the test). However, the scope for effects occurs once the control group starts declining in performance. For example, if there are treatment effects in deciles 4-5 of the tests, then focusing on the second half the test alone will not fully capture treatment effects. We therefore complement the above with a more flexible, higher-powered approach based on this intuition. To obtain a data-driven proxy for expected declines at each point in time throughout the exam, we use data from the *baseline tests*. Specifically, for each school, we compute how much worse students do in later questions relative to their performance at the start in baseline tests:[24]

$$PredictedDecline_l = E[Correct_{ils}|location = 1] - E[Correct_{ils}|location = l] \tag{2}$$

Since some tests have a small number of questions, we use quintiles as location bins to reduce noise. The first term, therefore, captures average baseline test performance in quintile 1. The second term captures this average for quintile l, where l takes the values 1-5. Consequently, $PredictedDecline_l$ serves as a proxy for how much worse we would expect students to do over the course of a test in the absence of any intervention. We then test whether receiving the Cognitive Practice treatment mitigates the rate of expected performance decline by estimating:

$$\begin{aligned} Correct_{ils} = \ & \alpha_0 + \alpha_1 CogPractice_s + \alpha_2 PredictedDecline_l \\ & + \alpha_3 CogPractice_s * PredictedDecline_l + \alpha_4 Baseline_s + \chi_{il} + \epsilon_{ils} \end{aligned} \tag{3}$$

where $PredictedDecline_l$ is as defined in Equation (2), and all other covariates are as defined in Equation (1). The main coefficient of interest is α_3, which captures the fraction of expected decline

value. When pooling across different test subjects, we allow both the $CogPractice_s$ term and decile fixed effects to vary by test subject, to allow for different level effects and declines across subjects. Finally, because the tests in higher grades had more questions (and therefore observations), we also weight by the inverse of the number of questions in the test so that each student-test receives equal weight.

[24] We omit the subscript indexing the predicted decline variable separately by school from the regression notation for simplicity. Results are similar if we do not compute this measure separately by school, or use alternate approaches (see Appendix Table A.11).

that is mitigated by the treatment. Under our hypothesis, α_3 will be positive: the treatment group will decline less steeply than the control group. α_1 captures the initial level effect: treatment effects in the early parts of the test before declines set in.

5.4 Impacts on Performance Declines

Table II presents the results from both estimation approaches. Col. 1 reports estimates from Equation (1), pooling across listening, Ravens Matrices, and math. Receiving Cognitive Practice increases average performance in the second half of the test by 1.31 percentage points (p.p.) ($p = 0.006$), corresponding to a 21.9% improvement relative to the control group. In addition, we see some evidence that treatment effects begin to emerge in deciles 2-5 of the tests, with an estimated effect of 0.79 p.p. ($p = 0.111$). In contrast, there is no discernible difference between the treatment and control groups at the start of the tests; the estimated effect in the first decile is -0.0012 ($p = 0.849$). In Panel B, Col. 1, we repeat the analysis disaggregating the two sub-treatments. The Math and Games arm each lead to higher performance in the second half of the tests by similar magnitudes, corresponding to treatment effects per question of 1.31 p.p. (21.9%, $p = 0.018$) and 1.32 p.p. (22.0%, $p = 0.015$), respectively. In addition, we detect no level effects for either sub-treatment: the estimated effects in the first decile of the tests are small and insignificant.

In Col. 2, we estimate Equation (3) to obtain a summary measure of treatment effects across the test, with a similar pattern in results. Receiving Cognitive Practice mitigates 9.25% of the expected decline over the test ($p = 0.002$).[25] Looking at each of the two sub-treatment arms separately in Panel B, the effects for each are similar in magnitude: the Math arm mitigates 9.62% of the expected decline ($p = 0.004$), and the Games arm mitigates 8.92% of the expected decline ($p = 0.008$). Because each experimental arm had different levels of math exposure, we also show overall results excluding the math test in Col. 3; the results remain similar to those in Col. 2.

The remaining columns in Table II disaggregate the results across subjects. In Panel A, receiving Cognitive Practice reduces expected declines in math by 10.8% ($p = 0.014$), listening by 7.1% ($p = 0.037$), and Raven's Matrices by 8.8% ($p = 0.054$).[26] We cannot reject that these three treatment effect coefficients are equal ($p = 0.765$). Panel B, Cols. 4-6 provide fully disaggregated results. The Math sub-treatment generates significant effects for each test subject, while some of the effects of the Games sub-treatment become noisier when results are fully decomposed. While it would be interesting to examine whether a sub-treatment has relatively larger impacts on performance declines when the test subject is more closely related to the content that was practiced, we are

[25] In the control group, pooling across tests, the average decline from the first to fifth quintile is 12 percentage points. In Col. 2, the coefficient on "Predicted Decline" is -0.29 ($p < 0.001$) (omitted from the table for space). Consequently, in percentage terms, the reduced form effects in Cols. 2-6 will be smaller in magnitude than those in Col. 1.

[26] In the April Raven's Matrices tests, there was a clerical error which led to test modules that were up to 80 questions, so that they could not be completed in the allotted time—making it impossible to use them to identify performance declines. These tests are excluded from the results, but the findings are robust to including these tests in the analysis (Appendix Table A.12). Test modules included in the analysis have, on average, 30 questions.

under-powered for such analysis. We cannot reject that each sub-treatment has the same effect on each test subject, but this may mask meaningful subject-specific differences in effects across the two training approaches.

In addition, as above, we see no evidence for any initial level effects from either treatment arm—particularly for the listening and Raven's tests. For the math test, we might expect some initial level effects since the Math and control arms spend their time solving math problems while the Games arm does not. Consistent with this, we see suggestive evidence the Games arm does worse on the math test relative to the other arms even at the start of the test (Panel B, Col. 4, coefficient $= -0.0191$, $p = 0.083$). However, recall from the discussion above that the math test was not constructed to be diagnostic of learning gains from the Math sub-treatment; the lack of strong level effects is therefore not surprising.[27] The lack of effects at the start of the tests is consistent with our design objective of choosing test domains where the scope for learning from the sub-treatments is negligible, particularly in listening and Raven's. This greatly simplifies the interpretation of the decline results: since there has been no change in ability (i.e. levels), the change in declines can be interpreted as an improvement in cognitive endurance.

These treatment effects are robust to restricting to only attempted questions (Appendix Table A.13), alternate empirical specifications varying the controls included (Appendix Table A.14), and bootstrapping standard errors to adjust for the fact that "Predicted Decline" is a constructed variable (Appendix Table A.15). We do not find any significant heterogeneity in treatment effects by grade, gender, baseline average score, or baseline decline in performance (Appendix Table A.6, Panel B).

Overall, these patterns indicate that each of the two approaches for cognitive practice—academic and non-academic—reduce the severity of performance declines over time. The magnitude of the impacts on decline rates in Table II is meaningful. For example, if the average treatment effect of the Cognitive Practice treatments were applied to the TIMSS data, it would reduce the difference in decline rates between high and low-income countries by 30%, or between White and Black/Hispanic students by 50%.

Finally, the improvements in declines produce suggestive increases in test performance overall (Cols. 7-8). Cognitive Practice increases the overall probability of getting a given question correct by 0.85 p.p., or 1.7% across all subjects ($p = 0.089$), and by 0.93 p.p. or 1.8% in the non-math subjects ($p = 0.074$).[28] These overall effect sizes are modest—as may be expected given that, by construction, the potential for learning is minimal (leading treatment effects at the start of the tests to be zero), and students are given ample time to finish (to eliminate the scope for increased performance on

[27] Recall that the math software provided no instruction, and the software practice problems differed from those on the math declines test, both in terms of content and format. Because students are typically taught via rote memorization in this setting, when they encounter problems in a new format (e.g. 2+3 written horizontally versus being asked to add two balls with three balls), they would have difficulty translating concepts. Note that this does not undermine the potential to use tablet-based general math practice as a learning tool; rather, we explicitly constructed the use of the tool and assessments to minimize scope for level effects in the assessments.

[28] We report subject-wise estimates in Appendix Table A.16. In addition, in Appendix Table A.17 we report the effects of Math practice on math test performance by question difficulty.

the extensive margin of attempting additional questions). As we discuss in Section 5.1 above, these design choices are necessary to construct a clean positive test for cognitive endurance effects. Rather, these findings, along with effects on other dimensions such as classroom attentiveness below, support the idea that cognitive endurance is an input into myriad aspects of learning—from listening comprehension in a lecture to solving a long homework assignment. Consequently, even small effects, when aggregated across domains and over time, could generate large overall performance impacts—consistent with the sizable effects of Cognitive Practice on school performance in Table I.

5.5 Persistence of Cognitive Endurance Effects

To test for persistent effects on cognitive endurance, we implement a follow-up round of the performance decline tests. These tests are conducted 3-5 months after the end of treatment activities across schools. They take place after the vacation break—when students progress from one grade to the next—a time during which there is often substantial decay in academic skills (Cooper et al., 1996; Alexander et al., 2007).

Table III tests for persistence. In each column, we report the average treatment effects for the main experimental period and the follow-up round. In Col. 1, we examine treatment effects only in the second half of the test (based on the estimation strategy in Equation 1); the estimated effect in the follow-up round is positive, with a magnitude that is 76% the size of the effect during the main rounds, but statistically insignificant. Using the higher-powered approach that examines treatment effects across the test (based on the predicted decline approach in Equation 3), we find strong effects in the follow-up round. Specifically, being assigned to Cognitive Practice mitigates 9.2% of the predicted decline during the main intervention period ($p = 0.002$), and 9.2% of the predicted decline in the follow-up round ($p = 0.033$). We find similar evidence of persistence for both the Math and Games arms (Cols. 3 and 4, respectively). At the bottom of the table, we report the F-test p-values testing for the equality of treatment effects during the experiment and the follow-up; across all specifications and columns, we cannot reject that the two sets of impacts are equal.

These data suggest the presence of some persistence, but whether we should expect persistence over longer periods is unclear. Rather, by documenting that cognitive endurance is malleable, our study opens the possibility that environmental factors could perpetuate differences across individuals. For example, if richer students attend schools or have home environments that allow more time for practicing concentration, as suggested by Figure A.2, then this could continually reinforce differences in cognitive endurance.

6 Supplementary Tests: Mechanisms and Confounds

In this section, we present additional tests to help clarify the mechanisms behind our main results on school performance and cognitive endurance. First, we supplement our findings on cognitive

endurance with two additional measures drawn from the psychology literature: laboratory measures of sustained attention and a measure of students' classroom attentiveness. If our effects operate by expanding the ability to sustain attention toward a goal, then we would expect to see improvements in these measures. Second, we assess the potential relevance of motivation and perseverance in driving our effects. Finally, we evaluate alternate mechanisms that would operate outside of impacting cognitive endurance—such as confidence, fairness concerns, or alternate cognitive channels such as working memory.

6.1 Psychology Measures of Sustained Attention

Psychologists refer to the ability to sustain cognitive effort toward a goal or activity over time as "sustained attention" (also referred to as attentional "vigilance").[29] Sustained attention is therefore the core underlying cognitive process that enables cognitive endurance.

The canonical measure of sustained attention is the Sustained Attention to Response Task, or SART (Smilek et al., 2010; Robertson et al., 1997). In this 8-minute task, students look at a computer screen as 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.[30] If a student has lost focus (e.g. is daydreaming), this will result in a slower reaction time or reduced accuracy. This provides an abstract, context-free measure of the capacity for sustaining cognitive effort over time.

In addition, we also examine impacts on a supplementary task used in the psychology literature—the symbol matching task, which is an adapted version of a concentration endurance task (Bates and Lemay, 2004; Dean et al., 2019). Students are given pages containing a grid of randomly ordered pictorial symbols. A specific set of 1-3 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. Students repeat this for 10-15 additional grids over 15 minutes. Example screen shots from the SART test and an example page from the symbol matching task are shown in Appendix Figure A.10. For each of these two tasks, the standard outcome measure used in the psychology literature is the z-score of the student's correct true positive rate minus the false-positive rate across the test (Youden, 1950).

Table IV presents intent-to-treat estimates of the impact of the intervention on these measures. Relative to the control group, the treatments increase average performance on the psychology measures of sustained attention by 0.048 SDs (Table IV, Col. 1, $p = 0.018$). We decompose these

[29] Psychologists decompose attention into three core functions: i) selection among competing items, ii) modulation of the selected item (i.e. processing efficacy and efficiency), and iii) sustained attention or vigilance: sustaining focus towards a chosen goal (Chun et al., 2011).

[30] Note that this test is not an intellectually demanding one (in contrast, for example, to a Raven's fluid intelligence test or cognitive control task). Rather, it captures the level of attentiveness via perception of stimuli in the environment. This task is part of a broader category of measures called "continuous performance tasks". We modify the traditional SART task slightly to make it less challenging and more child-friendly. For example, we adjust the frequency of the target stimuli and use shapes rather than numbers.

effects in the remaining columns. Notably, performance on SART increases by 0.067 SDs (Col. 2, $p = 0.028$). The average effect on the symbol matching task is 0.036 SDs, but this difference is not significantly significant (Col. 3, $p = 0.140$). We further decompose these results by examining effects on true positive and false positive rates separately in Appendix Table A.19.

In Col. 4, we examine impacts separately by each sub-treatment. The Math and Games arms improve average performance by 0.054 SD and 0.043 SD, respectively ($p = 0.023$ and $p = 0.071$), and we cannot reject equality of the effects of both sub-treatments ($p = 0.641$). Overall, these findings indicate that the treatments improved sustained attention as measured by psychologists.

6.2 Attentiveness in the Classroom

We monitor students' behavior during their regular class periods and assess them on measures of attentiveness. Our rubric draws on components of a diagnostic teacher rating scale commonly used to measure attention in the classroom: the Vanderbilt Attention-Deficit/Hyperactivity Disorder (ADHD) Diagnostic teacher rating scale, which is commonly used to assess students for signs of ADHD prior to a formal diagnosis. We adapt this scale to our setting. We examine student behavior along three dimensions: (1) whether students attend to and carry out several instructions from the teacher (e.g. writing down their information in a particular location on a paper and turning it in five minutes later as they transition to a new activity);[31] (2) their response to auditory stimuli (noticing a sound from a program staff member made in the hallway meant to get students attention); and (3) their physical signs of inattention (fidgeting, looking out the window, or off-task behavior during the teacher's lecture).

These observations are conducted with students organized in their normal class section (i.e. students from different treatment arms are mixed in the room), while students listen to a lecture and are engaged in common classroom activities. Student behavior along each dimension is rated by classroom observers who are blind to treatment status and sit quietly in the back of the room for the entire class. Note that students are unlikely to think their behavior is being observed; it is common to have a teaching assistant or head teacher sit at the back of the room and observe the class to provide teacher feedback. We conduct these observations for a subset of class sections across all schools.[32]

Students who receive Cognitive Practice exhibit increased classroom attentiveness on average, with a mean improvement of 0.083 SDs on the index overall (Table V, Col. 1, $p = 0.010$). We see

[31] Teachers asked students to bring their paper and pencil with them as they transitioned from one classroom to another and to write their name at the top of the page once they got to the other classroom. For a student to successfully complete this, they need to have listened and attended to the specific details in the teacher's instruction (i.e. which materials, where to write their name, etc), and then had the presence of mind to carry it out five minutes later rather than forgetting.

[32] Due to operational constraints, we did not conduct classroom observations for a cohort of fifth graders in three of the schools. In addition, students who were absent on their class's observation day would not be included in the analysis. Unlike the other individually administered test measures, where we tested absent students when they returned to school, it was not possible to do make-up observations for absentees.

evidence of improvements in the Math sub-treatment of 0.108 SD ($p = 0.003$), and in the Games sub-treatment of 0.058 SD, though this latter effect is not significant at conventional levels ($p = 0.117$) (Col. 5). These results suggest that practicing focused cognitive activity improves how attentive students are in the classroom.

6.3 Motivation and Perseverance

Should we understand our cognitive endurance effects as reflecting cognitive improvements, versus a broader view of endurance that could include factors such as motivation or perseverance? Psychologists consider such "cognitive" and "non-cognitive" channels to be inherently related. For example, sustained attention is an upstream input into perseverance, self-control, and other behaviors that involve sustaining focus towards a goal (Chun et al., 2011; Mischel, 2014; Zelazo et al., 2016). Consequently, we do not attempt to draw a strong line between them. This informs our choice of the more general term "endurance" to describe performance declines during cognitive tasks.

In this section, we explore the potential relevance of factors such as motivation and perseverance. For example, did the Cognitive Practice treatments prompt students to try harder in school, or improve their perseverance in continuing to work rather than give up when tired? This is relevant for interpreting the mechanisms behind both the effects on cognitive endurance, as well the school performance results more broadly. We draw on three supplementary pieces of evidence.

First, the declines results in Table II indicate that students are not simply trying harder in general, since that would lead to improvements at the start of the tests as well. Mean control group performance in the first decile of the declines tests is roughly 50%, leaving ample scope for treatment effects in the beginning. This would require a more specific type of motivational mechanism: one that operates specifically when students start to become cognitively fatigued. Alternately, test performance may not be elastic early in the test; it is possible that increased motivational drive only becomes relevant later when fatigue has set in, and effort is needed to keep going.

Consequently, second, we implement a more direct test for whether being more driven reduces performance declines. For one testing round, we randomize incentives so that students earn toys and other prizes for higher test scores on the listening and Raven's Matrices tests.[33] Specifically, students are told they will be able to choose a specific prize based on their quartile of performance, with higher quartiles having more attractive prize options. These prizes range from stickers to colored pencil sets to highly coveted toys. We used focus groups with students to come up with the prizes and rank order them in quartile sets to ensure the effectiveness of the incentives. This design provides an incentive for all students to try harder, across the skill distribution. We randomize at the school-grade-test level: within a school, students within the same grade have the same incentive

[33]In Appendix Table A.20, we verify that our main treatment effects on declines are similar if we restrict the data to observations where no incentives were offered. Note that this test relates to a broader literature on the effects of incentives on intrinsic and extrinsic motivation (e.g., Deci, 1971; Lepper, 1988), as well as using incentives to examine the role of motivation in test performance (Gneezy et al., 2019).

treatment status within a given test subject (e.g. Raven's).

Receiving incentives significantly increases average performance on the tests (Table VI, Col. 1). In addition, there is no differential impact of the incentives among students who receive Cognitive Practice and those who do not. In Col. 2, we examine effects over the course of the exam. When students are more motivated to do well, their performance in the first decile of the test increases sharply by 16.0 percentage points ($p = 0.016$). This indicates that performance is highly elastic to effort even at the start of the tests when students are cognitively fresh.

However, we see no evidence that being more driven reduces performance declines. If being more motivated led to less decline, the "Decile 6-10 x Incentive" coefficient in Col. 3 would be positive and significant. However it is actually negative in sign and insignificant. We find a similar pattern using the predicted decline specification in Col. 3: at the beginning of the test there is a sizable effect of the incentive (13.4 percentage points), but there is no improvement in the rate of decline; the interaction is actually negative and marginally significant.[34] These patterns are consistent with previous work in the decision fatigue literature that shows that increased cognitive effort early in a task will more quickly drain cognitive reserves, leading to worse performance later in the task (e.g., Levav et al., 2010). Moreover, there is no differential impact of incentives for students who receive Cognitive Practice. Overall, these patterns suggest that when students are motivated to try harder, this does not produce the same pattern of results as in Table II: they do better throughout the test (i.e. from the beginning), and do not exhibit less decline.[35]

Third, we consider whether the treatment effects on declines are coming from students just persevering in the face of difficulty. It is not clear why the treatments should necessarily affect this, since program classes were not set up to train perseverance or grit. For example, the practice sessions were open-ended: there no set number of questions that students needed to finish, and students were told to simply stop if they were cognitively fatigued (and especially in the Games arm, students regularly did so). Consequently, there was no feeling of persevering to get to the "end" of a set of questions or activity.

None the less, to undertake a very suggestive test for perseverance, we examine whether treated students respond differently after facing a question that most students get wrong. On average, when control students encounter a difficult question, they do worse on the subsequent question as well: after receiving a question in the bottom quartile of accuracy, students are about 2 p.p. (4.3%) less likely to get the subsequent question correct ($p < 0.001$) (Appendix Table A.21, Col. 3). Note that because question order is random and we include question fixed effects, this pattern is not simply an artifact of serial correlation in question difficulty. Such discouragement effects have been previously

[34] In Col. 3, "Predicted Decline" takes an average value of 0.2 at the end of the tests. Consequently, the total effect of incentives by the end of the test is the "Incentive" coefficient plus the "Predicted Decline x Incentive" coefficient multiplied by the expected value of "Predicted Decline" at the end of the test: 0.134 + (-0.350)*(0.2) = 0.064.

[35] If we run the estimation in Cols. 2 and 3 without the triple interaction, we can reject that incentives and Cognitive Practice have the same effect on declines at the 10% level in each case (the F-test p-values are 0.071 and 0.059, respectively).

documented in the education literature (e.g. Vos and Kuiper, 2003). However, Cognitive Practice does not mitigate discouragement effects: the interaction of treatment with a dummy for whether the previous question was difficult is insignificant across specifications. This is also true in the listening subject, which only asks quick simple recall questions (Col. 5).[36] Consequently, by at least this one specific metric, treated students do not exhibit greater resilience than control students.

The above tests, while informative, may not capture all dimensions of motivation, perseverance, or grit. Consequently, we view these exercises as exploratory. However, coupled with the positive evidence on sustained attention above, they indicate at least some role for cognitive improvements.

6.4 Potential Confounds

The Cognitive Practice treatments could arguably boost students' performance through other channels. For example, they may have increased confidence or excitement about school, affected the morale of control students due to fairness concerns, or improved alternate cognitive abilities such as working memory.

Note that a priori, the most straightforward version of all these explanations should lead to improvements across the duration of the declines tests, including at the beginning—in contrast to our findings. Such explanations could affect the slope of performance declines, but if they were driving the results, it is unclear why they should not manifest as students also doing somewhat better early in the tests. For example, if the treatments improved the ability to hold an object in working memory and manipulate it, this may make a student better at a Raven's or math test, but should also affect performance on the first question. Note that for the listening test, the randomized question order ensures that later questions are not based on content that appeared later in the story passage; thus, if students became better at holding a detail from the story in working memory, this should help them with performance across the (randomly placed) questions in the test.

While the arguments above apply to a wide variety of potential confounds, we supplement these with additional tests to further assess some alternate channels one by one below:

Student morale. We examine attendance data as a supplementary test for an excitement or morale channel. We might expect such a channel to affect the desire to come to school. Since we do not have administrative attendance data from schools (and some of the schools did not keep good attendance records), we proxy for this by examining whether control students were more likely to be absent on the day of surprise experimental endline tests—which occurred on 12 separate days per student on average over the course of the experiment. Similarly, they may be more likely to leave the school (either during the school year, or in the transition between one school year to the next). We test for treatment effects on attendance and re-enrollment in Appendix Table A.22 and Panel D of Appendix Table A.4, respectively, and find no evidence the treatment had an effect on either

[36]This test is especially interpretable for the listening test, since the questions are short multiple-choice questions of knowledge (requiring recall rather than significant cognitive effort to solve them). Consequently, a decline in performance after a hard listening question more likely reflects discouragement effects.

outcome. For example, the estimated impact of Cognitive Practice on attendance is -0.0045, and we can reject at the 5% level that the treatment had an effect of 1.4 p.p. (or greater) on attendance.

Grader bias. Another potential concern is that, despite our efforts to maintain confidentiality, teachers may have learned students' treatment status and given more favorable grades to treated students. As a potential test for this concern, in Appendix Table A.23, we examine heterogeneity by the degree of latitude teachers had in assigning grades for a given subject or group. We find no differential treatment effects by the level of subjectivity in grade assignment—assuaging concerns around teacher manipulation of scores.[37] In addition, this concern could not explain the improvements in performance on the decline tests, the psychology measures, or classroom attentiveness, which were all measured by the research team. Finally, schools and teachers had no dynamic incentives to make the intervention look successful: as per our ex ante agreement with schools, this was a one-year study, and schools would be allowed to keep all tablets and software free of charge at its conclusion.

Test-taking skills. Our results also cannot be explained by treated students developing better test-taking skills, such as skipping hard questions, due to the design of both the treatments and outcome measures. First, these skills could not be developed via the Games sub-treatment. Second, we directly mitigate this concern by ensuring sufficient time and high completion rates for tests, and by verifying that results are similar when we restrict analysis to attempted questions. Finally, recall that a subset of our tests—listening and SART—mechanically do not permit students to skip around or move faster through the tests. In addition, given that the declines tests are based on traditional paper-pencil assessments, increased familiarity with tablets also cannot explain our results.

Household inputs. Finally, we examine whether parents responded to the treatments by changing inputs at home. This is not a confound but is relevant for interpreting mechanisms and the magnitude of our treatment effects. While our data on out-of-school activities is limited, we collected some measures of time use from students in an endline survey, presented in Appendix Table A.24. We find no reported impacts on the amount of time spent at home in cognitive practice, on homework, on homework help from parents, or on whether students eat breakfast. Of course, these results do not rule out the general relevance of students' home environments in developing cognitive endurance.

Overall, the treatments could plausibly have affected various other channels. However, any potential channel would need to explain effects for both the Math and Games sub-treatments, and across the disparate tests such as listening and Raven's Matrices. In addition, it would need to explain why there are no detectable changes in performance at the start of the tests.

[37] In our sample schools, in upper grades (i.e. 3-5), Hindi and English assessments often involve writing sentences or paragraphs—making grading more subjective relative to lower grades (where students write in a simple word into a blank) or relative to math (where individual questions are either wrong or right). We show example pages from students' school assessments for math and English in Appendix Figure A.11. In Appendix Table A.23 Cols. 1-4, we find no differential effect along these dimensions. In addition, two of our sample schools were part of a school chain, where tests and strict grading rubrics were developed centrally at headquarters. We also see no differential treatment effects among chain vs. non-chain schools (Panel A, Col. 5).

6.5 Discussion

In summary, our study indicates two main sets of findings. First, the act of effortful thinking alone has broad benefits—proxied by improved school performance across unrelated domains. Second, effortful thinking changes a particular capacity: cognitive endurance. Our findings suggest many potential ways higher cognitive endurance could raise school grades, offering potential links between these two sets of results.[38] For example, it may improve one's ability to be attentive in class, listen to and retain material from a lecture, maintain focus while reading a textbook, think through a challenging concept at the end of a long class, or get questions right at the end of an exam.

These findings do not preclude the possibility that our treatments may have benefits through channels other than cognitive endurance. While our design allows us to evaluate some prominent alternate channels, we do not assess an exhaustive set of measures. However, our goal is not to provide a full accounting of the school performance results. Rather, we view our two main sets of findings as offering complementary evidence on the potential link between schooling and generalized mental capacity.

7 Impact of Schooling on Cognitive Endurance

Our paper is motivated by the hypothesis that schooling may expand our underlying capacity to engage in cognition itself. Our intervention improves cognitive endurance, but it does so by introducing an external intervention into schools. In this section, we augment our experimental evidence by examining whether the traditional experience of schooling helps develop cognitive endurance. We construct a suggestive test by exploiting quasi-random variation in years of schooling.

We use a sample of 5,300 nine to eleven-year-olds across 66 schools in Pakistan (Brown and Andrabi, 2021). In this school system, students born on January 1 or later are supposed to wait an additional year to enroll in school. By examining performance declines among students born just before versus after this cut-off, we compare students who are nearly the same age, but differ in their current years of schooling.[39] Because compliance with the birth month cut-off is imperfect, we use a fuzzy regression discontinuity (RD) approach, instrumenting for years of schooling with expected years based on birth month. Students who are born just after the enrollment cut-off have

[38] We use a simple back-of-the-envelope exercise to get a suggestive sense of whether the effects on school grades are due to learning vs. simply improved exam performance. This exercise assumes that the impact on performance declines in school assessments would be similar to the impacts in our declines tests—a strong assumption. Using the estimates for non-math tests in Table II, Col. 1, the treatments mitigated the decline by 0.0244 SD in deciles 6-10. This magnitude corresponds to only 27% of the 0.0897 SD treatment effect on school grades in Table I. This would imply that increased learning accounts for over 73% of the impact on school grades. These estimates are similar if we instead compute the numerator using all test subjects in the declines tests or restrict to only the listening test. However, given the assumptions required, these estimates should be viewed as rough and only suggestive.

[39] This exercise has limitations—for example, enrolling in school earlier versus later could also change other inputs, alter ones' relative emotional maturity, or provide different amounts of time at home. We therefore view it as only suggestive, but nonetheless complementary to our field experiment results.

0.22 fewer years of schooling ($p < 0.000$).[40] To estimate effects on performance declines, we use the same regression specification as Equation 3 but replace the treatment dummy with the predicted years of schooling (see Appendix B for details on the data and specification).

We first replicate the presence of substantial performance declines in the test data, which include the subjects of math, science, English and Urdu. In all tests, question order is randomized across students. On average, the probability of answering a given question correctly declines by 16 percentage points (24%) from the first to last decile of the test (Appendix Figure A.12).

Conditional on age, an additional year of schooling mitigates the rate of performance decline by 31% (Table VII, Col. 1, $p = 0.030$). These results are similar under both linear and quadratic functional forms of the running variable (Cols. 1-2).[41] This improvement is substantial: it would shift the average student from a 16 p.p. performance decline to a 11 p.p. decline in performance over the length of the exams. As an alternative benchmark, the impact of a full year of schooling is 3.4 times larger than the 9% estimated effect of our more limited experimental intervention.

We next examine whether these effects vary with school quality and pedagogy. Quality measures in the dataset are based on video recordings of students' classes, coded using the Classroom Assessment Scoring System (CLASS) rubric, a common tool for assessing pedagogical quality (Araujo et al., 2016; Pianta et al., 2012). Among the schools in the bottom quartile of the quality distribution, we cannot reject that additional schooling has no effect on cognitive endurance (Cols. 3-4). However, as we move up the quality distribution for each quartile rank increase, an additional year of school reduces performance declines by 22% ($p = 0.034$). Similarly, the positive impacts of schooling on cognitive endurance are concentrated in schools where students spend more time in independent, focused practice in class (Cols. 7-8).

Overall, the results in Table VII provide further, albeit suggestive, evidence for the role of schooling in developing cognitive endurance. They also indicate that the school's institutional and teaching environment is important: better schools appear substantially more effective in enabling students to develop this ability. Coupled with the results from our experiment, the complementary results in this section suggest that differential access to good quality schooling could widen achievement gaps by hampering the development of a core cognitive capacity.

8 Conclusion: Discussion and Broader Relevance

We conclude by discussing the broader implications of our findings. We first begin by presenting evidence for systematic differences in cognitive endurance in behaviors outside of schooling. Using supplementary data, we present two examples from substantially different high-stakes activities:

[40] Results are robust to using a linear and quadratic specifications for the first stage, and we do not find evidence of manipulation in the birth month variable around the cut-off (McCrary test $p = 0.504$).

[41] Our running variable only takes 12 values, for each month of the year, so we are limited in our ability to employ local polynomial or optimal bandwidth selection procedures.

productivity among data entry workers and voting at the ballot box.[42]

In Figure IIIa, we plot the hourly performance of full-time data entry workers over nine months using data from Kaur et al. (2015). Workers' earnings are comprised of a piece rate for each *accurate* field entered. Mistakes are costly: an inaccurate entry means that the worker has exerted the effort to enter the field but is not compensated for it. On average, error rates increase roughly 12% between 10am and 4pm.[43] Less educated workers (i.e. those without a high school degree) experience a decline in accuracy that is twice as large as that of more educated workers. This accounts for 10% of the productivity gap between more and less educated workers in the sample.

We find similar patterns in voting behavior, building on the work by Augenblick and Nicholson (2015). Using quasi-random variation in the order of ballot initiatives, the authors find that, when items are further down-ballot, individuals are substantially more likely to vote the default option (i.e. less likely to make an active, non-default choice). These effects are substantial: an additional 6% of propositions would have become laws if they had appeared first on the ballot. Using data obtained from the authors, we use racial composition of a voting precinct as a proxy for socioeconomic advantage (since income is not available in their data). In the early items on the ballot, the likelihood of picking the default option is quite similar for neighborhoods with more white vs. more non-white residents (Figure IIIb). However, over time, the propensity to make an active (i.e. non-default) choice declines much more quickly for less advantaged neighborhoods. Specifically, more advantaged groups decline 29.3% less quickly between the first and last quartile of ballot positions.

These patterns are in line with previous work showing that cognitive endurance is relevant for many aspects of daily life. While the examples in Figure III are certainly not exhaustive, they indicate the possibility that those from disadvantaged backgrounds continue to exhibit worse cognitive endurance as adults—with potential implications for their labor earnings, decision-making, and myriad other outcomes. For example, could high traffic accident rates in developing countries be influenced by more rapid attentional declines while driving, especially in the long shifts worked by truck and taxi drivers? Could differences in cognitive endurance contribute to positive feedback loops in learning akin to "dynamic complementarities" (Cunha et al., 2006; Cunha and Heckman, 2007)? In addition, the patterns in Figure III raise the question of how we should understand the systematic differences in cognitive endurance among adults. For example, such differences may reflect persistence from childhood training; alternately, they may reflect the fact that those working in higher-skilled jobs may receive more cognitive practice through work—reinforcing and perpetuating

[42]The choice of these examples is driven by data availability to fulfill two requirements: (i) situations where declines over time are interpretable as cognitive fatigue effects due to the absence of obvious confounders (e.g. the task itself does not get harder over time); and (ii) situations where differences between more and less advantaged individuals are not severely confounded by differential selection into the task.

[43]In this study, workers are recruited irrespective of their experience or educational background, mitigating some of the differential selection into the sample by education level (the only proxy for relative advantage in this dataset). Analysis is conducted using 10 am - 4 pm to avoid compositional effects of workers arriving and departing. The piece rate would need to increase by an estimated 2.4% at the end of the day to undo the performance decline (based on the effort elasticity of 0.33 from Kaur et al. (2015)).

differences over the life cycle. These possibilities are of course only speculative, but given their implications, warrant further research.

Our study indicates that systematic differences in cognitive endurance are not a given; they can be ameliorated.[44] Our intervention exemplifies a policy lever that may be useful in improving endurance among lower-income children: incorporating opportunities for them to engage in effortful thinking for sustained periods of time at school or at home. The supplementary data we present in Sections 2 and 7 indicate that better quality schools are already employing such strategies, but those attended by less privileged students are not. This may be due to real barriers, such as disruptions, unruly peers, or heterogeneous achievement levels within a class. In our setting, using tablets to both engage students and address heterogeneous skill levels was an effective solution. In other settings, other approaches may be more appropriate for enabling students to undertake mentally challenging activities. As an example, in low-income US settings, some schools that have successfully improved student outcomes place strong emphasis on giving children frequent assessments (Angrist et al., 2013; Dobbie and Fryer, 2013). Because this approach creates regular periods where students must sit and concentrate for long stretches (i.e. during weekly tests), it may have the ancillary benefit of improving cognitive endurance. The fact that we find gains using both academic and non-academic practice suggests that a broad array of approaches could be effective.

More broadly, we view our study as tracing one of the many potential pathways through which schooling may shape human capacities—beyond its effects on academic skills. Additional work linking specific elements of schooling to these capacities can help further our understanding of why education has such broad and persistent benefits. It may also offer normative insights on how to address disparities in human capital development.

UNIVERSITY OF CHICAGO, UNITED STATES
UNIVERSITY OF CALIFORNIA, BERKELEY AND NATIONAL BUREAU OF ECONOMIC RESEARCH, UNITED STATES
UNIVERSITY COLLEGE LONDON, ENGLAND
UNIVERSITY OF PENNSYLVANIA, UNITED STATES

[44]In addition, while related work views lower performance among disadvantaged students as stemming from "preferences" (e.g. motivation, growth mindset, etc.), our study suggests some of these differences are due to "skills" such as cognitive endurance, which can be built with deliberate practice. However, note that our design does not speak to whether our intervention increased the total "attention budget".

References

Akyol, Pelin, Kala Krishna, and Jinwen Wang, "Taking PISA Seriously: How Accurate are Low-Stakes Exams?," *Journal of Labor Research*, 2021, pp. 1–60.

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.

_ , Teodora Boneva, and Seda Ertac, "Ever failed, try again, succeed better: Results from a randomized educational intervention on grit," *The Quarterly Journal of Economics*, 2019, *134* (3), 1121–1162.

Alexander, Karl L, Doris R Entwisle, and Linda Steffel Olson, "Lasting consequences of the summer learning gap," *American sociological review*, 2007, *72* (2), 167–180.

Almlund, Mathilde, Angela Lee Duckworth, James Heckman, and Tim Kautz, "Personality psychology and economics," in "Handbook of the Economics of Education," Vol. 4, Elsevier, 2011, pp. 1–181.

Angrist, Joshua D, Parag A Pathak, and Christopher R Walters, "Explaining Charter School Effectiveness," *American Economic Journal: Applied Economics*, October 2013, *5* (4), 1–27.

Araujo, M. Caridad, Pedro Carneiro, Yyannú Cruz-Aguayo, and Norbert Schady, "Teacher Quality and Learning Outcomes in Kindergarten*," *The Quarterly Journal of Economics*, August 2016, *131* (3), 1415–1453.

ASER, *Annual Status of Education Report India* 2019.

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, 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.

Bates, Marsha E and Edward P Lemay, "The d2 Test of attention: construct validity and extensions in scoring techniques," *Journal of the International Neuropsychological Society*, 2004, *10* (3), 392–400.

Becker, Gary S, *Human capital: A theoretical and empirical analysis, with special reference to education*, University of Chicago press, 2009.

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.

Bettinger, Eric, Sten Ludvigsen, Mari Rege, Ingeborg F Solli, and David Yeager, "Increasing perseverance in math: Evidence from a field experiment in Norway," *Journal of Economic Behavior & Organization*, 2018, *146*, 1–15.

Boksem, Maarten AS, Theo F Meijman, and Monicque M Lorist, "Effects of mental fatigue on attention: an ERP study," *Cognitive brain research*, 2005, *25* (1), 107–116.

Borghans, Lex and Trudie Schils, "The leaning tower of PISA," Technical Report, Working Paper. Accessed February 24. http://www. sole-jole. org/13260. pdf 2015.

_ , Bas Ter Weel, and Bruce A Weinberg, "People skills and the labor-market outcomes of underrepresented groups," *Ilr Review*, 2014, *67* (2), 287–334.

Borgonovi, Francesca and Przemyslaw Biecek, "An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test," *Learning and Individual Differences*, 2016, *49*, 128–137.

Bowles, Samuel and Herbert Gintis, "Schooling in Capitalist America: Educational Reform and the Contradictions of Economic Life," 1976.

_ , _ , and Melissa Osborne, "The Determinants of Earnings: A Behavioral Approach," *Journal of Economic Literature*, December 2001, *39* (4), 1137–1176.

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.

Brown, Christina and Tahir Andrabi, "Inducing Positive Sorting through Performance Pay: Experimental Evidence from Pakistani Schools," *Working Paper*, 2021, p. 83.

Burke, Raymond V., Robert G. Oats, Jay L. Ringle, Leah O'Neill Fichtner, and Mary Beth DelGaudio, "Implementation of a Classroom Management Program with Urban Elementary Schools in Low-Income Neighborhoods: Does Program Fidelity Affect Student Behavior and Academic Outcomes?," *Journal of Education for Students Placed at Risk (JESPAR)*, 2011, *16* (3), 201–218.

Chen, Weiwei, Wayne A Grove, and Andrew Hussey, "The role of confidence and noncognitive skills for post-baccalaureate academic and labor market outcomes," *Journal of Economic Behavior & Organization*, 2017, *138*, 10–29.

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan, "How does your kindergarten classroom affect your earnings? Evidence from Project STAR," *The Quarterly journal of economics*, 2011, *126* (4), 1593–1660.

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.

Cloney, Dan, Gordon Cleveland, John Hattie, and Collette Tayler, "Variations in the availability and quality of early childhood education and care by socioeconomic status of neighborhoods," *Early Education and Development*, 2016, *27* (3), 384–401.

Conti, Gabriella, James Heckman, and Sergio Urzua, "The education-health gradient," *American Economic Review*, 2010, *100* (2), 234–38.

Cooper, Harris, Barbara Nye, Kelly Charlton, James Lindsay, and Scott Greathouse, "The effects of summer vacation on achievement test scores: A narrative and meta-analytic review," *Review of educational research*, 1996, *66* (3), 227–268.

Cunha, Flavio and James Heckman, "The technology of skill formation," *American Economic Review*, 2007, *97* (2), 31–47.

_ , James J Heckman, Lance Lochner, and Dimitriy V Masterov, "Interpreting the evidence on life cycle skill formation," *Handbook of the Economics of Education*, 2006, *1*, 697–812.

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.

Dean, Emma Boswell, Frank Schilbach, and Heather Schofield, "2. Poverty and Cognitive Function," in "The economics of poverty traps," University of Chicago Press, 2019, pp. 57–118.

Deci, Edward L, "Effects of externally mediated rewards on intrinsic motivation.," *Journal of personality and Social Psychology*, 1971, *18* (1), 105.

Deming, David J, "The growing importance of social skills in the labor market," *The Quarterly Journal of Economics*, 2017, *132* (4), 1593–1640.

Deming, David J., "Five Facts about Human Capital," 2021. Draft from October 13, 2021.

Dewey, J, "Experience and education (Original work published 1938)," *John Dewey: the latter works*, 1938, *1939*.

Diamond, A. and K. Lee, "Interventions Shown to Aid Executive Function Development in Children 4 to 12 Years Old," *Science*, August 2011, *333* (6045), 959–964.

Dillon, Moira R, Harini Kannan, Joshua T Dean, Elizabeth S Spelke, and Esther Duflo, "Cognitive science in the field: A preschool intervention durably enhances intuitive but not formal mathematics," *Science*, 2017, *357* (6346), 47–55.

Dobbie, Will and Roland G Fryer, "Getting Beneath the Veil of Effective Schools: Evidence From New York City," *American Economic Journal: Applied Economics*, October 2013, *5* (4), 28–60.

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.

Endo, Toshio and Kazutaka Kogi, "Monotony effects of the work of motormen during high-speed train operation," *Journal of human ergology*, 1975, *4* (2), 129–140.

Figlio, David N, "Boys Named Sue: Disruptive Children and Their Peers," *Education Finance and Policy*, 2007, *2* (4), 376–394.

Finkbeiner, Kristin M, Paul N Russell, and William S Helton, "Rest improves performance, nature improves happiness: Assessment of break periods on the abbreviated vigilance task," *Consciousness and cognition*, 2016, *42*, 277–285.

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

Gneezy, Uri, John A List, Jeffrey A Livingston, Xiangdong Qin, Sally Sadoff, and Yang Xu, "Measuring success in education: the role of effort on the test itself," *American Economic Review: Insights*, 2019, *1* (3), 291–308.

Goldin, Claudia and Lawrence F Katz, *The race between education and technology*, harvard university press, 2010.

Heckman, James J and Tim Kautz, "Hard evidence on soft skills," *Labour economics*, 2012, *19* (4), 451–464.

_ , 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.

Heckman, James, Seong Hyeok Moon, Rodrigo Pinto, Peter Savelyev, and Adam Yavitz, "Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program," *Quantitative economics*, 2010, *1* (1), 1–46.

Hirshleifer, David, Yaron Levi, Ben Lourie, and Siew Hong Teoh, "Decision fatigue and heuristic analyst forecasts," *Journal of Financial Economics*, 2019, *133* (1), 83–98.

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

Kraft, Matthew A. and Manuel Monti-Nussbaum, "The Big Problem With Little Interruptions to Classroom Learning," *AERA Open*, 2021, *7*, 23328584211028856.

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.

Lepper, Mark R, "Motivational considerations in the study of instruction," *Cognition and instruction*, 1988, *5* (4), 289–309.

Levav, Jonathan, Mark Heitmann, Andreas Herrmann, and Sheena S Iyengar, "Order in product customization decisions: Evidence from field experiments," *Journal of Political Economy*, 2010, *118* (2), 274–299.

Meuter, Renata FI and Philippe F Lacherez, "When and why threats go undetected: Impacts of event rate and shift length on threat detection accuracy during airport baggage screening," *Human factors*, 2016, *58* (2), 218–228.

Mijović, Pavle, Vanja Ković, Ivan Mačužić, Petar Todorović, Branislav Jeremić, Miloš Milovanović, and Ivan Gligorijević, "Do micro-breaks increase the attention level of an assembly worker? An ERP study," *Procedia Manufacturing*, 2015, *3*, 5074–5080.

Mincer, Jacob, "Human capital and economic growth," *Economics of education review*, 1984, *3* (3), 195–205.

Mischel, Walter, *The marshmallow test: Understanding self-control and how to master it*, Random House, 2014.

Morrison, Frederick J, Matthew H Kim, Carol M Connor, and Jennie K Grammer, "The causal impact of schooling on children's development: Lessons for developmental science," *Current Directions in Psychological Science*, 2019, *28* (5), 441–449.

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.

Pianta, Robert C, Bridget K Hamre, and Susan Mintz, *Classroom assessment scoring system: Secondary manual*, Teachstone, 2012.

Rapport, Mark D, Sarah A Orban, Michael J Kofler, and Lauren M Friedman, "Do programs designed to train working memory, other executive functions, and attention benefit children with ADHD? A meta-analytic review of cognitive, academic, and behavioral outcomes," *Clinical psychology review*, 2013, *33* (8), 1237–1252.

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.

Robertson, Ian H, Tom Manly, Jackie Andrade, Bart T Baddeley, and Jenny Yiend, "Oops!': performance correlates of everyday attentional failures in traumatic brain injured and normal subjects," *Neuropsychologia*, 1997, *35* (6), 747–758.

Sala, Giovanni, N Deniz Aksayli, K Semir Tatlidil, Tomoko Tatsumi, Yasuyuki Gondo, Fernand Gobet, Rolf Zwaan, and Peter Verkoeijen, "Near and far transfer in cognitive training: A second-order meta-analysis," *Collabra: Psychology*, 2019, *5* (1).

Schultz, T Paul, "Education investments and returns," *Handbook of development economics*, 1988, *1*, 543–630.

Sievertsen, Hans Henrik, Francesca Gino, and Marco Piovesan, "Cognitive fatigue influences students' performance on standardized tests," *Proceedings of the National Academy of Sciences*, 2016, *113* (10), 2621–2624.

Simons, Daniel J, Walter R Boot, Neil Charness, Susan E Gathercole, Christopher F Chabris, David Z Hambrick, and Elizabeth AL Stine-Morrow, "Do "brain-training" programs work?," *Psychological Science in the Public Interest*, 2016, *17* (3), 103–186.

Smilek, Daniel, Jonathan SA Carriere, and J Allan Cheyne, "Failures of sustained attention in life, lab, and brain: ecological validity of the SART," *Neuropsychologia*, 2010, *48* (9), 2564–2570.

Vos, Pauline and Wilmad Kuiper, "Predecessor items and performance level," in "Studies in Educational Evaluation," Vol. 29 2003, p. 191–206.

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

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

_ , "The World Bank World Development Indicators Database," 2015. https://data.worldbank.org/indicator/NY.GDP.MKTP.PP.CD, 2021-11-19.

Youden, W. J., "Index for rating diagnostic tests," *Cancer*, 1950, *3* (1), 32–35.

Zamarro, Gema, Collin Hitt, and Ildefonso Mendez, "When Students Don't Care: Reexamining International Differences in Achievement and Student Effort," *Journal of Human Capital*, 2019, *13* (4), 000–000.

Zelazo, Philip David, Clancy B Blair, and Michael T Willoughby, "Executive Function: Implications for Education. NCER 2017-2000.," *National Center for Education Research*, 2016.

9 Tables

Table I: Treatment Effects on School Performance

	Dependent Variable: Z-score of Student's Grades				
Subject:	All	Non-Math	Hindi	English	Math
	(1)	(2)	(3)	(4)	(5)
Panel A: Pooled Treatment Arms					
Cognitive Practice	0.0897**	0.0923**	0.0989**	0.0919**	0.0849**
	(0.0348)	(0.0386)	(0.0393)	(0.0407)	(0.0377)
Panel B: Disaggregated Treatment Arms					
Math Practice	0.0916**	0.0926**	0.0962**	0.0978**	0.0902**
	(0.0402)	(0.0445)	(0.0452)	(0.0471)	(0.0437)
Games Practice	0.0877**	0.0920**	0.1015**	0.0860*	0.0795*
	(0.0399)	(0.0444)	(0.0453)	(0.0469)	(0.0428)
p-value: Math Practice = Games Practice	0.9232	0.9899	0.9063	0.8013	0.7999
Observations	11320	7539	3780	3759	3781

Notes: This table reports treatment effects on students' regular school performance (mid-year and end-of-year grades) in the three core subjects of Hindi, English, and math.

- Observations are at the student-subject-semester level. The dependent variable is the standardized z-score of the student's grade.
- "Cognitive Practice" denotes receiving either treatment. "Math Practice" and "Games Practice" denote the Math or Games sub-treatments, respectively.
- Col. 1 includes all three subjects. Col. (2) restricts to English and Hindi, and Cols. (3)-(5) present each subject separately.
- All regressions include class section fixed effects and a linear control for baseline school performance. Standard errors clustered by student. * p<0.10, ** p<0.05, *** p<0.01.

Table II: Treatment Effects on Performance Declines

	Dependent Variable: 1[Question Correct]						Dep. Var.: Avg. Score	
	Test Subject							
	All	All	Non-Math	Math	Listening	Ravens	All	Non-Math
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Pooled Treatment Arms								
Cog. Practice x Deciles 6-10	0.0131***							
	(0.0048)							
Cog. Practice x Deciles 2-5	0.0079							
	(0.0049)							
Deciles 6-10	-0.0445***							
	(0.0038)							
Deciles 2-5	-0.0113***							
	(0.0037)							
Cog. Practice x Predicted decline		0.0925***	0.0770***	0.1084**	0.0709**	0.0876*		
		(0.0293)	(0.0290)	(0.0440)	(0.0341)	(0.0454)		
Cog. Practice	-0.0012	-0.0035	-0.0000	-0.0091	0.0013	-0.0020	0.0085*	0.0093*
	(0.0061)	(0.0062)	(0.0064)	(0.0094)	(0.0070)	(0.0101)	(0.0050)	(0.0052)
Panel B: Disaggregated Treatment Arms								
Math Practice x Deciles 6-10	0.0131**							
	(0.0055)							
Games Practice x Deciles 6-10	0.0132**							
	(0.0054)							
Math Practice x Deciles 2-5	0.0027							
	(0.0058)							
Games Practice x Deciles 2-5	0.0132**							
	(0.0057)							
Math Practice x Predicted decline		0.0962***	0.0933***	0.0979*	0.0960**	0.0982*		
		(0.0337)	(0.0345)	(0.0501)	(0.0402)	(0.0537)		
Games Practice x Predicted decline		0.0892***	0.0603*	0.1204**	0.0456	0.0762		
		(0.0336)	(0.0337)	(0.0511)	(0.0397)	(0.0524)		
Math Practice	0.0019	-0.0027	-0.0034	0.0004	-0.0021	-0.0065	0.0104*	0.0077
	(0.0069)	(0.0071)	(0.0074)	(0.0106)	(0.0080)	(0.0119)	(0.0057)	(0.0059)
Games Practice	-0.0043	-0.0043	0.0034	-0.0191*	0.0047	0.0028	0.0066	0.0109*
	(0.0071)	(0.0072)	(0.0075)	(0.0110)	(0.0083)	(0.0118)	(0.0058)	(0.0061)
p-value: Math Decline = Games Decline		0.8309	0.3559	0.6522	0.2278	0.6899		
Control Decline	0.12	0.12	0.05	0.18	0.06	0.03		
Observations	329268	329268	129115	200153	66932	62183	14692	9151

Notes: This table examines the treatment effect on performance decline in the listening, Ravens and math tests.

- Observations are at the student-test-question level in Cols. (1)-(6) and the student-test level in Cols. (7)-(8). Question item order was randomized across students. The dependent variable is a binary indicator for whether the question is correct in Cols. (1)-(6) and the student's average probability of getting a question correct across the test in Cols. (7)-(8).
- "Cog. Practice" denotes receiving either treatment. "Math Practice" and "Games Practice" denote the Math or Games sub-treatments, respectively. "Deciles 2-5" and "Deciles 6-10" are binary indicators for if the question appears in the given decile range. "Predicted Decline" is defined at the quintile-school level as the difference in the percent of questions correct in the first quintile minus the given quintile, and varies by school. In Col. (2), the coefficient on "Predicted Decline" is -0.15 ($p < 0.001$) (omitted from table for space).
- Cols. (1) and (2)-(6) correspond to the specification in Equations 1 and 3, respectively. Cols. (1), (2), and (7) estimate treatment effects for all three tests pooled. Cols. (3) and (8) show effects for the non-Math tests (listening and Ravens); Cols. (4)-(6) show effects for the Math, Listening, and Ravens tests, respectively.
- All regressions control for class section and test version id fixed effects, and a linear control for the student's baseline average score. Cols. (1)-(6) also include question fixed effects and a linear control for the fraction of students in the same school who got the question correct, computed restricting to the control group only and excluding the student's own observation.
- Standard errors are clustered by student. * p<0.10, ** p<0.05, *** p<0.01.

Table III: Persistence of Treatment Effects on Performance Declines

	Dependent Variable: 1[Question correct] Definition of Treat Variable			
	Cognitive Practice		Math Practice	Games Practice
	(1)	(2)	(3)	(4)
Cog. Practice x Deciles 6-10 x Endline	0.0144***			
	(0.0052)			
Cog. Practice x Deciles 6-10 x Follow-up	0.0110			
	(0.0102)			
Cog. Practice x Predicted decline x Endline		0.0921***	0.1071***	0.0773**
		(0.0305)	(0.0350)	(0.0351)
Cog. Practice x Predicted decline x Follow-up		0.0917**	0.0911*	0.0923*
		(0.0431)	(0.0505)	(0.0493)
Dep. Var. Mean	0.47	0.47	0.47	0.46
F-test p-value: Diff. of 2 coefficients = 0	0.7658	0.9935	0.7529	0.7624
Observations	329268	329268	219260	217142

Notes: This table examines the persistence of treatment effects on performance declines in the listening, Ravens and math exams.

- Observations are at the student-test-question level. Question item order was randomized across students. The dependent variable is a binary indicator for whether the question is correct.
- "Cog. Practice" denotes receiving a treatment, where the column header describes which treatment arm is being tested. Cols. (1) and (2) pool both sub-treatments, and Cols. (3) and (4) present the "Math Practice" and "Games Practice" separately.
- "Deciles 6-10" is a binary indicator that equals one if the question appears in the second half of the test. "Endline" is a binary indicator that equals 1 if the testing round is during the main experimental period (December or February rounds). "Follow-up" is a binary indicator that equals one if the test is administered during the follow-up round, 3 to 5 months after the end of the intervention. "Predicted Decline" is defined at the item quintile-school level as the difference in the percent of questions correct in the first quintile minus the given quintile, and varies by school.
-
- P-values of an F-test of equality of coefficients is presented at the bottom of the table.
- All regressions contain question, class section, and test version fixed effects, a linear control for baseline average score, and a linear control for the fraction of students in the same school who got the question correct, computed restricting to the control group only and excluding the student's own observation. Standard errors are clustered by student. * p<0.10, ** p<0.05, *** p<0.01.

Table IV: Measures of Sustained Attention

	Dependent Variable: Z-score Test Subject			
	Pooled	SART	Symbol Matching	Pooled
	(1)	(2)	(3)	(4)
Cognitive Practice	0.0483**	0.0672**	0.0357	
	(0.0203)	(0.0306)	(0.0242)	
Sub-treatments:				
Math Practice				0.0538**
				(0.0236)
Games Practice				0.0427*
				(0.0236)
p-value: Math Pratice = Games Practice				0.6411
Observations	9699	3895	5804	9699

Notes: This table examines the treatment effect on traditional measures of attention drawn from the psychology literature: the Sustained Attention to Response Task (SART, the standard measure of sustained attention) and a symbol matching persistence task.

- Observations are at the student-test level.
- The outcome variable is the standard metric in the psychology literature: the z-score of the students average true positive rate minus false positive rate across the test, winsorized at the 99th percentile. True positive rate is the fraction of stimuli in the test correctly identified by the student; false positive rate is the fraction of stimuli chosen (i.e. space bar presses in SART) where there was no actual target stimuli.
- "Cognitive Practice" denotes receiving either treatment. "Math Practice" and "Games Practice" denote the Math or Games sub-treatments, respectively.
- Cols. (1) and (4) pools across the two tasks. Col. (2) and (3) respectively present the SART and Symbol Matching results separately.
- All regressions contain class section fixed effects and a linear control for the student's average baseline performance on the two tasks. Standard errors are clustered by student. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table V: Attentiveness in the Classroom

	Dependent Variable: Z-score				
	Pooled	Task Completion	Response to stimuli	Physical signs (Reverse sign)	Pooled
	(1)	(2)	(3)	(4)	(5)
Cognitive Practice	0.0828***	0.0813	0.1260**	0.0380	
	(0.0319)	(0.0523)	(0.0617)	(0.0549)	
Sub-treatments:					
Math Practice					0.1075***
					(0.0366)
Games Practice					0.0577
					(0.0368)
p-value: Math Practice = Games Practice					0.1704
Observations	1197	1197	1196	1195	1197

Notes: This table examines the treatment effect on student attentiveness during a class session, measured using treatment-blind observers rating students on three components adapted from the Vanderbilt ADHD diagnostic teacher rating scale.

- Observations are at the student level and were conducted for a subset of class sections across all schools. In each column, the dependent variable is the standardized z-score of the given rating scale component (denoted at the top of the column).
- The three components of the rating scale are: i. *Task completion*, whether students completed two simple instructions given by their teacher, ii. *Response to stimuli*, response to auditory stimuli (noticing and attending to new sound during class), and iii. *Physical signs*, their physical signs of inattention (e.g., fidgeting, looking out the window), which is presented with a reverse sign, so a higher number corresponds to more attentiveness in each of the three components. The *Pooled* measure is a simple average of the z-scores for the individual components within the scale.
- "Cognitive Practice" denotes receiving either treatment. "Math Practice" and "Games Practice" denote the Math or Games sub-treatments, respectively.
- Cols. (1) and (5) pools across the three components. Cols. (2)-(4) present each of the three components separately.
- All regressions include class section and observer fixed effects, and controls for students' location within the classroom. There is no baseline measure for these outcomes. Robust standard errors are in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table VI: Effect of Incentives on Test Performance

	Dep. Var.: 1[Question correct]		
	(1)	(2)	(3)
Incentive	0.0934**	0.1604**	0.1336**
	(0.0464)	(0.0569)	(0.0579)
	[0.025]	[0.016]	[0.018]
Incentive x Cog. Practice	-0.0270	-0.0478	-0.0374
	(0.0334)	(0.0466)	(0.0443)
	[0.437]	[0.389]	[0.443]
Incentive x Dec. 6-10		-0.0592	
		(0.0412)	
		[0.166]	
Cog. Practice x Dec. 6-10 x Incentive		0.0181	
		(0.0414)	
		[0.732]	
Incentive x Predicted Decline			-0.3498*
			(0.2002)
			[0.066]
Cog. Practice x Predicted Decline x Incentive			0.0937
			(0.1854)
			[0.673]
Dep. Var. Mean	0.51	0.51	0.51
Observations	11515	11515	11515

Notes: This table tests whether offering students an incentive for their performance on declines tests affects the rate of performance declines, and whether the effect of incentives varies with Cognitive Practice treatment status.

- Observations are at the student-test-question level. Question item order was randomized across students. The dependent variable is a binary indicator for whether the question is correct.
- "Incentive" is a binary indicator denoting if the student was randomized to receive an incentive (a toy) for higher test performance. "Cog. Practice" denotes receiving either treatment. "Deciles 6-10" is a binary indicator that equals one if the question appears in the second half of the test. "Predicted Decline" is defined at the item quintile-school level as the difference in the percent of questions correct in the first quintile minus the given quintile, and varies by school.
- Cols. (2) and (3) correspond to the specification in Equations 1 and 3, respectively, but add a set of interactions with "Incentive." All main and interaction effects were included in the regression; the table only displays the coefficients that show the treatment effects of the randomized incentives (i.e. the Incentive level effect and interactions with the Incentive dummy).
- Data is restricted to the listening and Ravens Matrices tests in a single (April) round of data collection—the sample in which the incentive randomization was conducted. N=703 students, 26 grade*school*test clusters.
- All regressions contain question, class section, and test version fixed effects, a linear control for the student's baseline average score, and a linear control for the fraction of students in the same school who got the question correct, computed restricting to the control group only and excluding the student's own observation.
- OLS standard errors, clustered at the test-school-grade level (the unit of randomization for the incentive treatment) are reported in parentheses. Randomization inference p-values are reported in square brackets; for each draw, both cognitive practice and incentive receipt was re-randomized.

Table VII: Effect of an Additional Year of Schooling on Performance Declines

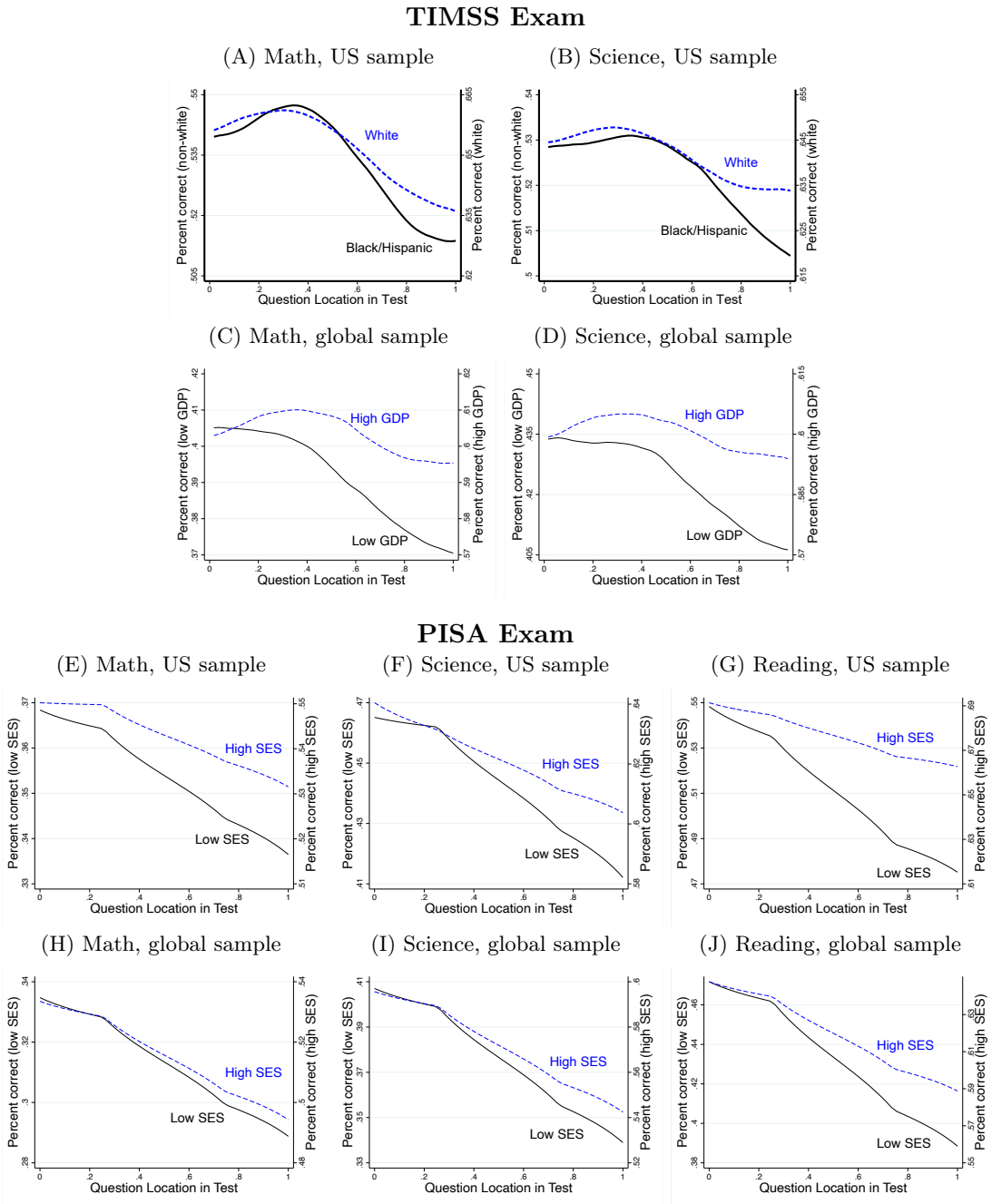
Dimension of Quality:	**Dependent Variable: 1[Question Correct]**							
			School Quality		Class Pedagogy		Independent Practice Time	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Yrs of Schooling x Predicted Decline	0.309**	0.337**	-0.0132	0.0103	0.0458	0.0664	0.0320	0.0547
	(0.143)	(0.144)	(0.183)	(0.184)	(0.158)	(0.160)	(0.177)	(0.178)
Yrs of Schooling x Predicted Decline x Higher Quality			0.223**	0.222**	0.219*	0.219*	0.188**	0.185**
			(0.105)	(0.104)	(0.115)	(0.113)	(0.0920)	(0.0911)
Dep. Var. Mean	0.59	0.59	0.59	0.59	0.59	0.59	0.59	0.59
Observations	276043	276043	276043	276043	276043	276043	276043	276043
Running variable func. form	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic

Notes: This table reports the effect of an additional year of school on student performance declines on exams using data from a sample of 5,353 9-11 year olds in Pakistan (Brown and Andrabi, 2021).

- Observations are at the student-test-question level. Question order was randomized. The dependent variable is a binary indicator for whether the question is correct.
- "Yrs of Schooling" is instrumented with the kindergarten entrance cut-off, controlling for birth month (the F-statistic on the first stage is 15.9). "Predicted Decline" is defined at the item decile-grade-subject level as the difference in the percent of questions correct in the first decile minus the given decile.
- Cols. (1)-(2) present the results of an additional year of school across all schools in the sample.
- Cols. (3)-(8) show heterogeneity in the effect of an additional year by different aspects of school/class quality (denoted in the column header). Quality measures are based on the CLASS rubric, comprised of 12 components such as classroom climate, time on task, time in independent practice, and use of higher-order thinking skills (Araujo et al. (2016); Pianta et al. (2012); see Appendix B for details). *School Quality* is the school's average score on the 12 components of the CLASS rubric (Cols. 3-4). *Class Pedagogy* captures the average score on all 12 components in the student's current grade level (Cols. 5-6). *Independent Practice Time* captures the quantity and quality of time students spend working independently on cognitively challenging material (Cols. 7-8). In each column, "Higher Quality" is a linear variable denoting quartile rank in the given quality dimension, with a value of 0 for the bottom quartile up to 3 for the top quartile.
- Odd-numbered columns include birth month as a linear control, and even-numbered columns include the quadratic of birth month as well. Standard errors are clustered by student. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

10 Figures

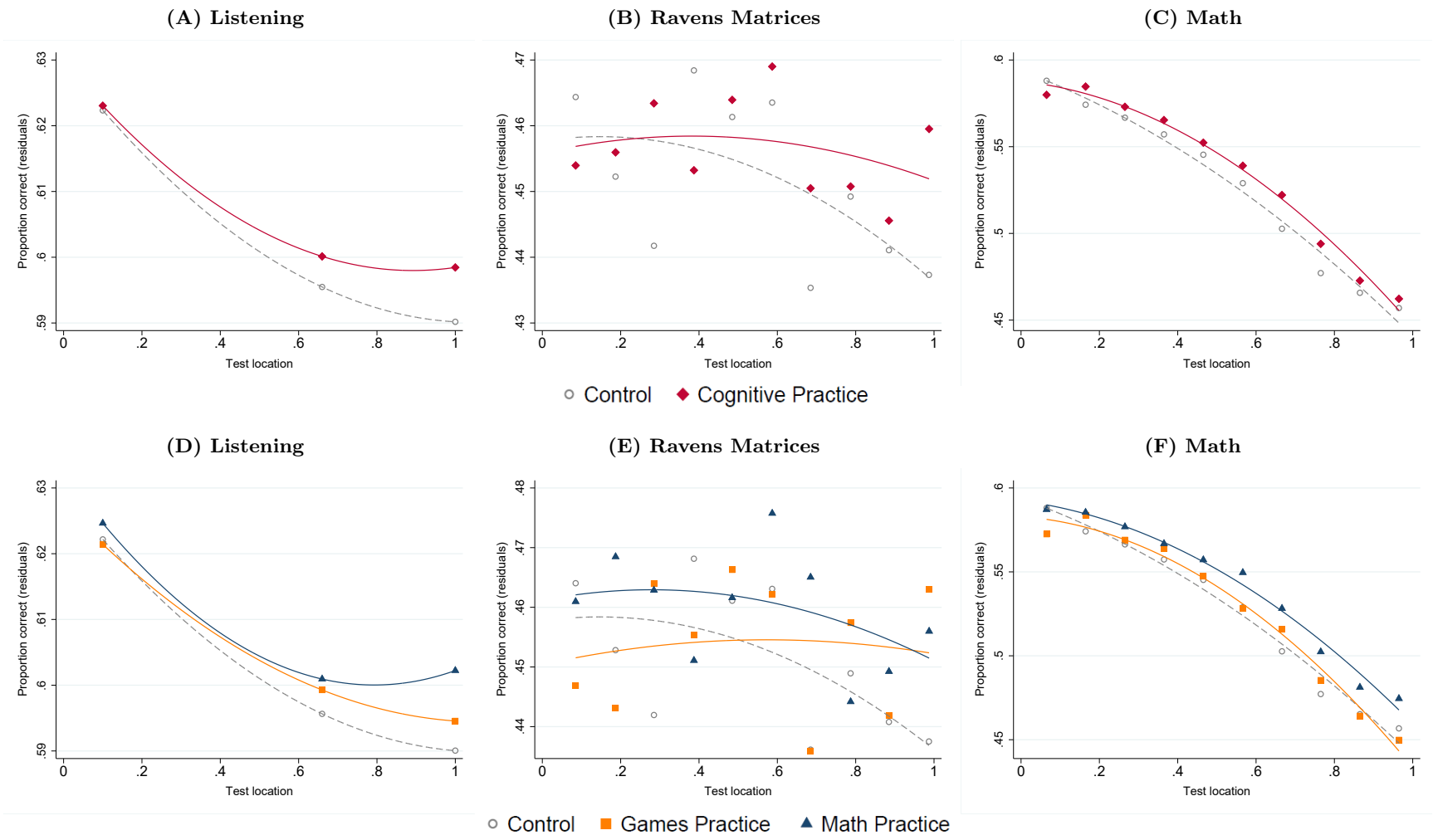
FIGURE I: Performance Declines in Achievement Tests



Notes: The figures show student performance over the length of the TIMSS and PISA tests.

- TIMSS (PISA) is administered to fourth graders (15 year-olds) in more than 50 countries. All subjects administered are presented here. Observations are at the student-question level.
- For TIMSS, question order is block randomized within each test subject; graphs plot residuals after removing question fixed effects. For PISA, randomization is across test subjects (with 4 randomization blocks per exam); we remove question block fixed effects.
- The x-axis is "Question location in test", which denotes where in the exam the question item appeared normalized on a scale of 0 to 1 (i.e. question number within subject in TIMSS and question block number across the exam in PISA). The y-axes plot the average score (i.e. percent answered correctly) for each question location on the test.
- The plots display the smoothed values of a kernel-weighted local polynomial regression, with a bandwidth of 0.15 for TIMSS and larger bandwidth of 0.33 for PISA (due to the smaller number of randomization blocks).
- The intercept of the y-axis varies by group—with more (less) advantaged students on the left (right) axis. In the TIMSS US sample (A-B), relative advantage is proxied by race (white and non-white, respectively). In the TIMSS global sample, these differences are proxied by the top (bottom) quartile of GDP/capita (C-D). In the PISA data (E-J), high (low) SES is proxied by the top (bottom) quartile of the ESCS measure, an index capturing parental income, occupation, and education.

FIGURE II: Performance Declines on Experimental Tests by Treatment Group

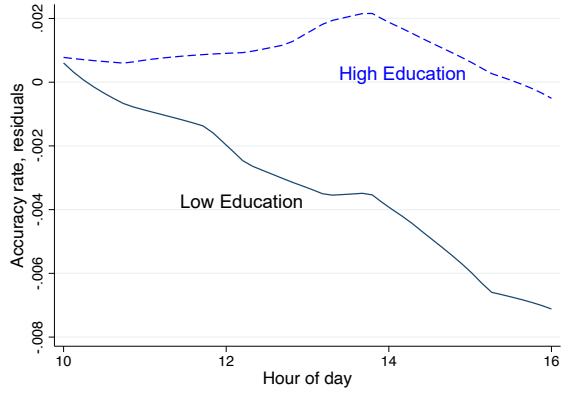


Notes: This figure plots declines in performance over the course of three tests: listening, Ravens Matrices, and math.

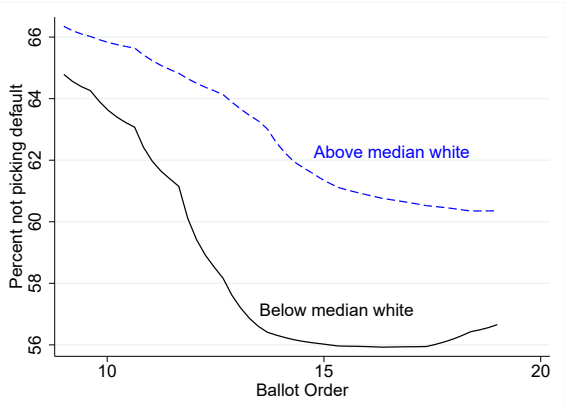
- Question order is randomized in each test. Observations are at the student-test-question level.
- Each panel displays a binscatter plot, where the x-axis is the percent location of the test, grouped in deciles (where 0.1 is the first decile of the test and 1 is the last decile), and the y-axis is the proportion of students who answer the question correctly. Data is residualized to remove question and test version fixed effects. All plots are overlaid with a quadractic best-fit curve by experimental arm.
- The top row plots the Control group (dashed gray, open circles) versus any Cognitive Practice (solid red, diamonds). The bottom row plots the Control (dashed gray, open circles) and each sub-treatment—Math arm (solid blue, triangles) and Games arm (solid orange, squares).
- For ease of interpretation of decline magnitudes, in each plot, the decile 1 control group mean is added to all residuals.
- Table II presents the full set of corresponding treatment effects estimates.

FIGURE III: Cognitive Endurance among Adults

(A) Worker accuracy



(B) Cognitive effort in the voting booth



Notes: These figures show performance over the length of two tasks (data entry and voting) by proxies for socioeconomic status.

- Panel A plots declines in entry accuracy among full-time data entry workers over the course of the work day.
  - Data are from Kaur et al. (2015). The sample is 8,382 worker-hours of data entry (90 workers).
  - The x-axis is the hour of the day, and y-axis is the accuracy rate (proportion of fields entered with no errors).
  - Relative advantage is proxied by a high school education (corresponding to the median split of education in the sample).
  - Data are residualized after removing worker fixed effects. 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.
- Panel B plots declines in active decision-making while voting in elections.
  - Data are from Augenblick and Nicholson (2015) and the United States census. Ballot item order in the voting data is quasi-random.
  - The x-axis is the location of an initiative (Proposition 35, the example provided in (Augenblick and Nicholson, 2015)) on the ballot, and the y-axis is whether the voter selects a choice other than the default option.
  - Above (below) median white denotes polling precincts where the fraction of White non-Hispanic residents is above (below) the median.