COGNITIVE ENDURANCE AS HUMAN CAPITAL*

Christina Brown‡ Supreet Kaur

Geeta Kingdon Heather Schofield

November 11, 2021

Abstract

Schooling may build human capital not only by teaching academic skills, but by expanding the capacity for cognition itself. We explore this hypothesis with a focus 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 a variety of 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 psychology 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. *JEL Codes*: I24, I25.

*We thank Ned Augenblick, David Deming, Ernst Fehr, Caroline Hoxby, Lawrence Katz, Patrick Kline, Matthew Kraft, Sendhil Mullainathan, Jesse Rothstein, Andrei Shleifer, and Christopher Walters for helpful comments and discussions. We thank Pixatel for 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, and Isadora Frankenthal, Medha Aurora, Joaquin Fuenzalida, Pranay Kapoor, 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.

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

‡Corresponding author. Address: 1126 E. 59th St., Chicago, IL 60637. Phone:(408) 835-8494, Fax:(773) 702-8490

1 Introduction

A large body of work documents far-reaching, persistent benefits of schooling (Bowles and Gintis, 1976; Goldin and Katz, 2010).[1] While it's clear that schooling affects cognitive abilities, the pathways through which it does so are less well understood. Researchers have long recognized that its role in shaping cognition could go beyond teaching academic content or skills. Schooling may expand the mind's capacity for cognition, for example by expanding our fundamental capacity to engage in sustained effortful thinking (e.g. Dewey, 1938; Morrison et al., 2019). This constitutes a more expansive view of how education shapes general human capital.

In this paper, we focus on one specific feature of schooling: formal education engages students in effortful thinking for continuous stretches of time.[2] 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—suggesting the possibility that schooling could play a role in its development.

To motivate the empirical relevance of these ideas, we begin with a set of illustrative examples. Using data from PISA and TIMSS, two prominent global academic achievement tests, we examine a key prediction of limited cognitive endurance: performance will decline over time during an intellectually effortful task. Each test has the feature that question order is block randomized across students, with ample time for students to finish the test. This enables us to examine how likely students are to get a given question correct when it appears earlier in the test (when they are still fresh) versus later in the test (when they may be more cognitively fatigued).

In Figure I, we plot performance over the length of these exams, separately by each test subject and by geography. In each of these ten plots, we consistently find two stark patterns. First, in line with cognitive fatigue, performance declines markedly over time: students are more likely to get a given question wrong if it appears later in the test. Note that such fatigue effects are not unique to academic tests; as we discuss below, they appear in myriad settings, from paramedics at work to voters at the ballot box. Second, performance declines are considerably more severe for students with lower socioeconomic status (SES). In the US, Black and Hispanic students show 72% more decline than White students—accounting for 10% of the total White/Non-white test score gap.

[1] This includes long-term effects on income, health, crime, and social well-being (e.g. Heckman et al., 2006b; Lance, 2011; Deming, 2011), as well as aggregate economic growth (Goldin and Katz, 2010; Acemoglu and Autor, 2012).

[2] Schooling is comprised of many important facets, and these may map to and affect various mental abilities. Our paper highlights one specific pathway among these, but does not diminish the potential relevance of others.

There is similar heterogeneity by wealth in the global sample. These motivating examples are, of course, only suggestive. In our experiment, we provide more carefully controlled measurement of cognitive fatigue.

While there may be many contributing factors for these SES differences, one potential interpretation is that cognitive endurance could be malleable—with the potential to be shaped through practice. Time use data from TIMSS teacher surveys suggest that schooling may be relevant for such training: both globally and in the US, higher SES schools allocate more time to independent focused practice. In other words, richer students spend more time in effortful thinking on their own during the school day.[3] In schools with such pedagogy, we also see fewer performance declines over time—even after controlling for wealth. While only correlations, these stylized patterns point to the potential for practice at school to affect cognitive endurance.

To test whether cognitive endurance is indeed malleable, we conduct a field experiment in a setting where the time spent in focused cognitive activity is limited: low-income primary schools in India. As is common in many such low-income environments, students spend little time on focused practice in school or at home. With the exception of exams, it is rare for them to sit and concentrate for 10-15 minutes at a time without distractions. We conduct our experiment with 1,636 low-income Indian students in grades 1-5.

We randomize a subset of students to receive sustained cognitive practice—engaging them in intellectually challenging content during the school day. 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: focused activity within the context of academic learning. 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. There is absolutely no academic content, such as numbers or letters, present at any point in these games—providing a pure test of our mechanism. For these treatments to be effective, the content must be difficult so that concentration is effortful, but also sufficiently engaging to retain student participation. To achieve this balance, and to overcome the hurdle of heterogeneous student ability in each class, we deliver each sub-treatment on simple tablet applications—enabling students to receive content appropriate to their skill level.[4]

The control group receives a status-quo math "study hall" period. As is standard in this setting, control students are provided with a small number of math problems copied from the chalkboard and can spend the remainder of the study hall session as they'd like. This results in little effective time spent in cognitive practice. Students are randomized at the individual level to either the control

[3]This may be due, at least in part, to the fact that facilitating such practice is more difficult in the crowded, more disruptive environments of lower-income schools (see, e.g., Burke et al., 2011; Kraft and Monti-Nussbaum, 2021).

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

group or one of the two cognitive practice sub-treatments. The experiment is implemented during study hall or an elective period 1-3 times per week between August and January, with practice sessions typically lasting about 20 minutes at a time. In total, treated students receive 10-20 hours of additional cognitive practice.

We examine impacts on two sets of outcomes. First, we test for improvements in cognitive endurance—as measured by the rate at which performance declines. Second, we test for impacts on students' school performance, in subjects unrelated to the content of the treatments. This validates whether the ability developed through the treatments is relevant for field behavior. To better understand mechanisms, we augment our findings with traditional psychology measures of attentional capacity, and a supplemental exercise on the role of motivation.

To test for effects on cognitive endurance, we measure impacts on the rate of decline in performance in three distinct domains—listening, Raven's Matrices (IQ), and mathematics—allowing us to test for broad-based impacts. For example, for listening, 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 not only captures an important input into learning in school, the content of this test is completely unrelated to the treatments: there is no sense in which they required students to practice listening. In each domain, students take a test with randomized question order and ample time to finish—enabling us to cleanly identify performance declines over time.

Consistent with finite cognitive endurance, in each of the three domains, control students exhibit significant fatigue effects: the probability of getting a question correct declines by 12% from the beginning to the end of a test on average. In line with our predictions, the treatments reduce the severity of these fatigue effects in each of the three domains we test. On average, cognitive practice mitigates the rate of performance decline across these domains by 21.6% (p=0.006), with similar average effects across the Math arm (21.2%, p=0.021) and Games arm (21.9%, p=0.015). Applying these gains to low SES groups in the TIMSS data would cut the gap in performance declines between high and low-income countries by 35%, or between Black and White students in the US by 38%.

In addition, the treatments have little impact in the beginning of the tests when students are still mentally fresh (e.g. the first quintile)—for example, in the listening or Ravens Matrices tests.[5] Rather, treatment effects only emerge later in the test, when control students become more cognitively fatigued. This pattern is especially consistent with improved cognitive endurance. In addition, it helps distinguish our effects from confounding mechanisms that would raise performance across all questions—for example, increased confidence, motivation to try harder, or working memory.[6]

[5]This pattern is consistent with the fact that the treatments did not teach subject content in these tests. As one would expect, the Math treatment improved math test performance even at the start for challenging math questions.

[6]Under such channels, it is unclear why students should not try harder or perform better early in the tests also, versus only later in the tests. Mean control group performance in the first decile of the listening, Raven's, and math tests is roughly 50%, leaving ample scope for treatment effects at the start of the test. In Section 5.4, we also discuss other potential channels such as complementary parental inputs.

The improvements in cognitive endurance persist 3-5 months after the end of treatment activities—after students return from end of academic year vacations. We cannot reject that effects at this follow-up round are equal to those at the end of the intervention.[7] As a whole, our findings support the view that the treatments reduce the severity of cognitive fatigue in a broadly applicable manner.

This, in turn, could affect students' academic performance both by improving their ability to learn (e.g. sustaining focus longer while listening to the teacher, reading a textbook, or thinking through a challenging concept) and also by reducing performance declines on exams and assignments (conditional on the academic knowledge they have). We examine overall impacts on grades in students' regular school classes—enabling us to test whether cognitive practice affects students' normal field behavior and outcomes.

Each of the two sub-treatment arms improves students' regular school performance in the core academic subjects taught in all schools. On average, student grades improve 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 ($p = 0.025$). Since the treatments could not have directly taught students Hindi or English, this points to improvements in a generalized mental resource. In addition, these impacts are similar for both the Math and Games sub-treatments. Using a simple back of the envelope exercise described in Section 6, we estimate that changes in performance declines on assessments can account for 1/3 of the impact on grades, implying that 2/3 of the effects stem from increased learning.

Our treatment effects on school performance indicate that simply spending time in effortful thinking—*without learning any subject content*—improves traditional measures of human capital. Moreover, such thinking need not even be academic in nature: even the students who receive the Games sub-treatment do substantially better in their academic classes. These findings imply that receiving an education—through the experience of cognitive practice—could reinforce the process of human capital accumulation, even outside of teaching content.

Should we view endurance as operating through a cognitive channel, or through motivation? Finding that cognitive endurance is malleable constitutes an advance under either of these views. Moreover, because psychologists consider these channels to be inherently related, we do not attempt to draw a strong line between them.[8] This informs our choice of the more general term "endurance" to describe performance declines during cognitive tasks. However, to explore the forces driving our effects, we augment our core findings with three supplementary measures.

First, we test for impacts on attentional capacity using traditional measures from the psychology literature. This includes the canonical measure of sustained attention, the Sustained Attention to Response Task (SART), which captures focus via reaction times to stimuli. Cognitive practice

[7] This provides evidence for some persistence, but of course does not speak to longer horizons. Our ability to collect data for further follow-up was halted by the Covid pandemic, which led schools to stop operating and another has shut-down completely since the intervention. Note that, irrespective of their longevity, by demonstrating malleability, our results open the possibility that SES differences in attentional practice at home, school, or the workplace could perpetuate differences in cognitive endurance even in later years—a possibility that warrants further research.

[8] Sustained attention is viewed as an upstream requirement for exerting perseverance, self-control, and other behaviors that involve sustaining focus towards a goal (Chun et al., 2011; Mischel, 2014; Zelazo et al., 2016).

improves performance on traditional attention lab measures, with an average effect of 0.088 SD in the Math arm ($p = 0.040$) and 0.075 in the Games arm ($p = 0.085$).

In addition, we measure effects on classroom behavior, adapted from a diagnostic teacher rating scale used to assess attention. This captures students' attentiveness in class, rated by observers that are blind to treatment status. We also see improvements in this index, with an average effect of 0.117 SD in the Math arm ($p = 0.003$) and 0.070 SD in the Games arm ($p = 0.074$). These findings support the idea that cognitive practice may bolster students' ability to attend to, and therefore learn in , the classroom.

Third, we undertake an additional exercise to examine the potential role of motivation. For a subset of the declines tests, we randomize the chance to earn toys for higher test scores. This sharply increases test performance, even at the beginning of the test—indicating that performance is highly elastic to effort even when students are cognitively fresh. However, the incentives do not reduce the severity of performance declines—indicating that an internal drive to do better does not mitigate observed fatigue effects. This test may not capture all dimensions of motivation; but this, along with the positive evidence on attentional measures, suggests a likely role for cognitive improvements.

While effortful thinking is an inherent feature of formal education, our test relies on an outside intervention introduced into schools. As a complement to our experimental evidence, we examine whether the natural experience of schooling does indeed develop cognitive endurance. We exploit quasi-random variation in years of schooling, due to birthday cut-offs for school enrollment, to construct a suggestive test. Using supplementary data on elementary school students from Brown and Andrabi (2021), we first replicate the presence of large performance declines in academic tests. We then use a regression discontinuity approach to show that, conditional on student age, an additional year of schooling does indeed mitigate performance declines—at a magnitude about three times as large as the effects from our more limited experimental intervention. These effects are considerably stronger for better quality schools, and those that engage students in independent practice in class. In contrast, among the worst quality schools, an additional year of school produces no discernible improvement in cognitive endurance. This suggests that initial disparities in schooling quality, through their impact on core mental capacity, could exacerbate achievement gaps among students.

We conclude by examining the broader relevance of cognitive endurance among adults: costly production errors among full-time piece-rate data entry workers, and deterioration in decision-making among voters at the ballot box. In each case, we document substantial performance declines over time—over the work shift or further down the ballot—and show that declines are considerably more severe among those with lower socioeconomic status. While only suggestive, these patterns provides impetus for more work on socioeconomic differences in cognitive endurance.

Our paper contributes to two sets of literatures. First, we advance a growing body of work on cognitive fatigue effects, including decision fatigue. Recent studies document specific instances of performance declines in numerous field settings (e.g. Endo and Kogi, 1975; Levav et al., 2010; Danziger et al., 2011; Brachet et al., 2012; Augenblick and Nicholson, 2015; 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).[9] These papers show that fatigue effects are meaningful for high stakes behaviors—for example, whether a judge grants parole, if a proposition becomes law, or how well a student does on a standardized test. We augment this work by documenting that cognitive fatigue exhibits more quickly among lower socioeconomic status (SES) groups, and this partially accounts for performance gaps by SES, across a variety of settings.[10] This suggests, for 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 malleable and can be improved—advancing work in both economics and psychology.[11]

Second, this study furthers our understanding of how schooling builds general human capital. Research in the economics, education, and psychology literatures 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).[12] This argument is typically based, for example, on looking at the impacts of an additional year of schooling on diverse outcomes. We make three 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 the broad benefits of schooling, there has been less work unpacking the education black box: what exact features of schooling are relevant, and how do they engender particular skills? Understanding such specific pathways would enable targeted policies to improve varied dimensions of human capital. We provide the first empirical demonstration of one such pathway: we isolate a specific feature associated with formal education (sustained effortful thinking) and establish its causal impact on a specific mental capacity (cognitive endurance). Third, our results suggest that worse schools are less likely to inculcate this capacity. This offers

[9] 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. Some of these studies interpret declines as reflecting motivation rather than cognitive fatigue. It is of course not possible to distinguish these in observational data. See discussion above for the role of motivation in our experimental results.

[10] While many studies in the education literature examine performance declines, particularly in PISA, there has been limited work on SES heterogeneity. A notable exception is Borgonovi and Biecek (2016), who explore heterogeneity along various dimensions, including 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.

[11] 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. In this literature, attempts to "train" sustained attention have not found "far transfer"—improvements outside of the exact task or game that was practiced—likely due to small sample sizes (typically 10-40 individuals per arm). See Rapport et al. (2013) for a meta-analysis of programs training attention, Simons et al. (2016) for a broader review of the cognitive training literature, and Chun et al. (2011) for an excellent review of the psychology literature on attention. Our findings, such as on traditional psychology measures of sustained attention, advance this literature as well by demonstrating far transfer.

[12] Relatedly, a growing body of work demonstrates the importance of non-academic skills—such as higher order cognitive skills or non-cognitive skills—for worker productivity, underscoring that human capital is broad and multifaceted (e.g. Heckman et al., 2006a; Almlund et al., 2011; Chetty et al., 2011; Heckman and Kautz, 2012; Borghans et al., 2014; Chen et al., 2017; Deming, 2017, 2021).

a new channel through which educational disparities could handicap more disadvantaged children, widening achievement gaps. In addition, we document that just the practice of thinking itself equips students to 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 further work.[13]

2 Motivation and Background

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. This offers a key empirical implication: when a person is engaged in a task that requires intellectual resources, performance during that task will decline over time. Note that this definition does not inherently take a stance on the specific psychological mechanism that produces performance declines. We probe potential mechanisms within the context of our field experiment below. In this section, we highlight suggestive patterns in field data to motivate our experiment.

2.1 Performance Declines and Socioeconomic Status

To motivate the empirical relevance of cognitive endurance, we begin with a set of illustrative examples. We focus on the key prediction that performance will decline over time during effortful cognitive activity. We look for this prediction within two prominent global academic achievement tests: TIMSS and PISA. TIMSS is a math and science test administered in over 50 countries to fourth graders during the school day. Question order is block randomized within each subject, and students are given ample time during each 30-minute test subject exam, so that declines are not driven by changes in question difficulty or test completion.[14] Similarly, PISA is administered to 15-year olds globally, covering math, science, and language. The test has four 30-minute sections or blocks, and these are administered in random order across students. Once each 30-minute block ends, students must move onto the next block.

In Figure I, we plot performance over time. For TIMSS, we compare performance on a given question when it appears earlier in the test versus later, including question fixed effects. For PISA, we examine performance on a given 30-minute block when it appears earlier versus later in the exam, including block fixed effects. The figure shows the US sample in the top row and the global sample

[13] We also relate to studies that introduce new interventions during the school day to improve generalized skills such as mindset, patience, grit, or working memory (Bettinger et al., 2018; Alan and Ertac, 2018; Alan et al., 2019; Berger et al., 2020).

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

in the bottom row, separately for each test subject in the TIMSS and PISA tests, respectively. In each graph, the x-axis denotes the location in the test (as a percentage of the total test length), and the y-axes denote the average probability that the question was answered correctly.

In each of these 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 TIMSS exam, among low socioeconomic status students, the rate of performance decline is 16% in the global sample. Note that such performance declines are not unique to these tests. They also arise in other academic achievement tests, as well as myriad settings where the stakes are high—including among voters, paramedics, data entry workers, judges, financial analysts, airport baggage security inspectors, train operators, and consumers buying cars (Borghans and Schils, 2015; Balart et al., 2018; Augenblick and Nicholson, 2015; Brachet et al., 2012; Kaur et al., 2015; Danziger et al., 2011; Hirshleifer et al., 2019; Meuter and Lacherez, 2016; Edkins and Pollock, 1997; Levav et al., 2010). These motivating examples are only suggestive. However, the ubiquity of performance declines across domains supports the premise that cognitive endurance matters for economic outcomes. In our field experiment, we provide more carefully controlled measurement of cognitive fatigue effects.

Second, more disadvantaged students exhibit markedly stronger cognitive fatigue effects. Across each of the ten plots in Figure I, performance declines are more severe among lower SES students. For example, in the US, Black and Hispanic students show 72% more decline than White students; this difference in decline accounts for 9% of the total White/Non-white test score gap (Panels A-B and E-G). We see similar patterns by wealth globally (Panels C-D and H-J). Below, we document similar systematic heterogeneity by wealth among adults in behaviors outside of schooling.

2.2 Cognitive Practice and Schooling Environments

Good schooling engages students in effortful thinking for continuous stretches of time. For example, in many schools, this feature is explicitly incorporated into pedagogy: students are required to sit and independently work on academic problems on their own. In classroom time use data from the TIMSS teacher survey, the average student engages in some independent practice one out of every three school days. To varying degrees, other aspects of schooling—taking a test, reading a textbook, doing homework, possibly even listening to a lecture—may also engage students in effortful thinking for extended periods. In other words, school involves more than just learning content; the act of learning the content often requires periods of sustained concentration.

However, the degree to which students engage in sustained concentration varies across schools—and does so systematically by socioeconomic status. As an example, both globally and in the US, poorer students spend less time in focused independent practice during the school day (Figure II, Panels A-B). This amounts to 40% less independent practice among students in poorer countries compared to richer ones, or 10% less practice among more disadvantaged students in the US compared

to more advantaged ones. In addition, the environmental conditions faced by poorer students—more crowded classrooms, disruptions from peers, and less ability to focus on homework at home—may make it less likely that they can effectively engage in concentration, even when it is attempted (Kraft and Monti-Nussbaum, 2021; Figlio, 2007). For example, poorer students attend schools with considerably more disruptions during class (Figure II, Panels C-D). These environmental conditions could also discourage teachers from attempting to engage a class in focused practice work.[15]

Work in psychology hypothesizes that exposure to periods of effortful thinking could be consequential for "training" sustained attention (e.g. Chun et al., 2011). Such independent activity requires self-driven focus—as opposed, for example, to external stimuli that can capture your attention, such as listening to a lecture. Potentially consistent with this view, students who are exposed to more independent practice time in school exhibit much less steep performance declines over the length of the TIMSS exams (Appendix Table A.1, Col. 2). This relationship holds even controlling for income differences across students (Col. 3).[16] While these are simply correlations and therefore not causal, they provide motivational support for the possibility that exposure to independent practice could affect cognitive endurance.

Consequently, our experimental intervention increases the amount of time students spend solving cognitively challenging problems on their own. This approach reflects common practices already used in many wealthier schools. Of course, this does not negate the potential role of other activities, at school or home. Rather, our choice of independent practice as the basis for our design is informed by both the psychology literature, and the motivational patterns above.

3 Experimental Design

The primary goal of our study is to construct a field experiment to test whether spending time in effortful thinking expands cognitive endurance, with downstream effects on academic achievement. We supplement this with ancilliary tests on underlying mechanisms.

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, the teaching approach focuses on 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, when teachers

[15]For example, in our sample teachers cite these conditions, along with heterogeneous ability among students, as factors that prevent independent practice in class.

[16]We conduct this motivational analysis within the global sample, where there is greater power to examine these correlations due to the larger sample size.

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.1 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. In order to construct a robust test of our hypothesized mechanism, we use two different approaches for this cognitive practice—academic or non-academic—and compare this with a control arm:

1) **Treatment: Cognitive practice.** Students solve intellectually challenging problems.
    a) ***Math***: Students practice academic math problems.
    b) ***Games***: Students play cognitively demanding games, such as mazes and tangrams, with no academic content.
2) **Control: Study hall.** Students attend a status-quo study hall period, with limited cognitive practice.

The control group receives a status-quo math "study hall" period. As is standard practice in this setting, in this group, 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. marking exams). Students can decide whether to attempt the questions, and spend the remainder of the study hall session as they'd like. Because the work does not feel engaging, the size of the chalkboard limits the number of problems, and many students do not have the skill to attempt grade-level content, little effective time is spent in cognitive practice for most students. The math problems assigned to students are drawn from the same question bank as those in the Math sub-treatment arm.

The Cognitive Practice treatment is divided into two sub-treatments. The Math sub-treatment mimics what good schooling does: focused cognitive practice within the context of academic learning. Students solve a series of math problems on their own in each session. However, under our hypothesis, practicing any intellectually demanding task should improve cognitive endurance—regardless of whether students learn anything from it. This motivates the design of the Games sub-treatment,

which does not entail any academic learning or practice. Students play intellectually challenging games, like tangrams or Flow Free (see Appendix A.2). These games are chosen so that they: 1) contain absolutely no academic content, such as numbers or letters—providing a more pure test of our mechanism, and 2) require effortful thought to complete. Treated students receive about 20 minutes of focused practice per program class period, compared to 0-10 minutes for students in the control group (reflecting wide heterogeneity across students). This resulted in 10-20 hours of cognitive practice in the treatment arms on average (see details below).

For the cognitive practice treatments to be effective, they require an activity that is cognitively demanding so that concentration is taxing, but 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 each treatment on simple tablets—enabling students to receive content appropriate to their skill level. In each sub-treatment arm, the tablet software provides no instruction, only the practice of problems or games. Appendix figure A.3 show pictures of example classes implementing the treatment.

For the Math sub-treatment, we use the imagine Math software, developed by Pixatel. This displays math practice problems via a simple interface, with no graphics, animations, or other visual features (see Appendix figure A.2a). One problem appears on the screen at a a time; students are asked to solve the problem, and then select the correct answer on the tablet.[17] Depending on the student's performance, subsequent problems become easier or more challenging. Overall, we selected topics to prioritize practice rather than learning new content.

For the Games sub-treatment, we use simple games with limited animation downloaded from the Android app store. These should not be viewed as "fun" video games, but rather traditional stimulating puzzles and games delivered through a bare-bones tablet interface (see Appendix figure A.2b). The specific games were chosen to meet three criteria: 1) they should be dynamically adaptive to continue to challenge all students regardless of initial skill (so students do not get bored over time); 2) they should not be related to test outcomes we would measure later (e.g. no games with sound or listening were selected); 3) they should be challenging and require concerted effort, but still sufficiently engaging that students would work for an extended period. The final criteria relied heavily on piloting a variety of potential games and selecting those which appeared effective by visually judging the children's engagement.

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 student engagement in intellectually challenging material, while solving the practical challenge of heterogeneous ability. Consequently, we view this implementation approach as simply a convenient way to achieve our goal of increasing the amount of time in effortful thinking in this context. Below,

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

we assess whether this approach may have generated other impacts, such as increased motivation or confidence, which could explain our results. Finally, note that our main outcome measures are collected using traditional pencil-and-paper tests.

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 treatment groups for the duration of the study. Randomization was at the individual level, stratified by class section (i.e. classroom) and baseline math test scores.[18]

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 classroom 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 that due to the fixed transition cost across classrooms, the effective intervention time was roughly 20 minutes per session. Appendix Figure A.4 shows the timeline of the experiment. Each school dedicated 1-3 elective periods per week from August to January for the intervention. The number of sessions varied across weeks based on other activities such as festivals, planned assemblies, or exams, and across schools based on when the intervention began in the school and the schedule agreed upon with the school administration based on the number of free elective periods available. This resulted in 10-20 hours of cognitive practice in the treatment arms on average.

At each school, we placed three study staff members who were responsible for splitting students up into the correct intervention classroom, overseeing activities in these classrooms, and then returning students back to their normal class sections as a group.[19] 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. Across all experimental groups, they did not engage in any instruction during the practice sessions. They typically corrected other homework or did administrative tasks at a desk while students practiced—as is customary among teachers overseeing study hall periods in our setting.

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

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

[19] Having our own research staff implement the intervention during elective classes helped ensure study protocols were followed, such as no instruction or extra help beyond basic technical support (e.g. swapping out a malfunctioning tablet) to treatment students. This helps ensure a clean test of our research question.

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, in line with limited interaction and engagement with school activities among parents in our population. Because the randomization assignment was controlled by the study team and overseen by study staff, it was not possible for students to switch across treatment arms—as verified by our administrative data and random spot checks.

Finally, while not essential for our experiment design, we endeavored to keep the regular school teachers blind to students' treatment status. Treatment assignment rosters were never shared with teachers. 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), so that teachers would not have directly observed which students were in which group. This helps reduce concerns that teachers could systematically have treated students differentially in some way based on their knowledge of treatment status.[20]

We do not observe any imbalance in student covariates or baseline test scores (Table A.2). Of the 44 pair-wise t-tests comparing treatment arms, none are statistically significant. In addition, attrition was low – 11% for school administered exams, 3% for experimental exams – and balanced across experimental arms (Table A.3).

3.4 Outcome Measures and Mechanisms

Our outcome measures are summarized below and explained in more detail in the results section. We examine impacts on two primary outcomes: cognitive endurance as measured through performance declines, and students' academic performance. We supplement this with additional measures and tests to better understand mechanisms. Note that the outcomes below were pre-registered.[21] Across our measures, we test students in domains that are unrelated to the content they practice as part of the treatment arms—enabling us to draw conclusions about whether our results capture a change in core cognitive capacity.

3.4.1 Primary Outcomes

(i) ***Cognitive endurance: performance declines.*** We measure changes in cognitive endurance by estimating performance declines during intellectual activity. We test whether the treat-

[20] Of course it is possible that some teachers may have learned of treatment status for some individual students through conversation or by walking past the program classrooms.

[21] Specifically, we pre-registered the performance declines tests (in the subjects of listening, Raven's Matrices, and math), including the fact that we would be looking at effects on declines (i.e. slopes), and the the traditional psychology measures (SART and symbol crossing). We also pre-registered looking at 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.

ments mitigate the severity of performance declines across three unrelated domains—listening, IQ tests (Ravens Matrices), and math—over a 20-30 minute period. In each test, we randomize question order and allot ample time for test completion, allowing us to cleanly identify performance declines.

(ii) ***Academic achievement: school grades.*** Second, we examine students' regular school performance in their core academic subjects of Hindi, English, and math. For this, we use the end-of-term grades provided by the schools. This offers a direct test for whether the ability developed through the treatments is relevant for field behavior. In addition, these scores are intended to capture a combination of improved cognitive endurance on the exam itself as well as direct learning effects through improved attention in the classroom. Note that both the performance declines and school grades include domains that were not practiced during treatment sessions. For example, neither sub-treatment arm involved students practicing listening, and neither could have taught students Hindi or English. Consequently, examining impacts on non-math subjects provides a test for changes in a generalized, transferable ability.

The performance decline tests (outcome i) were administered during the school day at four times: Baseline (September), Mid-line (December), Endline (February), and Follow-up 3-5 months after the end of the intervention (April-September). All tests were administered via paper and pencil. Students in each class section cohort were tested together, so that each test batch had students from across the treatment arms. Certain tests were randomly not administered in all rounds due to time constraints.[22] For each outcome, we pool across testing rounds in tables and figures to present average impacts, unless otherwise stated.

School performance measures (outcome ii) were provided by all schools for the treatment period, and by a subset of schools for the year before treatment as a baseline measure. 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.

3.4.2 Mechanisms: Additional Outcomes and Tests

Outcome (i) above enables us to examine our primary goal: whether cognitive endurance is malleable. However, this does not shed light on whether we should understand endurance through a primarily cognitive lens, versus through a broader view of perseverance, which could include factors such as motivation. To better understand mechanisms, we augment the above with two additional sets of analyses. First, we examine measures of attentional capacity used in the cognitive psychology literature: canonical psychology lab measures, and assessments of classroom behavior. Second, we

[22] In addition, there was a clerical error in some of the April Ravens Matrices tests, which led to test modules that were up to 60 questions long and therefore unusable for looking at performance declines. These tests are excluded from the data analysis. The results are robust to including these tests in the analysis. Test modules included in the main analysis have, on average, 30 questions.

introduce supplemental variation using incentives to explore the role of motivation. Again, we provide an overview here, with more details when we present the results.

Traditional Psychology Measures. We assess students' attentional capacity as proxied by the canonical measure of sustained attention in cognitive psychology: the Sustained Attention to Response Task (SART) (Smilek et al., 2010). This provides an abstract context-free measure of whether students' ability to sustain focus has improved. We supplement this with a secondary lab measure used in psychology, a symbol matching task, which is an adapted version of a concentration endurance task.

Classroom Behaviors. To examine effects on classroom behavior, we adapt the Vanderbilt ADHD Diagnostic teacher rating scale, a commonly used assessment used by teachers to evaluate student behaviors, to our local environment. This measures students' attentiveness level in the classroom while a teacher is lecturing and while students are engaged in common classroom activities. Students' behavior is rated by treatment-blind observers.

Role of Motivation. To understand the potential relationship between motivation and cognitive endurance, we exogenously induce students to be more driven to perform well. Specifically, for a subset of the performance declines tests (outcome (i)), we randomize whether students are incentivized to perform better on the test: they receive increasingly desirable toys as their performance is higher up the distribution of scores. We use this approach so that students randomized to receive the toy incentives are motivated to work hard on the test, regardless of initial skill. We can then test whether this raises performance both early in the test (enabling us to check whether early performance is elastic to motivation), and the effect on performance declines (enabling us to check whether improved motivation mitigates performance declines). We randomize these incentives at the grade-school-exam level during one round of testing.

Alternate Mechanisms. The cognitive practice treatments could arguably boost performance through other channels unrelated to cognitive endurance. In Section 5 below, we discuss potential confounds—including changes in confidence, a desire to do well in school, or alternate cognitive mechanisms such as working memory.

Balance. The randomization was successful. We find no significant differences in demographic characteristics or baseline performance across experimental arms (Appendix Table A.2). In addition, attrition was relatively low: averaging 10% for school administered tests and 3% for the experimental exams used to capture cognitive fatigue. There is no differential attrition by treatment status for any outcome (Appendix Table A.3).

4 Results I: Cognitive Endurance

4.1 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, allowing us to look for broad generalizable impact:

(1) *Listening:* This task measures students' attentiveness while listening to a passage—mimicking an activity that is required in nearly all typical classroom settings. Using headphones, each student listens to a pre-recorded set of short simple stories. After each story, the student is asked simple factual questions about the content of the story, for example, "what color was the dolphin?" Each question is presented one at a time, with fixed time pacing between consecutive questions about a story, after which the next story begins. In order to avoid any concerns about literacy, answers are multiple choice and visual (e.g. in the above example, green, blue, black, and grey squares to denote the color of the dolphin).

(2) *Ravens Progressive Matrices:* This is a non-verbal multiple-choice test of reasoning in which the participant is asked to identify the element that completes a pattern in a figure (Raven, 1936, 2000). This test is 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.

(3) *Math:* A standard paper-and-pencil test of math problems. These tests are constructed to include a mix of both remedial and more grade-appropriate problems—enabling us to test for declines effects regardless of whether students are at grade-level.

For each test of the three tests, we randomize the order of questions across different test packet versions.[23] We then randomize the test packet version across students to enable student-level randomization of question order. Appendix Table A.5 and Appendix Figure A.5 verify the balance in test version and question difficulty by treatment status.

In addition, we ensure that students have sufficient time to finish the tests without time pressure. Consistent with this, nearly all students are able to respond to questions near or at the very end of the exam; the last question completed was on average 99.7%, 93.3%, and 99.3% of the way through the exam for the listening, math, and Ravens Matrices tests, respectively (Appendix Table A.6). Consequently, declines over time are not confounded by changes in question composition or non-completion. Below we verify that results are similar when restricted to only attempted questions.

[23]For the listening test, the question order randomization ensures that questions later in the test are not more likely to be based on content that appeared later in the story passage. This helps assuage concerns that performance on later question items could be related to the students' working memory instead of cognitive endurance.

Each set of tests is adapted to grade level. Students take only one test per day, so are cognitively fresh at the start of each test. All tests are conducted during the school day, and are interpreted by students as being regular school tests.

4.2 Performance Decline Patterns

Figure III plots performance on each test over time—separately for the control group and each sub-treatment arm. In each panel, the x-axis is the percent location of the test (where 0 is the beginning of the test and 1 is the final question of the test), and the y-axis is the proportion of students who answer the question correctly. The data is residualized to remove question fixed effects. Consequently, the plots can be interpreted as showing changes in average performance when the same question appears earlier in the test versus later.[24]

The solid black line displays control group performance in each plot. Across tests, students are 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 shows substantial declines in test performance over time: 18 p.p., 6 p.p. and 3 p.p. for math, listening, Ravens, respectively. Because test completion is high, note that these patterns persist even when restricting the data to only attempted questions (Appendix Figure A.7). This replicates the patterns seen in the TIMSS and PISA data documented above in Figure I—supporting the empirical relevance of cognitive fatigue.

Consistent with our hypothesis, cognitive practice mitigates performance declines. In the plots, the blue dashed line shows average performance for the Math sub-treatment, and the green long dashed line for the Games sub-treatment. On average, relative to the control group, being assigned to the treatment reduces the rate of decline in the second half of the test—by 21.2% in the Math arm and 21.9% in the Games arm.[25]

4.3 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)

[24] Because initial performance is similar and statistically indistinguishable across treatment arms (see Table I), initial levels are normalized to the control group mean in the first quintile of each test in order to more clearly visualize declines. Appendix Figure A.6 reproduces these plots without this normalization.

[25] See Section 4.4 below for an explanation; Table I, Col. (1) for the corresponding regression results.

fixed effects, which flexibly capture declines over time in the control group. The χ_{il} is a vector of question fixed effects. We also control for the student's baseline score.[26] In addition, we can run the above regression to separately estimate effects for each subtreatment, replacing the $CogPractice_s$ dummy with two separate dummies for the Math and Games practice sub-treatments. 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, by focusing on the second half of the test, 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 treatment effects occurs once the control group starts declining in performance. 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 of the test:[27]

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

Since some tests have a small number of questions (e.g. some Ravens tests have 10 questions only), we use quintiles as location bins to reduce noise.[28] 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 do over the course of a test in the absence of any intervention. We then test whether receiving Cognitive Practice mitigates the rate of expected performance decline:

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

[26] We also include 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, which captures question difficulty. In addition, when pooling across different test subjects, we allow the $CogPractice_s$ term to vary by test subject, to allow for different level effects across subjects. Finally, because the tests in higher grades had more questions (and therefore observations), we also weight by the number of questions in the test so that each student-test receives equal weight.

[27] This allows for differences in baseline ability across schools. Results are similar if we do not compute this measure separately by school, or use alternate approaches (see Table A.7).

[28] In our baseline specification, predicted declines were estimated within schools to allow for differences in decline patterns across schools. We omit the school index from the regression notation for simplicity. Results are robust to alternative predicted decline specifications (see Table A.7.

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 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 level effect: treatment effects at the start of the test before declines set in. To account for the fact that $PredictedDecline_l$ is estimated from baseline data, we bootstrap standard errors.

4.4 Impacts on Cognitive Endurance

Table I presents the results from both estimation approaches. Col. (1) presents estimates from Equation (1), pooling across subjects. Receiving Cognitive Practice improves average performance in the second half of the test by 1.29 percentage points (pp) ($p = 0.006$), corresponding to a 21.6% reduction in the amount of decline relative to the control group. 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 -0.0027 ($p = 0.45$). 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 reductions of 1.27 pp (21.2%, $p = 0.021$) and 1.31 pp (21.9%, $p = 0.015$), 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.

The estimation approach in Equation (3) gives a similar pattern of results, shown in Col. (2). receiving Cognitive Practice mitigates 9.27% of the expected decline over the test ($p = 0.001$).[29] 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 9.8% of the expected decline ($p = 0.003$), and the Games arm mitigates 8.8% of the expected decline ($p = 0.007$). As before, there is no discernible level effect; for example, the estimated impact of Cognitive Practice in the beginning of the tests, before declines set in, is -0.005 and insignificant (Panel A, Col. 2, $p = 0.412$).

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. Moreover, the treatments have no discernible impact at the start of the tests when students are still cognitively fresh. These patterns are especially consistent with cognitive endurance, and help distinguish our effects from mechanisms that would raise performance across all questions, such as improved confidence or a desire to try harder (discussed further Section 5 below).

The remaining columns in Table I disaggregate the results across subjects. Because each of the experimental arms had different levels of math exposure, we first show overall results excluding the math test in Col. (3). The results are similar to those in Col. (2). In Cols. (4)-(6), we report effects for each of the three test subjects separately. In Panel A, receiving Cognitive Practice reduces expected declines in math by 10.4% ($p = 0.015$), listening by 6.8% ($p = 0.040$), and Raven's Matrices

[29] In the control group, pooling across tests, the average decline from the first to the fifth quintile is 12 percentage points.

by 9.7% ($p = 0.031$). We cannot reject that these three treatment effect coefficients are equal. These findings indicate an improvement in cognitive endurance even in domains that are unrelated to what was practiced in the treatments, such as the listening test. This is consistent with broad impact, and supports the idea that cognitive endurance is a generalizable resource that is applicable in various activities.

Panel B Cols. (3)-(6) provides 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. As above, we see no evidence for any level effects from either treatment arm—particularly for the listening and Ravens tests.[30] 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 underpowered 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. Finally, we document that treatment effects are very similar when restricting analysis to attempted questions only (Appendix Table A.8), and are robust to alternate empirical specifications varying the controls included (Appendix Table A.9). We do not find any significant heterogeneity in treatment effect by student grade, gender, baseline average score or baseline decline in performance (Appendix Table A.10, Panel A).

The magnitude of these effects 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 among Black versus White students by 37.5%. As a whole, we take the findings in Table I as proof that cognitive endurance is malleable, and can be expanded through spending time in effortful thinking—regardless of whether it is academic or non-academic in nature.

4.5 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 between when students progress from one grade to the next. This enables us to examine effects after a break when students are not attending school—a time during which there is often substantial decay in academic learning Cooper et al. (1996); Alexander et al. (2007).

Table II tests for persistence. Cols. (1) and (2) show results for the pooled sample using each

[30]For the math test, we might expect some 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 some 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. 3, coefficient=-0.0165, $p = 0.12$). These patterns become stronger when we examine level effects for non-remedial math questions; for such questions, students in the math sub-treatment show a significant improvement in performance relative to the other experimental arms. However, since we see no such evidence for learning effects for the listening and Ravens tests, the results in Col. (4) offer an easily interpretable test for cognitive endurance changes.

of our two estimation strategies (based on Equations 1 and 3, respectively). In each column, the interaction term with the follow-up round dummy gives the differential effect of the treatment during the follow-up, relative to the effect during the main experimental period.[31] We see no evidence for a decline in treatment effects 3-5 months after the cognitive practice intervention ends: in each column, the interaction term is essentially zero and insignificant. For example, in Col. (2), being assigned to cognitive practice mitigates 9.3% of the predicted decline during the main intervention period ($p = 0.002$), in line with the results in Table I, Panel A, Col. (2). In addition, the interaction term for effects in the follow-up round is essentially zero and insignificant (coefficient of -0.0012, $p = 0.978$)—indicating no detectable change in treatment effects. We further decompose these results to examine each sub-treatment independently and 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 for the total effect of the treatment relative to the control group in the follow-up period, and generally find evidence of sustained increases.

This provides evidence for some persistence, though of course does not speak to persistence over longer horizons. Note that our ability to collect longer-horizon follow-up data was disrupted by the Covid-19 pandemic. The schools stopped operating during the pandemic, and one school has shut-down since the completion of our study. Whether we should expect persistence over longer periods is ambiguous. 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 II, then this could continually reinforce differences in cognitive endurance.

5 Channels of Impact: Mechanisms and Confounds

Our findings indicate that cognitive endurance is malleable and can be improved through training. This advances the literature on cognitive fatigue, and is arguably interesting regardless of the specific mechanism underlying endurance effects. However, this does not shed light on whether we should understand our effects as reflecting cognitive improvements, versus a broader view of perseverance that could include factors such as motivation. 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.

However, it is interesting to better understand the forces driving our effects. Consequently, before turning to effects on school performance, we augment our findings above with supplementary

[31] The prior analysis pooled these follow-up tests with the main midline and endline tests for power.

measures of mechanisms. We use these secondary measures to both provide positive support for our results on cognitive endurance, and to explore channels. Finally, in Section 5.4, we examine factors that may operate outside of cognitive endurance, such as confidence or memory.

5.1 Psychology Measures of Sustained Attention

We begin with traditional measures of attentional capacity used in the cognitive psychology literature.[32] The canonical measure for sustained attention—the ability to sustain cognitive thought toward a goal over time—is the Sustained Attention to Response Task (SART). 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 (Peebles and Bothell, 2004).[33,34] If a student has lost focus (e.g. is daydreaming), then this will result in a slower reaction time in pressing the space bar, or reduced accuracy such as missing the stimulus entirely. This provides an abstract, context-free measure of whether the capacity for sustaining attention has improved. In addition, we also examine impacts on a supplementary task used in the psychology literature—a symbol matching task. Students are given pages containing a grid of randomly ordered pictorial symbols. A specific set of 2-3 target symbols is displayed at the top of the sheet above the grid. Students are asked to go through the grid, crossing out any of the target symbols they encounter. We define performance using the standard metric of performance on these tasks: student's correct true positive rate z-score minus the false positive rate z-score.

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 0.081 SDs (Table IV, Col. 1, $p = 0.028$). We decompose these effects in the remaining columns. Notably, performance on SART increases by 0.108 SDs (Col. 2, $p = 0.047$). The average effect on the symbol matching task is 0.065 SDs, but this difference is not significantly significant (Col. 3, $p = 0.148$).

In Col. 4, we examine impacts separately by each sub-treatment. The Math and Games arms improve average performance by 0.088 SD and 0.075 SD, respectively ($p = 0.040$ and $p = 0.085$). Moreover, we cannot reject that the effect of both sub-treatments is the same ($p = 0.756$). Overall,

[32] 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). In this section, we examine effects on measures of sustained attention to help provide positive evidence on mechanisms. However, in the paper as a whole, we use the broader phrase "cognitive endurance"—both because our goal is not to take a stance on whether our core results are driven by sustained attention versus other related mechanisms, and because these related mechanisms have similar implications for the field behaviors we examine.

[33] This task is the outcome we examine that is not a paper-and-pencil measure. In order to distinguish it to the extent possible from tablet based training, we conduct the task on computers with large screens and external keyboards where the space bar is marked in red tape to clearly identify the appropriate key to press when a stimulus appears.

[34] We modify the traditional SART task slightly to make it more child-friendly. For example, we adjust the frequency of the target stimuli and use shapes rather than numbers.

these findings suggest that the treatments improved sustained attention as measured by psychologists.

5.2 Attentiveness in the Classroom

Do improvements in cognitive endurance translate to changes in observed classroom behavior? 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, which we adapt to our setting.[35] We examine student behavior along three dimensions: (1) whether students attend to and carry out 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);[36] (2) their response to auditory stimuli (noticing and attending to new sound during class); and (3) their physical signs of inattention (fidgeting, looking out the window, or pestering a classmate during the teacher's lecture).

These observations are conducted with students while in their normal class section (i.e. so that students from different treatment arms are mixed in the room). 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 in this context. Rather, it is common to have a teaching assistant or head teacher sit at the back of the room and observe the class.

Students who receive cognitive practice are more attentive during their classes, with an average improvement of 0.094 SDs on the index measures overall (Table V, Col. 1, $p = 0.006$). In addition, the two sub-treatments, Math and Games, each improves classroom attentiveness by 0.117 and 0.070 SDs (Col. 5, $p = 0.003$ and $p = 0.075$, respectively); we cannot reject that these two coefficients are the same. These results suggest that practicing focused cognitive activity—whether academic or non-academic in nature—improves students' basic attentiveness in the classroom.

5.3 Motivation and Cognitive Endurance

What is the role for motivation in the endurance effects we see? For example, did the cognitive practice treatments prompt students to try harder in school, or train them to have an improved capacity to keep going when they are tired? To better understand the potential role for motivation, we examine two supplementary pieces of evidence.

[35] The measures draw on 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.

[36] Teachers asked students to move their school materials from one side of the classroom to another, and to write their roll number (i.e. their seat number, known to all students and used ubiquitously to identify them) on the upper right corner of a paper and turn it in five minutes later. 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. the number should be written in the upper left corner), and then had the presence of mind to carry it out five minutes later rather than forgetting.

First, the results in Table I indicate that students are not simply trying harder in general, since that would lead to improvements at the start of the test 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 a subset of the declines tests, we randomize incentives so that students earn toys and other prizes for higher test scores. 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 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 treatment status within a given test subject (e.g. Ravens).

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 (Col. 2). In Cols. (3), 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 0.168 SDs ($p = 0.001$). This indicates that performance is highly elastic to effort even at the start of the tests when students are fresh.

However, we see no evidence that being more driven reduces performance declines later in the test; if anything, in Col. (3), the coefficient on "Decile 6-10*Incentive" is actually negative in sign.[37] Rather, these patterns suggest that when students are motivated to try harder, they do better throughout the test (i.e. from the beginning), with similar levels of decline over the course of the test. We find similar results using our alternate estimation strategy using predicted decline in Col. (4).

The above tests, while informative, may not capture all dimensions of motivation. 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.

[37] If being more motivated led to less performance decline, this coefficient would be positive and significant. A negative coefficient may indicate that because students try harder earlier in the test, cognitive fatigue may set in faster for them (Iyengar and Kamenica, 2010; Kamenica, 2012).

5.4 Alternate Mechanisms

The cognitive practice treatments could arguably boost students' performance through other channels. For example, they may have taught some subject content, increased confidence or excitement about school, 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 declines results, it is unclear why they should not manifest as students also doing somewhat better early in the tests.

In addition, our design and choice of outcomes also makes some of these other explanations less likely. For example, neither the Games or Math arm could have taught students subject content that would have made them do better in the listening comprehension test; this test asks students to recall simple details from short audio stories, and so by design is not a test of content knowledge or skills. Similarly, if the treatments improved working memory, it is not clear why there would be only slope effects and no change in levels at the start of the tests. For example, the listening test questions ask about details in random order (and so uncorrelated with the order in which the details appear in the story). An improved ability to hold an object in working memory and manipulate it may make a student better at a Ravens test,but this should also affect performance on the first Raven's question. Note that our results also cannot be explained by improvements in test taking strategy, or increased familiarity with tablets.[38,39] In summary, 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, across the disparate tests such as listening and Ravens Matrices, and why there are no changes in performance at the start of the tests.

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 (Table A.11). 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 before school. While not definitive, this indicates that it is unlikely that changes in home behaviors drive a substantial portion

[38] A potential concern is treated students intuit better test-taking strategies, such as skipping hard questions. It is difficult to see how this skill could be developed via the Games sub-treatment. We also directly mitigate this concern by ensuring sufficient time and high completion rates for tests. Consistent with this, results are very similar if we restrict to attempted questions (Appendix Table A.8). Finally, recall that a subset of our tests (e.g. listening or SART) mechanically do not permit students to skip around or move faster through the tests (see above).

[39] All of our primary outcomes—the declines tests (math, listening, and Ravens) and school grades—are based on traditional paper-pencil assessments. One of our supplementary measures, SART, which must be electronic to accurately measure reaction times, is computer-based. For this, students simply press the space bar when a stimulus appears on a screen, and we administer the task on a laptop with an external keyboard to make it as distinct as possible from the tablet-based interventions.

of our results. Of course, they do not rule out the general relevance of students' home environments in developing cognitive endurance.

6 Results II: School Performance

In this section, we examine overall impacts on students' school grades. This enables us to test for impacts on regular field behavior and outcomes. Changes in school performance would indicate that simply engaging in thinking for sustained periods can improve students' capacity to build traditional human capital—an interesting and meaningful policy outcome irrespective of mechanism.

The findings above suggest several ways in which our treatments could affect students' ability to learn and their academic performance in school. For example, our listening results indicate that treated students may be able to sustain attention toward a teacher while she is lecturing for longer. The broad applicability of cognitive endurance would suggest potential similar improvements during other activities—such as maintaining focus while reading a textbook, or thinking through a challenging concept at the end of a long class session. In addition, conditional on their academic knowledge, students may receive higher scores on exams and assignments simply due to their ability to sustain performance for longer. Note that these are all implications of improved cognitive endurance, regardless of one's view of the precise underpinning for cognitive endurance effects.

In our setting, the core academic subjects are Hindi, English, and Math. These are taken by all students in our sample, and are the only subjects offered universally to all students across our schools. In addition to their intrinsic importance for basic literacy and numeracy, these subjects also enable us to test for broad impact of our intervention. Neither treatment arm taught students Hindi or English—making these subjects wholly unrelated to the content of the Cognitive Practice treatments.[40] In addition, we had no engagement with students' regular academic classes or assessments; these academic grades are determined as usual by the school without any consultation or inputs from the research team.[41]

We examine effects in Table III. Overall, receiving Cognitive Practice improves school performance by 0.0897 SDs (Panel A, Col. 1, $p = 0.010$). These sizable gains are present even in Hindi and English, with impacts of 0.0989 SDs ($p = 0.012$) and 0.0919 SDs ($p = 0.024$), respectively (Panel A, Cols. 3-4). Panel B disaggregates these results by sub-treatment arm. Each of the Math and Games arms generally has significant effects on each of the three core academic subjects. The 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

[40] 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.

[41] As we discuss in Section 3.3 above, we did not share students' treatment status with teachers, and teachers played no role in getting students to their program classes.

that on Games Practice, but we cannot reject that they are the same.[42]

The magnitude of these effects 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 had a similar impact on academic gains of 0.12 SD of kindergarten through grade 3 students (Krueger and Whitmore, 2001).[43] 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 any academic learning (e.g. in the Games arm).

How much of these impacts stems from increased learning versus being less fatigued in later parts of homework assignments or tests? A back of the envelope exercise which applies the reductions in the rate of performance decline found in the Listening, Raven's Matrices, and Math exams to expected rates of decline on the school administered exams finds that approximately 1/3 of the impact on grades is driven by improved test performance. This suggests that the remaining 2/3 of the effect is likely due to a mixture of learning and or other related effects on effects (e.g., in assignments and school exams, later questions cover progressively more difficult concepts, and students often do not reach the end).

However, the composite measure of overall school performance is an important and policy relevant educational outcome. It is a standard proxy for academic attainment, and is consequential for students' future success—for example, it used by educators and employers to rank students for promotion to the next grade or college admissions. In addition, both the learning and fatigue channels are relevant beyond school; as our examples in Section 8 below show, productivity at work depends both on how much you know, and your ability to sustain performance over the course of a task or work shift.

As a whole, the results in Table III indicate that simply spending time in effortful thinking—*without learning any subject content*—improves traditional measures of human capital. Moreover, such thinking need not even be academic in nature: even the students who receive Games Practice do substantially better in their academic classes. This directly supports the possibility that the process of receiving an education, through the experience of cognitive practice, can improve human capacity.

[42] We do not find any significant heterogeneity in treatment effect by student grade, gender, baseline average score (Appendix Table A.10, Panel B). It does appear there is a marginally smaller treatment effect for students who had worse cognitive endurance at baseline ($p = 0.056$).

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

7 Supplementary Evidence: Impact of Schooling on Cognitive Endurance

Our intervention expands cognitive endurance by increasing time in effortful thinking—an activity that is inherent in the experience of formal education. However, in our study, this is accomplished through an outside intervention introduced into schools. In this section, we augment our evidence by examining whether the natural experience of schooling helps develop cognitive endurance. Using supplementary data, we exploit variation in years of schooling based on birthday cutoffs for kindergarten enrollment. If schooling improves cognitive endurance, we might expect reduced performance declines among students who are just above the enrollment cut-off (i.e. who have one additional year of schooling conditional on age).

We test this idea using data from Brown and Andrabi (2021), which includes a sample of over 5,300 nine to eleven year-olds across 66 schools in Pakistan. This larger dataset offers both the necessary power for a regression discontinuity analysis and also allows us to examine heterogeneity by school quality. In addition, the data has the unique feature that exams—covering math, science, English, and Urdu—have randomized question order across students, enabling us to identify effects on performance declines. Appendix Figure A.8 documents that we see substantial performance declines in each of the four test subjects. On average, the probability of answering a given question correctly declines by 16 percentage points (p.p.) from the first to the last decile of the test.

In Pakistan, the birthday cutoff for kindergarten admission is December 31; 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. This exercise has its limitations—for example, enrolling in school earlier versus later could also change other inputs—and we therefore view it as only suggestive.[44] However, despite these limitations, we view it as offering a helpful signal that complements our field experiment results. Because we do not have perfect compliance with he birth month cutoff, we use a fuzzy regression discontinuity approach, where we instrument for years of schooling with the birth month cutoff. Our key specification is:

First stage:

$$YrsofSchooling_s = \alpha_0 + \alpha_1 MOB_s + \alpha_2 MOB_s^2 + \alpha_3 1[MOB_s \leq 6] + \mu_s \tag{4}$$

Second stage:

[44] The "treatment" is entering kindergarten at age 5 versus staying home for an additional year and entering at age 6. Note that this approach necessarily compares a student who is the youngest in their class (i.e. the "treated" group with an additional year of school) to those who are the oldest students in their class (i.e. the "control" group with one fewer year of schooling). Lower relative age (and therefore emotional maturity) could, for example, impede learning, with the direction of effects on cognitive endurance unclear. Being at home longer before enrolling in school could also change non-school inputs.

$$Correct_{ils} = \beta_0 + \beta_1 \widehat{Yrsof Schooling}_s + \beta_2 PredictedDecline_l + \beta_3 \widehat{Yrsof Schooling}_s PredictedDecline_l + +\beta_4 MOB_s + \beta_5 MOB_s^2 + \beta_6 MOB_s * PredictedDecline_l + \beta_7 MOB_s^2 * PredictedDecline_l + \epsilon_{ils} \quad (5)$$

where $YrsofSchooling_s$ captures the total years of schooling student s has received at the time of the exam. MOB_s is the student's month of birth, which is the running variable in the regression discontinuity framework. $1[MOB_s \leq 6]$ is an indicator that equals one if the student was born in the second half of the year (i.e. July to December, before the cut-off) and zero otherwise.[45] $Correct_{ils}$ is a binary variable that captures whether student s correctly answered question item i appearing in location (decile) l. $PredictedDecline_l$ is calculated in a parallel fashion to Eq. 2, where for each test subject we take the difference between the average score in decile 1 minus the average score in decile l.[46] The coefficient of interest is β_3, which captures the extent to which an additional year of schooling mitigates the predicted performance decline.

Controlling for month of birth, we find students who are just older than the cutoff have 0.22 more years of schooling ($p < 0.000$) at age 9-11. The imperfect compliance is due to some parents choosing to hold their child back to start when they are older or schools choosing to implement their own birthday cutoff. The coefficient and statistical significance remain the same when we control for month of birth linearly or using a quadratic. We also do not find evidence of manipulation in the birth month variable around the cutoff (McCrary test $p = 0.504$).

Table VII presents the effect of an additional year of schooling on cognitive endurance. In Col. (1), *YrsofSchooling* x *PredictedDecline* provides the estimate of β_3. We find that, conditional on age, an additional year of schooling mitigates the rate of performance decline by 31% ($p = 0.030$). This estimate is similar regardless of whether we control for a linear or quadratic function of the running variable (Cols. 1-2).[47] Since the average student sees a 15 p.p. performance decline over the course of a given test, these results indicate that those with an additional year of school would see only a 10 p.p. decline on the same exam. The magnitudes in Cols. (1)-(2) imply that the impact of a full year of schooling is 3.4 times larger than the effect of our more limited experimental intervention (Table I, Col. 2).

Do these benefits vary by school quality? The data include measures based on video recordings of students' classes; observers code the videos using the CLASS rubric, a common tool for assessing pedagogical quality (Araujo et al., 2016; Pianta et al., 2012).[48] Table VII, Columns (3)-(8) show

[45] The sample is restricted to students born from July 2007 to June 2009. MOB begins at 1 for July 2007 goes up to 12 for June 2008 and then resets to 1 for July 2008 and goes up to 12 again for June 2009. The definition of $1[MOB_s \leq 6]$ identifies a discrete jump at January in expected years of schooling. We use the 6 months before/after each January 1 cut-off to create non-overlapping samples of treatment and control students for the stacked RD across grades.

[46] We use students who are born in May through August as our "control group" to calculate Predicted Decline.

[47] Our running variable only takes 12 values; despite this, results are similar regardless of whether we use a linear or quadratic functional form.

[48] The rubric contains rating of the class operations along 12 dimensions, from classroom climate, feedback provided

the heterogeneity in the effect of an additional year by three measures of quality. In each column, *Higher quality* captures the quartile rank in the given quality dimension from 0 to 3, with 0 denoting the bottom quartile and 3 denoting the top quartile. Note that the dataset does not contain direct information on socioeconomic status; however, in other contexts, higher SES schools are associated with higher quality scores on the CLASS rubric.

In Cols. (3)-(4), we examine heterogeneity by overall school quality: the school's average score on all 12 components of the CLASS rubric (including classroom climate, use of higher order thinking skills, time in independent practice, feedback, etc). Among the schools in the bottom quartile of the quality distribution, we cannot reject that additional schooling has no effect on cognitive endurance. However, as we move up the quality distribution for each quartile rank increase, an additional year of school reduces performance declines by 22% ($p = 0.034$). Results are similar when examining the quality of pedagogy in the student's specific grade (Cols. 5-6). Finally, in Cols. (7)-(8), we test for heterogeneity by time spent in independent focused practice during class. Again, we see a similar pattern of results: additional schooling only mitigates performance declines when students spend time in independent practice. Of course, this measure is correlated with, and therefore may reflect, other dimensions of quality.

Overall, the results in Table VII provide suggestive evidence for the role of schooling in developing cognitive endurance. However, 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. This strengthens the interpretation of the motivational patterns in Section 2. Coupled with the results from our experiment, the complementary results in this section suggest that differential access to good quality schooling could widen economic disparities through the development of cognitive endurance.

8 Conclusion: Discussion and Broader Relevance

We conclude by discussing the broader implications of our findings. We first begin by presenting evidence for socioeconomic differences in cognitive endurance in behaviors outside of schooling. Using supplementary data, we present two examples from substantially different high-stakes activities: productivity among data entry workers, and voting at the ballot box.[49]

to students, time on task, use of higher order thinking skills, etc. Brown and Andrabi (2021) describes the scoring and quality assurance process used in reviewing the classroom videos. Note that while all 66 schools are part of the same private school chain, there is substantial heterogeneity across them, with different schools serving different demographics and charging different levels of school fees.

[49]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 low versus high SES individuals are not severely confounded by differential selection into the task. As an example, looking at cognitive endurance among doctors would violate condition (ii), since the types of low SES individuals who select into being a doctor are surely positively selected. Their performance relative to high SES doctors would therefore be difficult to interpret as reflecting general performance decline differences between the two groups.

In Figure IVa, 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.[50] 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 a non-default choice). These effects are substantial: an additional 6% of propositions would have become laws if they had appeared earlier in the ballot. Using data obtained from the authors, we use racial composition of a voting precinct as a proxy for socioeconomic status (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 IVb). However, over time, the propensity to make an active (i.e. non-default) choice declines much more quickly for lower socioeconomic status individuals. Specifically, high socioeconomic status 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.[51] In addition, while the examples in Figure IV 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? In addition, the patterns in Figure IV raise the question of how we should understand SES differences in cognitive endurance among adults. For example, this may reflect persistence from childhood training, or reflect the fact that those working in higher skilled jobs may receive more cognitive practice through work—reinforcing and perpetuating 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. Our intervention exemplifies a policy lever that may be useful in improving endurance

[50] 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 SES 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)).

[51] While the previous literature has documented performance declines in various settings, the examples we present above are the first documentation of SES heterogeneity in cognitive endurance among adults of which we are aware.

among lower-income children: incorporating opportunities for them to engage in effortful thinking for sustained periods of time at school or home. The supplementary data we present in Sections 2.2 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 levels was an effective solution. In other settings, other approaches may be more appropriate for getting students to undertake mentally challenging activity. As an example, in low income US settings, some "testing schools", which place strong emphasis on giving children frequent assessment tests, have been successful at raising student outcomes. Because this approach creates frequent periods where students must sit and concentrate for long stretches (i.e. during weekly tests), it may have the ancillary benefit of also 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

References

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

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.

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.

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.

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.

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, 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.

Deming, David J, "Better schools, less crime?," *The Quarterly Journal of Economics*, 2011, *126* (4), 2063–2115.

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

_ , "The Growing Importance of Decision-Making on the Job," Technical Report, National Bureau of Economic Research 2021.

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

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.

Edkins, Graham D and Clare M Pollock, "The influence of sustained attention on railway accidents," *Accident Analysis & Prevention*, 1997, *29* (4), 533–539.

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.

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

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 J., Lance J. Lochner, and Petra E. Todd, "Chapter 7 Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond," in E. Hanushek and F. Welch, eds., *Handbook of the Economics of Education*, Vol. 1 of Handbook of the Economics of Education, Elsevier, 2006, pp. 307–458.

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.

Iyengar, Sheena S and Emir Kamenica, "Choice proliferation, simplicity seeking, and asset allocation," *Journal of Public Economics*, 2010, *94* (7-8), 530–539.

Kamenica, Emir, "Behavioral economics and psychology of incentives," *Annu. Rev. Econ.*, 2012, *4* (1), 427–452.

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.

Lance, Lochner, "Chapter 2 - Nonproduction Benefits of Education: Crime, Health, and Good Citizenship," in Eric A. Hanushek, Stephen Machin, and Ludger Woessmann, eds., *Handbook of The Economics of Education*, Vol. 4 of Handbook of the Economics of Education, Elsevier, 2011, pp. 183–282.

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.

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.

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

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.

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.

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," 2004.

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 Performance Declines

	Dependent Variable: 1[Question Correct]					
	Test Subject					
	All	All	Non-Math	Math	Listening	Ravens
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pooled Treatments						
Cog. Practice x Deciles 6-10	0.0129***					
	(0.0047)					
Cog. Practice x Deciles 2-5	0.0084*					
	(0.0049)					
Deciles 6-10	-0.0597***					
	(0.0051)					
Deciles 2-5	-0.0115***					
	(0.0037)					
Cog. Practice x Predicted Decline		0.0927***	0.0818***	0.104**	0.0677**	0.0972**
		(0.0285)	(0.0285)	(0.0428)	(0.0329)	(0.0451)
Cognitive Practice	-0.0027	-0.0050	-0.0018	-0.0089	-0.0013	-0.0049
	(0.0060)	(0.0061)	(0.0042)	(0.0091)	(0.0068)	(0.0100)
Panel B: Disaggregated by Sub-treatment						
Math Practice x Deciles 6-10	0.0127**					
	(0.0055)					
Games Practice x Deciles 6-10	0.0131**					
	(0.0054)					
Math Practice x Deciles 2-5	0.0034					
	(0.0057)					
Games Practice x Deciles 2-5	0.0135**					
	(0.0057)					
Math Practice x Predicted Decline		0.0976***	0.0993***	0.0955*	0.0895**	0.112**
		(0.0329)	(0.0341)	(0.0488)	(0.0393)	(0.0540)
Games Practice x Predicted Decline		0.0881***	0.0639*	0.115**	0.0457	0.0815
		(0.0329)	(0.0328)	(0.0501)	(0.0384)	(0.0516)
Math Practice	-0.0050	-0.0053	-0.0042	-0.0011	-0.0050	-0.0098
	(0.0070)	(0.0070)	(0.0049)	(0.0103)	(0.0078)	(0.012)
Games Practice	-0.0050	-0.0046	0.0008	-0.0165	0.0025	0.0003
	(0.0071)	(0.0070)	(0.0048)	(0.0107)	(0.0079)	(0.0115)
p-value: Math Decline = Games Decline		0.7273	0.3161	0.7707	0.2895	0.5824
Control Decline	0.12	0.12	0.05	0.18	0.06	0.03
Observations	329349	329349	129115	200234	66932	62183

Notes: This table examines the impact of cognitive practice on the rate of performance decline over time. Panel A estimates treatment effects for both treatments pooled relative to the control group. Panel B shows effects for the Math and Games sub-treatments (each relative to the control group) separately. Col. (1) corresponds to the specification in Equation 1. Col (2) corresponds to the specification in Equation 3. "Cognitive Practice" is a binary indicator that equals 1 if the student was assigned to a treatment (either the Math or Games Practice). "Predicted Decline" is the amount of average decline in each quintile of the test location, relative to the first quintile of the test, within each given school. "Deciles 2-5" and "Deciles 6-10" are binary indicators that equal one if the question appears in the second through fifth deciles of the test or the second half of the test, respectively. Cols. (1) and (2) estimate treatment effects for all three tests pooled. Cols. (3)-(6) show effects for the non-Math tests (Listening and Ravens), then Math, Listening, and Ravens tests separately, respectively. The Coefficients in Cols. (3)-(5) are estimated from a single regression on all the data. Question item order was randomized across students. All regressions contain question and test version fixed effects, and baseline controls. Observations are at the student-test-question level. Standard errors are clustered by student, and bootstrapped in columns (2)-(6). The dependent variable mean is 0.47 in the control group. The coefficient on "Predicted Decline" is -0.15 in Col. 2 ($p < 0.001$). * p<0.10, ** p<0.05, *** p<0.01.

Table II: Persistence of Treatment Effects

	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	0.0143***			
	(0.0051)			
Cog. Practice x Deciles 6-10 x Follow-up	-0.0044			
	(0.0113)			
Cog. Practice x Predicted Decline		0.0933***	0.1104***	0.0764**
		(0.0304)	(0.0341)	(0.0338)
Cog. Practice x Predicted Decline x Follow-up		-0.0012	-0.0182	0.0155
		(0.0434)	(0.0512)	(0.0504)
F-test p-value: sum of 2 coefficients = 0	0.3249	0.0325	0.0683	0.0628
Observations	329349	329349	219341	217223

Notes: This table examines the persistence of the treatment effects over time. "Follow-up" is a binary indicator that equals one if the test is a follow-up test, administered roughly 3 to 5 months after the end of the intervention. "Deciles 6-10" is a binary indicator that equals one if the question appears in the second half of the test. "Predicted Decline" is the amount of average decline in each quintile of the test location, relative to the first quintile of the test, within each given school. Question item order was randomized across students. All regressions contain baseline controls, question fixed effects, and test version fixed effects. Observations are at the student-test-question level. Standard errors are corrected to allow for clustering by student, and bootstrapped in columns (2)-(4). * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table III: Impacts on School Performance

	Dependent Variable: Z-score of Student's Grades				
Subject:	All	Non-Math	Hindi	English	Math
	(1)	(2)	(3)	(4)	(5)
Cognitive Practice	0.0897**	0.0923**	0.0989**	0.0919**	0.0849**
	(0.0348)	(0.0386)	(0.0393)	(0.0407)	(0.0377)
Sub-treatments:					
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 (pooling mid-year and end-of-year grades) in the three core subjects offered by all schools in the study (Hindi, English and math). The dependent variable is the student's z-score on the test. "Cognitive Practice" denotes receiving any treatment, "Math Practice" and "Games Practice" denote the Math or Games practice sub-treatments, respectively. Cols. (1)-(5) regress a z-score of student performance on a dummy for Cognitive Practice, in the first row, and on dummies for Math and Games Practice, in the second and third rows, respectively. All regressions include class section (strata) fixed effects and baseline controls. Observations are at the student-subject-exam level. Column (1) includes all three subjects. Columns (2) restricts to English and Hindi, and Col. (3)-(5) present each subject separately. Standard errors 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	0.0814**	0.1080**	0.0650	
	(0.0371)	(0.0543)	(0.0449)	
Sub-treatments:				
Math Practice				0.0883**
				(0.0429)
Games Practice				0.0747*
				(0.0434)
p-value: Math Pratice = Games Practice				0.7560
Observations	9704	3897	5807	9704

Notes: This table examines the impact of cognitive practice on traditional measures of attention in the psychology literature. The Sustained Attention to Response Task (SART) task measures focus via reaction times and accuracy to stimuli displayed on a computer. Symbols Matching is a task in which students cross out instances of symbols listed at the top of the page in a grid below. Outcomes are measured as *true positive z-score - false positive z-score*, winsorized at the 99th percentile. Clustered standard errors are in parentheses. Regressions control for baseline test scores classroom fixed effects. Observations are student-tests. * 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	Pooled
	(1)	(2)	(3)	(4)	(5)
Cognitive Practice	0.0940***	0.0971*	0.1363**	0.0452	
	(0.0340)	(0.0572)	(0.0623)	(0.0582)	
Sub-treatments:					
Math Practice					0.1174***
					(0.0393)
Games Practice					0.0703*
					(0.0394)
p-value: Math Practice = Games Practice					0.2335
Observations	1206	1198	1197	1196	1206

Notes: This table examines the impact of cognitive practice on classroom based measures of attention adapted from the Vanderbilt ADHD diagnostic teacher rating scale. Classroom observers were blind to students' treatment status. The "Pooled" measure is a simple average of the z-scores for the individual outcomes within the scale. The "Physical Signs" measure was reversed so that a larger number corresponds to a better outcome, as is true of the other two outcome measures. All regressions include grade fixed effects. Clustered standard errors are in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table VI: Effect of Incentives on Test Performance

	Dependent Variable: 1[Question Correct]			
	(1)	(2)	(3)	(4)
Incentive	0.0916**	0.111**	0.168***	0.146***
	(0.0414)	(0.0417)	(0.0505)	(0.0490)
Cog. Practice x Incentive		-0.0293	-0.0466	-0.0377
		(0.0334)	(0.0474)	(0.0446)
Decile 6-10 x Incentive			-0.0583	
			(0.0391)	
Cog. Practice x Decile 6-10 x Incentive			0.0152	
			(0.0421)	
Predicted Decline x Incentive				-0.354*
				(0.201)
Cog. Practice x Predicted Decline x Incentive				0.0765
				(0.186)
Observations	11515	11515	11515	11515

Notes: This table reports the effect of offering students an incentive for their performance on the test (randomized during a subset of the exams administered by the study). "Incentive" is a binary indicator that equals 1 if the student was provided a toy if they reached a certain score on the exam. "Cog. Practice" is a binary indicator that equals 1 if the student was assigned to treatment (either the Math or Games sub-treatment). "Predicted Decline" is the amount of average decline in each quintile of the test location, relative to the first quintile of the test, within each given school. "Deciles 6-10" is a each binary indicators that equals one if the question appears in the second half of the test. The omitted category are the questions in decile 1 (i.e. the beginning) of the test. Question item order was randomized across students. All regressions contain question and test version fixed effects, and baseline controls. Observations are at the student-test-question level. Standard errors are clustered at the test-school-class level, the unit of randomization for the incentive treatment. The dependent variable mean is 0.47 in the control group. * p<0.10, ** p<0.05, *** p<0.01.

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)
Observations	276043	276043	276043	276043	276043	276043	276043	276043
Running variable func. form	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic	Linear	Quadratic

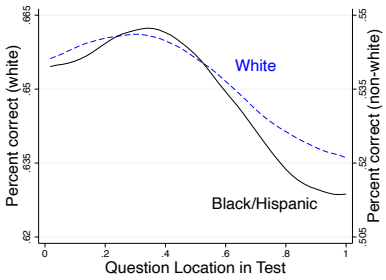
Notes: This table reports the effect of an additional year of school on student performance declines on exams and compares the effect of an additional year for higher and lower quality schools/classes. "Years of Schooling" is instrumented using whether the student's birthday is above or below the kindergarten entrance cutoff, controlling for birth month. "Predicted Decline" captures the average decline in performance for a given question item relative to the first decile of the test. Column (1) and (2) present the results of an additional year of school across all schools in our sample. Columns (3)-(8) show the heterogeneity in the effect of an additional year by school/class quality. "Higher Quality" captures the school or class's quartile rank in the given quality dimension from 0 to 3, with a value of 0 for schools/classes in the bottom quartile up to 3 for schools/classes in the top quartile. We use three different measures of school/class quality all of which are based on scoring 20 minute classroom videos using the CLASS observation rubric (Araujo et al., 2016; Pianta et al., 2012). "School Quality" captures the school's average score on all 12 components of the CLASS rubric (ranging from classroom climate, time on task, use of higher order thinking skills, etc). "Class Pedagogy" captures the average score on all 12 components in the student's current grade level. "Independent Practice Time" restricts to one of the 12 components which focuses on the quantity and quality of time students spend working independently on cognitively challenging material. Columns (1), (3), (5) and (7) include birth month as a linear control, and columns (2), (4), (6) and (8) include the quadratic of birth month as well. Data comes from a sample of 5,353 9-11 year olds in Pakistan (Brown and Andrabi, 2021). The F-statistic on the first stage is 15.9. Question order was randomized on the exams. Observations are at the student-test-question level. Standard errors are clustered by student. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

10 Figures

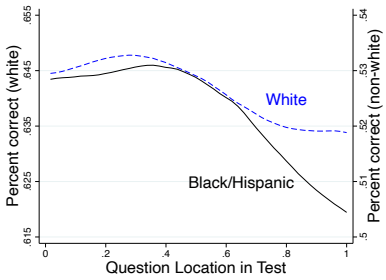
FIGURE I: Performance Declines in Achievement Tests

TIMSS Exam

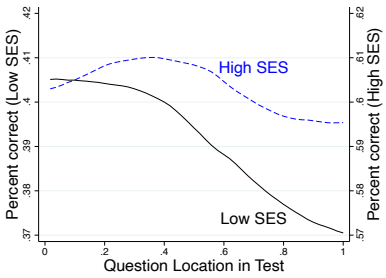
(A) Mathematics, US sample



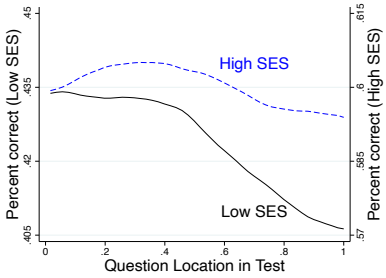
(B) Science, US sample



(C) Mathematics, global sample

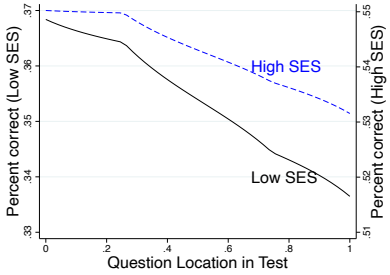


(D) Science, global sample

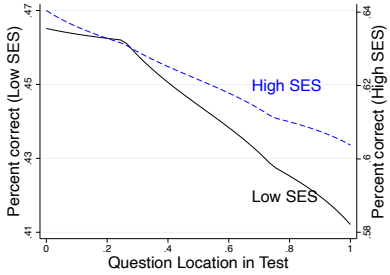


PISA Exam

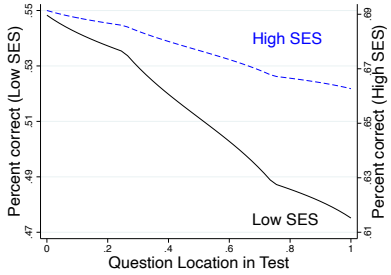
(E) Mathematics, US sample



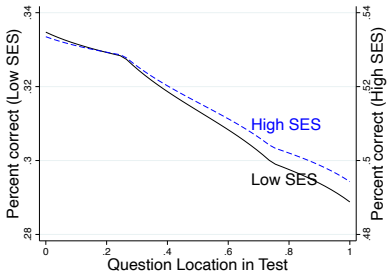
(F) Science, US sample



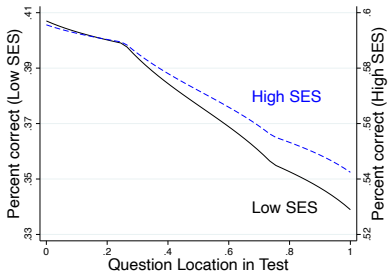
(G) Reading, US sample



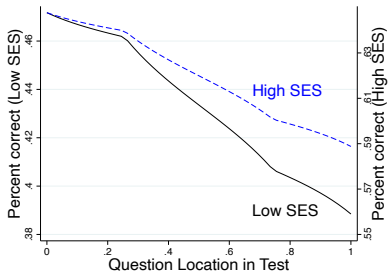
(H) Mathematics, global sample



(I) Science, global sample



(J) Reading, global sample

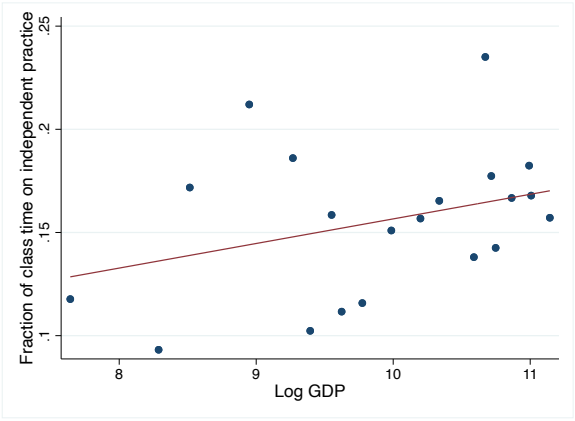


Notes: TIMSS data and PISA data, authors' calculations. Question order is block randomized. In the TIMSS US sample (A-B), high and low SES students are proxied by race (white and non-white, respectively). In the TIMSS global sample, (C-D) high (low) SES countries are proxied by the top (bottom) quartile of GDP/capita. In the PISA data (E-J), high (low) SES is proxied by the top (bottom) quartile of the ESCS measure, an index of SES included in the PISA data. Question location in test denotes where in the exam the question item appeared normalized on a scale of 0 to 1.

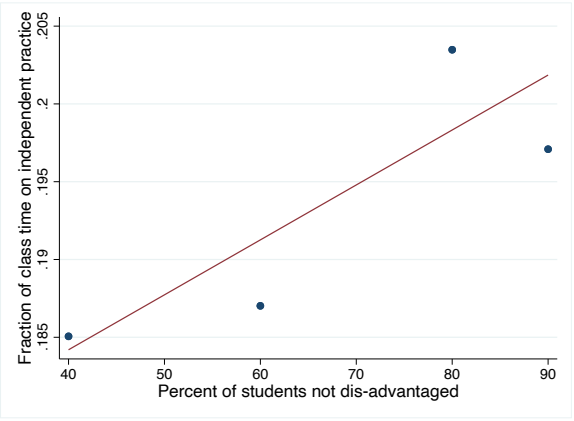
FIGURE II: Differences in Schooling Practices by Socioeconomic Status

Frequency of independent practice

(A) Global sample

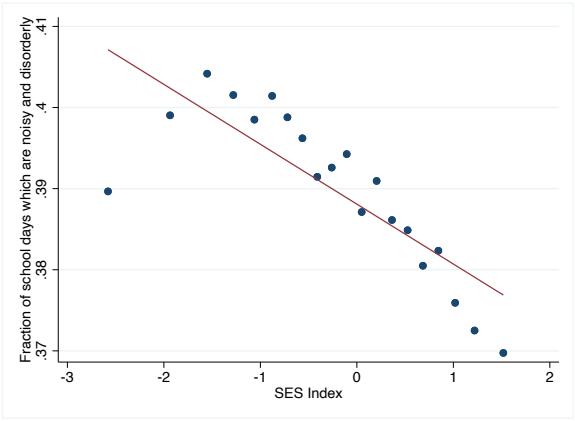


(B) US sample

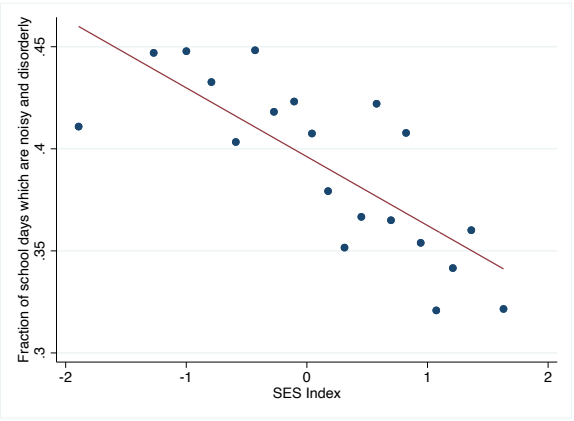


Frequency of classroom distractions

(C) Global sample

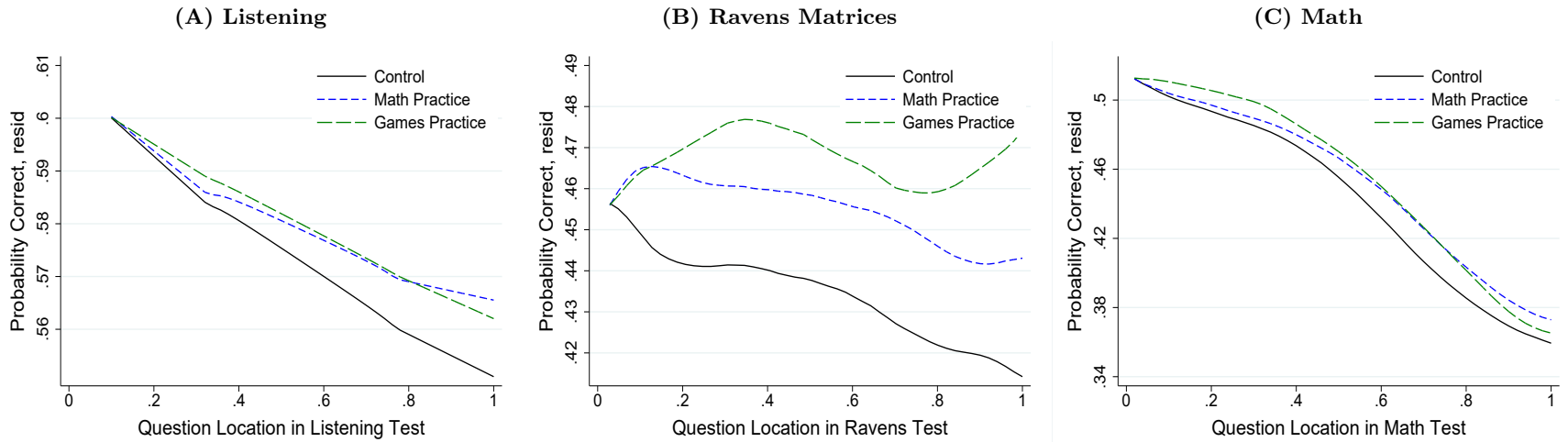


(D) US sample



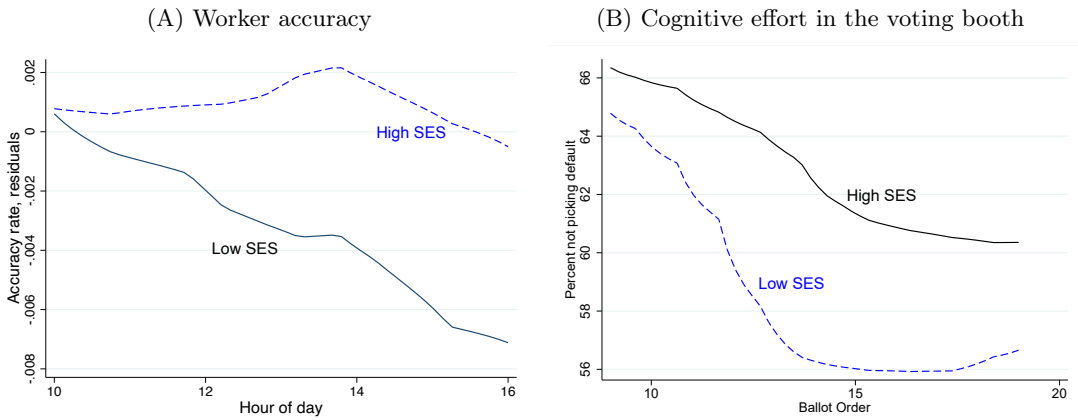
Notes: The figures show the relationship between income and schooling environment. Panels A and B present data from the TIMSS teacher survey on pedagogy used within the classroom. The y-axis is the fraction of class time spent on independent practice. Teachers rate how often students engage in this type of activity on a 4-pt scale from "never" (coded as 0) to "every or almost every lesson" (coded as 0.75). In Panel A, the sample is all countries, and the x-axis is Log of GDP. In Panel B, the sample is the US, and the x-axis is the percent of students within the school who are not disadvantaged (where the fraction of disadvantaged students is reported by school administrators from among 4 discrete options). Panels C and D present data from the PISA teacher survey. The x-axis is a student-level SES index constructed by PISA. The y-axis is the fraction of classes in which there is noise and disorder, rated by teachers on a 4-pt scale from "never" (coded as 0) to "every lesson" (coded as 1). The data is grouped into ventiles, presenting the average within each ventile (blue dots). In each plot, the red line is the line of best-fit.

FIGURE III: Experimental Treatment Effects: Performance Declines



Notes: This figure plots declines in performance over the course of three tests in the RCT: (a) listening, (b) Raven's Matrices, and (c) math. Question order is randomized in each exam. In each plot, the y-axis is the probability a question was answered correctly, and the x-axis is the percent location of the question on the test (where 0 is the beginning of the test and 1 is the end of the test). Data is residualized to remove question fixed effects. In each plot, the initial level at the start of the test is normalized to the control group mean in quintile 1 for that test for ease of interpretation of decline magnitudes. Each line displays performance over time for the control group (solid black line), Math Practice (short blue dashes), and Games Practice (long green dashes). Observations are at the student-test-question level; N = 66,932 (listening), 62,193 (Raven's Matrices), and 200,234 (math). Table I presents the full set of corresponding treatment effects estimates.

FIGURE IV: Cognitive Endurance among Adults: Differences by Socioeconomic Status



Notes: 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). Data are residualized after removing worker fixed effects. High SES is defined as 1 if the worker has above high school education (corresponding to the median split of the sample). The sample is 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. Item order in the voting data is quasi-random. The x-axis is the location of an initiative on the ballot and the y-axis is whether the voter selects a choice other than the default option. High (low) SES denotes polling precincts where the fraction of non-Hispanic white residents is above (below) the median. With the exception of the census data, all data was provided by the authors of each respective study.

A Supplementary Tables and Figures

Table A.1: Evidence on Pedagogy and Declines in Performance from the TIMSS Exam

	Time Spent on Indep. Practice	Item Correct	
	(1)	(2)	(3)
Log GDP	0.0105***		0.0704***
	(0.00230)		(0.00199)
Question Location		-0.0393***	-0.133***
		(0.00376)	(0.0121)
Time Spent on Indep. Practice		0.0885***	0.0695***
		(0.0204)	(0.0202)
Question Location x Time Spent on Indep. Practice		0.0317***	0.0278***
		(0.00682)	(0.00700)
Question Location x GDP			0.00954***
			(0.00113)
Constant	0.422***	0.513***	-0.175***
	(0.0224)	(0.0107)	(0.0218)
R^2	0.005	0.136	0.152
Observations	7476337	8217081	7476337

Notes: This table uses data from the TIMSS exam administered to fourth graders around the world, and GDP data collected from the World Bank, to show: 1) differences in pedagogy by income (col 1), and 2) differences in rate of decline of performance by SES and pedagogy (cols 2 and 3). "Time Spent on Independent Practice" is the fraction of study time students spend working independently. * p<0.10, ** p<0.05, *** p<0.01.

Table A.2: Baseline Balance

Variable	(1) Control N	(1) Control Mean/SE	(2) Cog. Practice (Pooled) N	(2) Cog. Practice (Pooled) Mean/SE	(3) Games Practice N	(3) Games Practice Mean/SE	(4) Math Practice N	(4) Math Practice Mean/SE	T-test P-value (1)-(2)	T-test P-value (1)-(3)	T-test P-value (1)-(4)	T-test P-value (3)-(4)
Panel A: Student Characteristics												
Grade	548	2.746	1115	2.706	555	2.739	560	2.673	0.594	0.931	0.403	0.452
		(0.062)		(0.044)		(0.062)		(0.061)				
School Income Tercile	548	2.254	1115	2.248	555	2.261	560	2.236	0.903	0.878	0.717	0.604
		(0.035)		(0.025)		(0.035)		(0.035)				
Income Tercile	548	1.960	1115	1.960	555	1.971	560	1.948	0.994	0.735	0.723	0.485
		(0.024)		(0.016)		(0.024)		(0.023)				
Baseline Ability Tercile	548	1.989	1115	1.992	555	1.991	560	1.993	0.933	0.961	0.922	0.962
		(0.028)		(0.019)		(0.028)		(0.027)				
Female	541	0.351	1097	0.364	548	0.349	549	0.379	0.620	0.927	0.343	0.297
		(0.021)		(0.015)		(0.020)		(0.021)				
Panel B: Student Baseline Scores												
Baseline Listening (mean)	499	0.559	1027	0.549	507	0.542	520	0.555	0.615	0.485	0.857	0.600
		(0.017)		(0.012)		(0.017)		(0.017)				
Baseline Math (mean)	493	0.398	995	0.412	493	0.411	502	0.413	0.218	0.325	0.253	0.892
		(0.010)		(0.007)		(0.010)		(0.009)				
Baseline Ravens Matricies (mean)	491	0.367	1004	0.370	493	0.360	511	0.379	0.841	0.671	0.452	0.242
		(0.012)		(0.008)		(0.012)		(0.012)				
Baseline Listening (decline)	487	-0.002	1001	-0.020	495	-0.026	506	-0.014	0.461	0.406	0.648	0.698
		(0.020)		(0.014)		(0.021)		(0.020)				
Baseline Math (decline)	493	-0.070	995	-0.064	493	-0.070	502	-0.058	0.771	0.998	0.614	0.623
		(0.017)		(0.012)		(0.017)		(0.017)				
Baseline Ravens Matricies (decline)	460	-0.071	951	-0.027	462	-0.024	489	-0.030	0.246	0.316	0.256	0.894
		(0.025)		(0.023)		(0.040)		(0.026)				

Notes: This table presents summary statistics for student baseline covariates by treatment group. Columns (1), (2), (3) and (4) present the sample size and mean for each covariate by treatment status. Column (2) pools students who are in either treatment group and Column (3) and (4) separates students by their sub-treatment. Columns (5)-(8) present p-values for the test of equality of means. Panel A includes student characteristics. Student, baseline ability and income were provided from school. Students for who income or ability tercile was not available for the school were coded as being in the middle tercile. Students' gender was determined based on their name. For 1.6% of the sample students' name was gender neutral, so we leave the variable missing. School Income Tercile is based on the school the student attended which two of the six program schools designated in each tercile. Panel B presents students' performance each of the decline tasks, showing the average score on the task (mean) and the performance at the end of the task minus the beginning (decline). * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table A.3: Attrition

	Treatment			Disaggregated Sub-Treatments				
	Control	Cog. Practice	p-value 1 = 2	Math Practice	Games Practice	p-value 1 = 4	p-value 1 = 5	p-value 4 = 5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: School Administered Exams								
Pooled	0.8957	0.8955	0.99	0.8980	0.8930	0.88	0.87	0.75
	(0.011)	(0.008)		(0.011)	(0.011)			
Math	0.8957	0.8955	0.99	0.8980	0.8930	0.88	0.87	0.75
	(0.011)	(0.008)		(0.011)	(0.011)			
Hindi	0.8957	0.8955	0.99	0.8980	0.8930	0.88	0.87	0.75
	(0.011)	(0.008)		(0.011)	(0.011)			
English	0.8957	0.8955	0.99	0.8980	0.8930	0.88	0.87	0.75
	(0.011)	(0.008)		(0.011)	(0.011)			
Panel B: Psychology and Classroom Measures								
Pooled	1.0000	1.0000	1.00	1.0000	1.0000	1.00	1.00	1.00
	(0.000)	(0.000)		(0.000)	(0.000)			
Symbol Matching	0.9969	0.9962	0.80	0.9939	0.9985	0.43	0.56	0.18
	(0.002)	(0.002)		(0.003)	(0.002)			
SART	0.9089	0.9377	0.05	0.9289	0.9469	0.26	0.02	0.25
	(0.013)	(0.008)		(0.012)	(0.010)			
Panel C: Experimental Exams: Listening, Ravens Matrices, and Math								
Pooled	0.9708	0.9777	0.32	0.9778	0.9776	0.39	0.41	0.98
	(0.006)	(0.004)		(0.005)	(0.005)			
Math	0.9615	0.9744	0.09	0.9752	0.9736	0.13	0.18	0.85
	(0.007)	(0.004)		(0.006)	(0.006)			
Listening	0.9668	0.9692	0.76	0.9687	0.9697	0.84	0.75	0.91
	(0.007)	(0.004)		(0.006)	(0.006)			
Ravens	0.9602	0.9698	0.23	0.9687	0.9710	0.37	0.25	0.79
	(0.007)	(0.004)		(0.006)	(0.006)			

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

Table A.4: Test Characteristics

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

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

Table A.5: Average Question Difficulty by Question Location and Treatment

	Test Quintile				
	1	2	3	4	5
Math Practice	-0.00571	-0.00493	0.00128	-0.000778	-0.00355
	(0.231)	(0.256)	(0.659)	(0.854)	(0.397)
Games Practice	0.00166	0.00303	0.00176	-0.00650	0.00155
	(0.734)	(0.473)	(0.545)	(0.129)	(0.726)
Mean of dependent variable	0.490	0.443	0.456	0.443	0.468
SD of dependent variable	0.276	0.249	0.261	0.246	0.276
Number of observations	77148	52698	71712	52698	80266

Notes: This table presents the average question difficulty in each quintile of the test by treatment status to confirm that difficulty was not differential over time by treatment status. The outcome variable is the control group mean performance on the question item. "Math Practice" is a dummy for whether the student was part of the math sub-treatment and "Games Practice" is a dummy for whether the student was part of the games sub-treatment. Data is at the student-question item level and is from the three tests used to measure student declines: Math, Listening and Ravens. Each column restricts to question items in the given quintile of the test (e.g., column (1) restricts to the first 20% of the test). Standard errors are clustered at the student level. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table A.6: Test Completion

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

Notes: This table presents information about how much of each test students completed. "% attempted" is the percentage of individual question items students provided an answer to. "% skipped" is the percent of questions in which students left a question blank but answered at least one subsequent question. "% of students completing last question item" captures the percent of students who provided an answer on the last question of the exam, proxying for "finishing" the exam. "Avg. last question completed location" captures the average location of the last question item a student completed on the test as a percent of the total test. The listening and ravens tests are multiple choice tests and the math exam is free response.

Table A.7: Declines in Performance - Robustness to Predicted Decline

	Dep. Var.: 1[Question Correct] Average Predicted Decline:		
	Overall	by School	by School-Test
	(1)	(2)	(3)
Panel A: Pooled Treatment Arms			
Cog. Practice x Predicted Decline	0.0684**	0.0927***	0.0662**
	(0.0273)	(0.0285)	(0.0269)
Cog. Practice	-0.0010	-0.0050	0.0012
	(0.0057)	(0.0061)	(0.0053)
Panel B: Disaggregated Treatment Arms			
Math Practice x Predicted Decline	0.0692**	0.0976***	0.0664**
	(0.0308)	(0.0329)	(0.0306)
Games Practice x Predicted Decline	0.0677**	0.0881***	0.0670**
	(0.0311)	(0.0329)	(0.0309)
Math Practice	-0.0006	-0.0050	0.0018
	(0.0065)	(0.0069)	(0.0060)
Games Practice	-0.0015	-0.0050	0.0006
	(0.0066)	(0.0070)	(0.0062)
Observations	329349	329349	329349

Notes: Panel A estimates treatment effects for both treatments pooled relative to the control group. Panel B shows effects for the Math Practice and Games Practice sub-treatments (each relative to the control group) separately. "Cognitive Practice" is a binary indicator that equals 1 if the student was assigned to a treatment (either the Math or Games Practice)."Predicted Decline" is the amount of average decline in each quintile of the test location, relative to the first quintile of the test, overall (column 1), within each school (column 2), and within each school-test (column 3). Question item order was randomized across students. All regressions contain question difficulty controls and baseline controls. Observations are at the student-test-question level. Standard errors clustered by student.

Table A.8: Treatment Effects on Decline in Performance Over the Length of the Test - Restricting to Attempted Questions

	Dependent Variable: 1[Question Correct]				
	Test Subject				
	All	All	Math	Listening	Ravens
	(1)	(2)	(3)	(4)	(5)
Panel A: Pooled Treatment Arms					
Cog. Practice x Deciles 6-10	0.136***				
	(0.0045)				
Cog. Practice x Predicted Decline		0.0922***	0.1064**	0.0671**	0.0983**
		(0.0279)	(0.0432)	(0.0327)	(0.0455)
Cognitive Practice	-0.0006	-0.0025	-0.0021	-0.0014	-0.0043
	(0.0060)	(0.0061)	(0.0092)	(0.0067)	(0.0101)
Panel B: Disaggregated Treatment Arms					
Math Practice x Deciles 6-10	0.0133**				
	(0.0053)				
Games Practice x Deciles 6-10	0.0138***				
	(0.0051)				
Math Practice x Predicted Decline		0.1014***	0.1005**	0.0883**	0.1192**
		(0.0323)	(0.0484)	(0.0390)	(0.0543)
Games Practice x Predicted Decline		0.0829**	0.1132**	0.04657	0.0769
		(0.032)	(0.0504)	(0.0380)	(0.0519)
Math Practice	0.0009	-0.0038	0.0026	-0.0050	-0.0089
	(0.0068)	(0.0071)	(0.0103)	(0.0077)	(0.0120)
Games Practice	-0.0021	-0.0015	-0.0071	0.0022	0.0004
	(0.0070)	(0.0071)	(0.0107)	(0.0078)	(0.0115)
Observations	279570	279570	150777	66929	61864

Notes: This table examines the impact of cognitive practice on the rate of decline in performance over the course of the exam, restricting to only question items in which the student provided an answer. Panel A estimates treatment effects for both treatments pooled relative to the control group. Panel B shows effects for the Math Practice and Games Practice sub-treatments (each relative to the control group) separately. Col. (1) corresponds to the specification in Equation 1. Col (2) corresponds to the specification in Equation 3. "Cognitive Practice" is a binary indicator that equals 1 if the student was assigned to a treatment (either the Math or Games Practice)."Predicted Decline" is the amount of average decline in each quintile of the test location, relative to the first quintile of the test, within each given school. "Deciles 6-10" is a binary indicator that equal one if the question appears in the second half of the test. The omitted category are the questions in decile 1 (i.e. the beginning) of the test. Cols. (1) and (2) estimate treatment effects for all three tests pooled. Cols. (3), (4), and (5) show effects for the Math, Listening, and Ravens tests separately, respectively.The Coefficients in Cols. (3)-(5) are estimated from a single regression on all the data. "Control Decline" captures the average score in the first quintile of the test minus the fifth quintile of the test for students in the control group,controlling for question fixed effects. Question item order was randomized across students. All regressions contain question and test version fixed effects, and baseline controls. Observations are at the student-test-question level. Standard errors clustered by student. $^{*}p < 0.10,^{**}p < 0.05,^{***}p < 0.01$.

Table A.9: Declines in Performance - Robustness to Choice of Controls

	Dependent Variable: 1[Question Correct]					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Pooled Treatment Arms						
Cognitive Practice x Deciles 6-10	0.0121**	0.0129***	0.0125***			
	(0.00474)	(0.00474)	(0.00474)			
Cog. Practice x Predicted Decline				0.0903***	0.0927***	0.0527**
				(0.0292)	(0.0285)	(0.0219)
Panel B: Disaggregated Treatment Arms						
Math Practice x Deciles 6-10	0.0112**	0.0127**	0.0122**			
	(0.00548)	(0.00548)	(0.00547)			
Games Practice x Deciles 6-10	0.0130**	0.0131**	0.0127**			
	(0.00541)	(0.00539)	(0.00540)			
Math Practice x Predicted Decline				0.0967***	0.0976***	0.0466*
				(0.0338)	(0.0329)	(0.0251)
Games Practice x Predicted Decline				0.0842**	0.0881***	0.0591**
				(0.0336)	(0.0329)	(0.0246)
Observations	329349	329349	329349	329349	329349	329349

Notes: Panel A estimates treatment effects for both treatments pooled relative to the control group. Panel B shows effects for the Math Practice and Games Practice sub-treatments (each relative to the control group) separately. Col. (1) - (3) correspond to the specification in Equation 1. Col (4) - (6) correspond to the specification in Equation 3. "Cognitive Practice" is a binary indicator that equals 1 if the student was assigned to a treatment (either the Math or Games Practice). "Deciles 6-10" is a binary indicator that equal one if the question appears in the second half of the test. The omitted category are the questions in decile 1 (i.e. the beginning) of the test. "Predicted Decline" is the amount of average decline in each quintile of the test location, relative to the first quintile of the test, within each given school. Col. (1) and (4) control for version, col (2) and (5) add in question fixed effects and col (3) and (6) add in student fixed effects. Question item order was randomized across students. All regressions contain question difficulty controls and baseline controls. Observations are at the student-test-question level. Standard errors clustered by student. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

Table A.10: Heterogeneous Treatment Effects on Declines and School Performance

Panel A: Decline on Listening, Math and Ravens Tests

	Dependent Variable: 1[Question Correct]			
	(1)	(2)	(3)	(4)
Covariate:	Grade	Female	Baseline Mean	Baseline Decline
Cog. Practice x Predicted Decline	0.1430**	0.0919**	0.1295	0.0809***
	(0.0724)	(0.0378)	(0.0907)	(0.0290)
Cog. Practice x Predicted Decline x Covariate	-0.0173	0.0086	-0.0814	0.0543
	(0.0199)	(0.0583)	(0.1601)	(0.1553)
p-value: Cog. Practice x Pred. Decline + Cog. Practice x Pred. Decline x Covariate = 0	0.0213	0.0216	0.540	0.381
Observations	329349	325311	317346	316676

Panel B: School Tests

	Dependent Variable: Z-score of Student's Grades			
	(1)	(2)	(3)	(4)
Covariate:	Grade	Female	Baseline Mean	Baseline Decline
Cog. Practice	0.0793	0.0452	0.1138	0.0849*
	(0.1020)	(0.0561)	(0.1058)	(0.0459)
Cog. Practice x Covariate	-0.0008	0.0821	-0.0958	-0.4354*
	(0.0301)	(0.0938)	(0.1979)	(0.2276)
p-value: Cog. Practice + Cog. Practice x Covariate = 0	0.303	0.0909	0.869	0.120
Observations	11320	11162	10983	10965

Notes: The table shows whether there was a heterogeneous treatment effect of the program on decline in performance (Panel A) or students' school tests (Panel B) by student covariate. "Covariate" varies by column. In column (1), it is grade, which ranges from 1-5. In column (2), it is a binary indicator for whether the student is female. In column (3), it is student's baseline average percent of questions correct on the listening, math and Ravens tests, and in column (4), it is the difference between student's performance in the first quintile of the test versus the last quintile on the baseline tests.

Panel A: Panel A uses the same specification as in Table I col (2) and adds in an interaction term with each covariate. The dependent variable is whether the student got the question item correct. "Cog. Practice" is a binary indicator that equals 1 if the student was assigned to a treatment (either the Math or Games Practice). "Predicted Decline" is the amount of average decline in each quintile of the test location, relative to the first quintile of the test, within each given school. Question item order was randomized across students. All regressions contain question and test version fixed effects, and baseline controls. Observations are at the student-test-question level.

Panel B: Panel B uses the specification from Table III, col (1) and adds in an interaction with each covariate. The dependent variable is the students' endline score on their regular school test in z-score. All regressions include class section (strata) fixed effects and baseline controls. Observations are at the student-subject-exam level.

In both panels, standard errors are clustered by student. Additional level and interaction coefficients are not included for brevity. * $p<0.10$, ** $p<0.05$, *** $p<0.01$.

Table A.11: Effect of Treatment on Behaviors Outside of School

	(1) Breakfast	(2) Practice	(3) HW Time	(4) HW Help
Cognitive Practice	-0.0370	0.0588	-0.0174	-0.0179
	(0.0629)	(0.0761)	(0.0598)	(0.0625)
Dep. Var. Mean	1.656	3.194	1.603	1.055
Dep. Var. SD	0.760	0.927	0.743	0.735
Observations	706	706	706	706

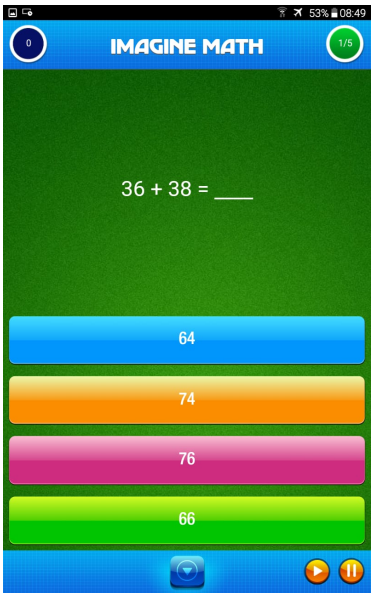
Notes: This table shows the effect of treatment assignment on behaviors outside of school, such as homework practices. "Cognitive Practice" is dummy for whether the student was assigned the treatment. "Assets" is a measure of the total number of assets students have at home from the following list: book, fan, mixer, refrigerator, phone, computer, car, motorbike. "Breakfast" is a measure of the total items a student had for breakfast that day from the following list: egg, bread, rice, paratha, cereal, milk, tea, fruit, meat, other. "Practice" is a measure of how much time the student spent after school on cognitively-focused practice activities, ranging from 0-5. "HW Time" is a measure for how much time students spend on homework from 0 (less than 45 minutes), 1 (about 45 minutes), 2 (more than 45 minutes). "HW Help" captures the number of individuals (family members, tutor, etc) that help them with their homework. All outcomes are based on student self-report from a survey conducted at the end of the study. The survey was conducted with a subset of the total sample of students who were re-surveyed at endline. Specifications include controls for student school, grade, and classroom. $^{*}p < 0.10,^{**} p < 0.05,^{***} p < 0.01$.

FIGURE A.1: Example Classrooms from Study Schools

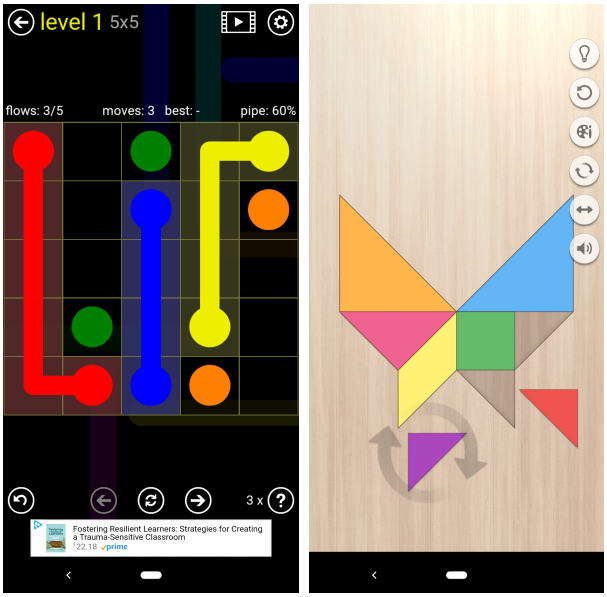
Notes: The photographs show two example classrooms from our study's schools to provide context.

FIGURE A.2: Treatment Tablet Software

(a) Math practice treatment



(b) Games practice treatment

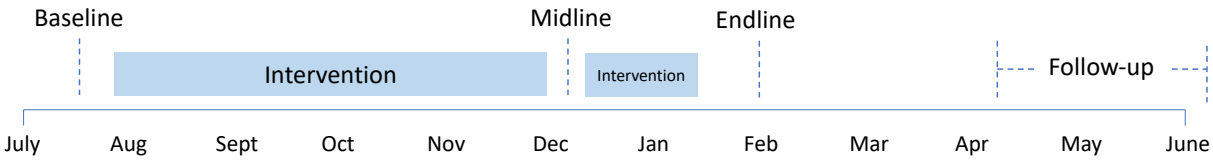


Notes: These figures show example screenshots from the treatment tablet software used throughout the intervention. For the Math Practice, we use the imagineMath software, developed by Pixatel. For the Games Practice, we use simple games with limited animation downloaded from the Android app store.

FIGURE A.3: Program Treatment Classes

Notes: The photographs show two example treatment program classes.

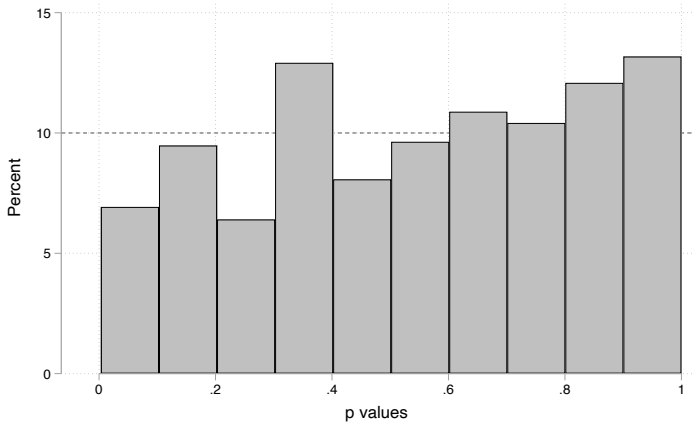
FIGURE A.4: Experiment timeline



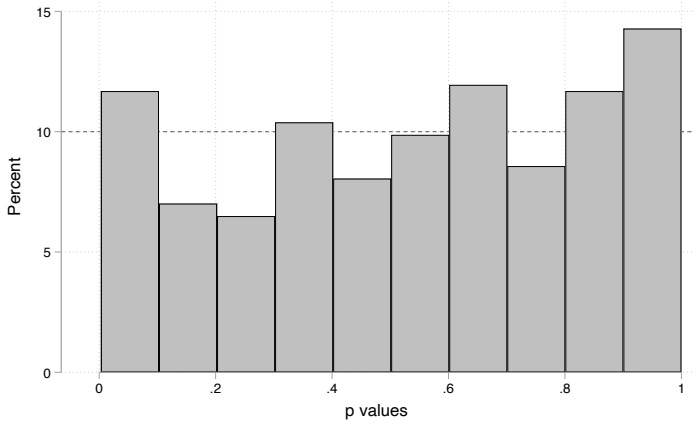
Notes: This figure shows the timeline of the intervention. Program treatment and control classes were administered from August to early December and again in January. Baseline tests were conducted in July and August (before the start of program classes). Midline tests were conducted during the intervention break in December, and endline tests were conducted in early February. Follow-up tests were conducted from late April through June. The experiment was administered from July 2017-June 2019.

FIGURE A.5: Randomization Balance of Test Versions

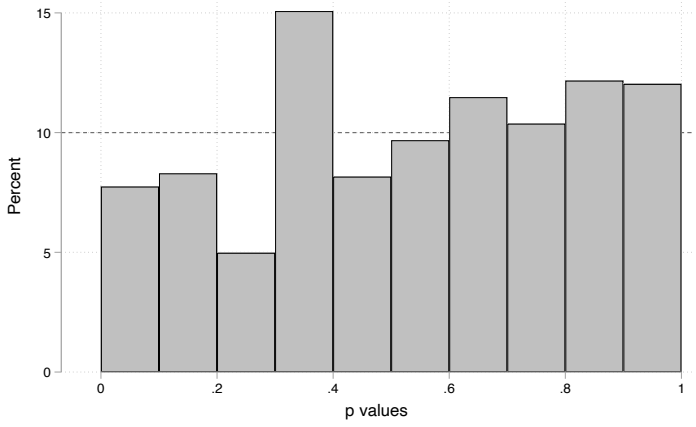
(a) Math Test



(b) Listening Test

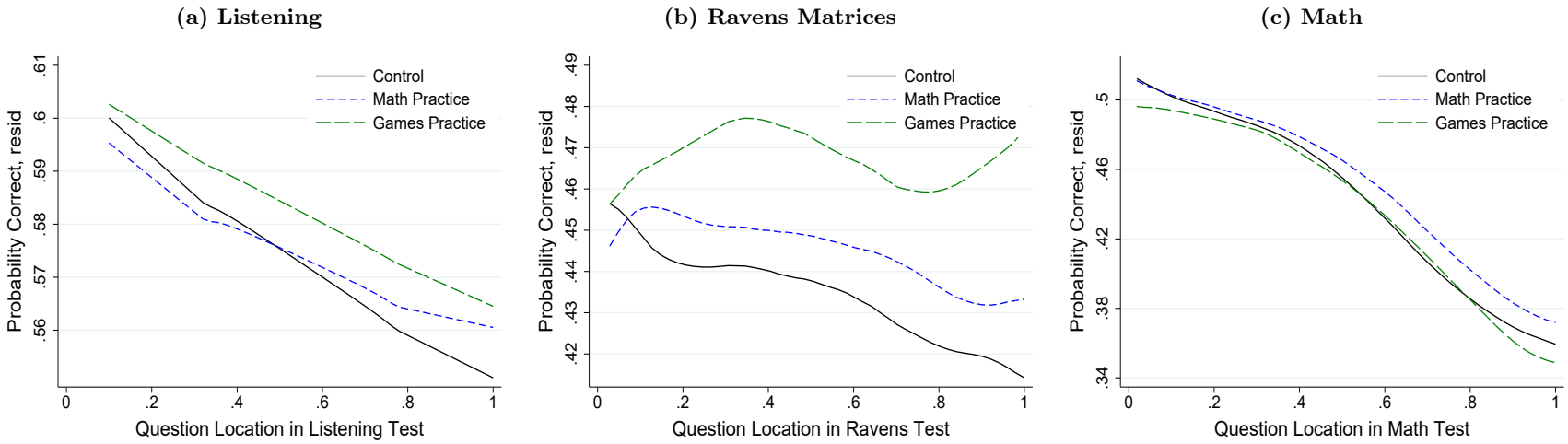


(c) Ravens Test



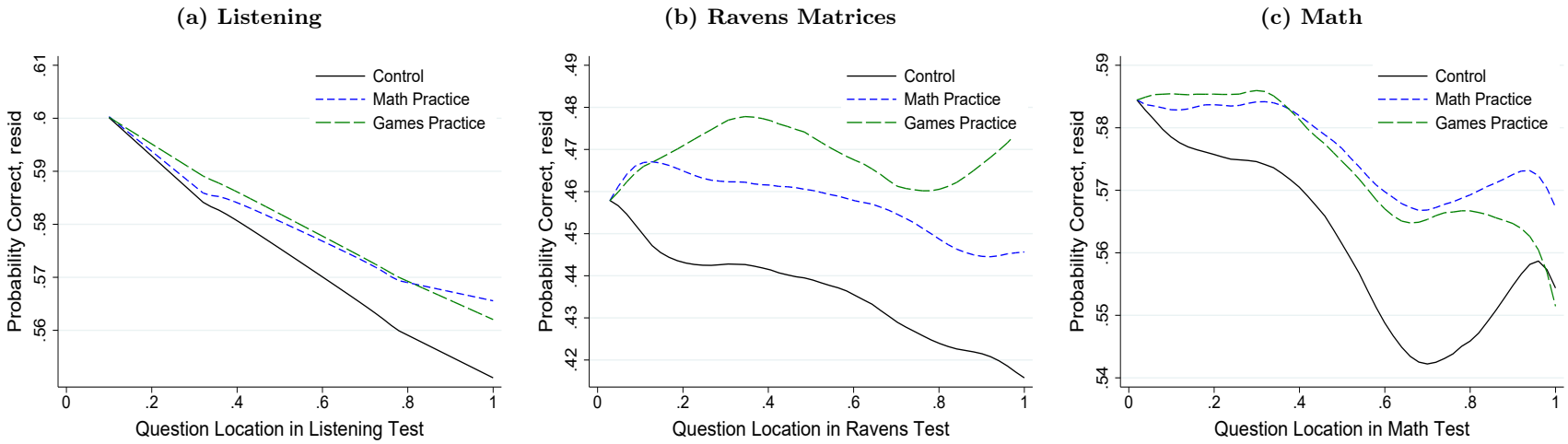
Notes: These figures show the distribution of p-values for 4,478 coefficients of whether the student received a given test version on dummies for treatment status. These regressions are calculated within each test-round-school-grade. This allows to test whether allocation of test versions as balanced across treatment status. For a perfectly random allocation of test versions, in the limit, we would expect each bar to approach 10%.

FIGURE A.6: Performance Over the Length of the Test by Treatment



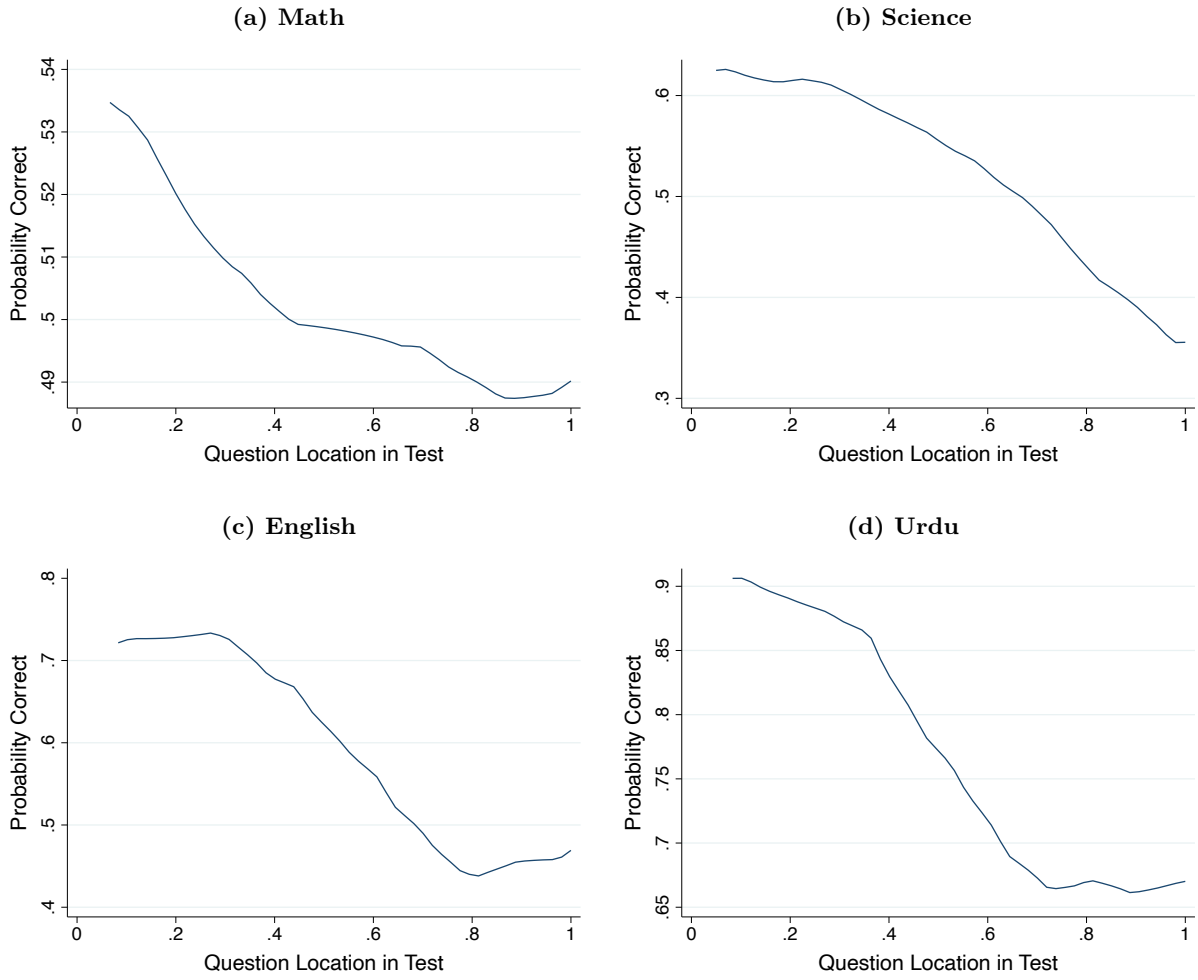
Notes: This figure plots the declines in performance across time on three tests administered as a part of the study: (a) listening, (b) Raven's Matrices, and (c) math. Question order is randomized in each exam. Each figure plots the probability a question was answered correctly (y-axis) against the percent location of the question on the test (where 0 is the beginning of the test and 1 is the end of the test, x-axis). Data is residualized to remove question fixed effects. Each line displays performance over time for the control group (solid black line), Math arm (short blue dashes), and Games arm (long green dashes), respectively. Observations are at the student-test-question level; N = 66,932 (listening), 62,183 (Raven's Matrices), and 200,234 (math). Table I presents the full set of corresponding treatment effects estimates.

FIGURE A.7: Attempted Questions Only: Performance Over the Length of the Test by Treatment



Notes: This figure plots the declines in performance on **attempted questions only** across time on three tests administered as a part of the study: (a) listening, (b) Raven's Matrices, and (c) math. Question order is randomized in each exam. Each figure plots the probability a question was answered correctly (y-axis) against the percent location of the question on the test (where 0 is the beginning of the test and 1 is the end of the test, x-axis). Data is residualized to remove question fixed effects. For each plot, the initial level at the start of the test is normalized to the control group mean in decile 1 for that test for ease of interpretation of decline magnitudes. Each line displays performance over time for the control group (solid black line), Math Practice (short blue dashes), and Games Practice (long green dashes), respectively. Observations are at the student-test-question level; N = 66,929 (listening), 61,864 (Raven's Matrices), and 150,777 (math). Table A.8 presents the corresponding treatment effects estimates.

FIGURE A.8: Performance Over the Length of the Test by Subject



Notes: This figure plots the declines in performance over the length of the exam by test subject using data from Brown and Andrabi (2021). Question order on the tests is randomized. The figure plots the probability a question was answered correctly (y-axis) against the percent location of the question on the test (where 0 is the beginning of the test and 1 is the end of the test, x-axis). Observations are at the student-test-question level; N = 217,516.