Digital Self-Command*

Mridul Joshi[†1], Jalnidh Kaur[‡2], and Lena Song[§3]

[1] *Stanford University*
[2] *University of Glasgow*
[3] *University of Illinois Urbana-Champaign*

November 4, 2026

Abstract

Can self-command, the capacity to exercise informed judgment and follow through on one's intentions, be taught for digital technology use? We design a classroom intervention teaching digital self-command and test it in a cluster-randomized trial with college students in India. The intervention reduces objectively measured social media use by around ten percent. It modestly improves a mental health index by lowering digital addiction and loneliness, with no effect on anxiety or depression. It also increases discernment between true and false claims, but only with targeted instruction on misinformation. Our findings suggest education could complement social media regulations while preserving individuals' agency over their technology use.

JEL Classification: C93, D91, I12, I21, I25, I31, L86
Keywords: digital literacy, misinformation, self-control, smartphones, social media

*The study is pre-registered at the AEA RCT Registry (AEARCTR-0013341). We thank Aaron Chen, Harshiv Chhabra, Hermogenes David Rios, Rishav Ray, and especially Sukhmani Kaur, Durgesh Khade, and Sara Hercules Nunez for outstanding research assistance, and Archana M V and Manvi Govil for excellent research management. We are grateful to J-PAL South Asia, and our partner universities for making this study possible, and to Dr Sarabjeet Singh for his support throughout the study. We are grateful to Emily Weinstein from Center for Digital Thriving, Harvard University, Manik Inder Singh from NIMHANS India, and to Gauravdeep Singh from Initiators of Change for providing feedback on the curriculum. We extend our thanks to Chris Karr and Audacious Software for building Dashboard. This work was supported by the USAID Development Innovation Ventures; the Weiss Fund; the University of Illinois Urbana-Champaign; UIUC Campus Research Board; Stanford King Center for Development; Stanford McCoy Family Center for Ethics in Society; Stanford Freeman Spogli Institute for International Studies; and Center for Social & Behavioral Science at UIUC. Joshi gratefully acknowledges support from the Dixon and Carol Doll Graduate Fellowship through a grant to the Stanford Institute for Economic Policy Research. We obtained IRB approval from Krea University/IFMR, Stanford University, and University of Illinois Urbana-Champaign. All errors are our own.

[†]Email: `mriduljoshi@stanford.edu`
[‡]Email: `jalnidh.kaur@glasgow.ac.uk`
[§]Email: `lenasong@illinois.edu`

1 Introduction

People do not always use digital technologies as they intend. Smartphones and social media are important to how people connect, but they also create environments in which it is hard to exercise informed judgment and follow through on one's intentions. Features such as infinite scroll, autoplay, and engagement-maximizing algorithms make it easier to keep using social media than to stop, and to accept claims than to verify them. Concerns about the adverse effects of these technologies have prompted supply-side policy responses, such as restrictions on platform design and access (OECD, 2026). Another proposed response is education (Weinstein and James, 2022; Office of the Surgeon General, 2023), but there is little causal evidence on its effects.

This paper studies whether an educational program can build the capacity to use smartphones and social media more intentionally. Borrowing the term self-command from Adam Smith, we refer to this capacity as *digital self-command*.[1] We define digital self-command as the capacity to exercise informed judgment, to anticipate and prepare for situations in digital environments where the immediately appealing response conflicts with one's intentions, and to follow through on those intentions when such situations arise.

We design a program for college students that builds awareness and gives them practical strategies to manage their smartphone and social media use and critically assess information they encounter online. The curriculum consists of three sessions. The first session provides information about the design of digital platforms, raises awareness of self-control problems, and develops attentiveness to habit formation. The second session equips learners with tools and strategies, such as commitment devices to align their behavior with the preferences of the long-run self. The third session offers information and tools to discern misinformation online. While social media and smartphone use can be addictive (Allcott, Gentzkow, and Song, 2022), these technologies also serve important productive and social purposes. The program does not suggest that all levels of digital use are harmful, but instead encourages active decisions and greater agency in place of passive consumption. It targets not whether, but how students use technology.

We develop a simple framework to conceptualize digital self-command and what an educational intervention could impact. At any point, a student can continue using a smartphone app or stop and do something else. We summarize the circumstances of this choice by a state. Some states are more tempting than others, raising the value the student places on use above the value she would have assigned in advance. The student may misperceive temptation or future payoffs. Teaching self-command can therefore operate through better recognition of tempting states, better understanding of the future payoffs, and greater adoption of strategies that allow the student to avoid tempting states, weaken the pull within them, or make it easier to choose something else.

We evaluate the program using a cluster-randomized trial with 3,430 college students across five higher-education institutions in Punjab, India. We randomize 86 clusters at the institution-major-

[1] In *The Theory of Moral Sentiments*, Smith distinguishes knowing a rule from having the self-command to act on it (Smith, 1759, VI.iii). He argues that knowledge may not be sufficient because passions can "drive" or "seduce" a person to violate rules he approves of in his "sober and cool hours" (Smith, 1759, VI.iii.1).

year level into one of control, *Deliberate Use*, or *Deliberate Use & Discernment* arms. Students assigned to *Deliberate Use* receive the first two sessions, while students assigned to *Deliberate Use & Discernment* also receive the third session on misinformation. We randomize at the level at which the program would be taught in practice, so our estimates capture the policy-relevant effect of treating an entire class.

Our pre-analysis plan specifies social media consumption, a mental health index (subjective well-being, digital addiction, depression, and anxiety), and misinformation discernment as primary outcomes. A key feature of our study is the objective measurement of smartphone and social media use, rather than reliance on self-reports, which are often noisy and inaccurate (Parry et al., 2021). Students in both treatment and control are encouraged to install a custom Android app that collects timestamped system usage records from the preceding week, from which we construct measures of total screen time, usage by app and app category, and usage by time of day. We complement these data with student surveys and administrative records on class schedules and academic performance. We survey students at baseline, endline, and approximately a year after the intervention, and administer three short check-in surveys during the intervention to track short-run dynamics.

The program reduced time spent on the phone relative to the control group, driven by a reduction in social media use. At endline, objectively measured total phone use is nine percent lower than in the control group, or about 32 minutes per day. Social media use is 10 percent lower, a reduction of about 25 minutes per day. We find no detectable impact on non-social media use.

We find that the program leads to a small improvement in the mental health index, driven primarily by reduced digital addiction. Students also report feeling less lonely. We find little change in symptoms of anxiety and depression, happiness, or life satisfaction. The *Deliberate Use & Discernment* arm also improves students' accuracy in evaluating online claims, raising discernment by 0.18 SD, while the *Deliberate Use* arm has no detectable effect on discernment.

Beyond the primary outcomes, treated students allocate more time to sleep and perform better on exams. Students report sleeping longer, potentially due to time freed up by reduced screen time, with suggestive improvements in sleep quality. Using administrative data, we find that students in the treatment group are 5 percentage points less likely to fail their university exams.

We test the channels in the conceptual framework and find evidence that teaching digital self-command improves understanding of digital environments, encourages the adoption of strategies and tools to manage use, and enhances students' sense of agency. At endline, treated students score higher on a quiz covering the curriculum's core concepts, such as platform design, habit formation, and commitment devices. They also perform better on vignettes that test their ability to diagnose and respond to digital situations, consistent with better recognition of tempting states and understanding of future payoffs. They are more likely to express interest in setting limits and to report using commitment devices, consistent with greater strategy adoption. Students also report greater internal locus of control, consistent with a stronger sense of agency over their choices. Combining objective use with course timetables, we find suggestive evidence that treated students

reduce social media use more during scheduled class time than during breaks, relative to other phone use, consistent with students managing phone use when it could conflict with their academic goals. At the same time, the payment students require to deactivate Instagram, elicited with real stakes, does not fall. The reduction in use therefore reflects better self-command over a technology students still value.

The patterns over time further shed light on how self-command can be built. The addiction index initially rises during the program before decreasing at endline. This suggests that the program makes students more aware of how difficult it is to manage their phone use before they learn and adopt relevant strategies and tools. On discernment, the *Deliberate Use* arm has no detectable effect at any point during the intervention, while *Deliberate Use & Discernment* improves it only after the misinformation module is taught, consistent with a role for targeted instruction.

Nearly a year after the program, some effects persist. Students in the treatment group report lower digital addiction, score higher on digital literacy, and report better sleep quality. The gains in misinformation discernment are no longer detectable. The estimates for objectively measured total and social media use are negative but imprecise, and we cannot distinguish them from either zero or the endline estimates. By follow-up, generative AI use had become widespread among students. We use AI to test whether learning transfers to a technology the curriculum never covered. Treated students who use it report using it more purposefully, especially to prepare for exams. This suggests that digital self-command may extend beyond technologies directly covered in the curriculum.

We also provide evidence on how the program could be delivered at scale. To test whether the program works simply by providing information and reminders, we embedded in the follow-up survey a microlearning experiment that delivered the key points from the curriculum through short, self-guided lessons. The information intervention raises digital literacy but does not move beliefs or behavioral intentions, suggesting that the effects of the program are not driven by information alone. Greater exposure to treated friends is associated with higher internal locus of control among students in the treatment group, while the analogous estimate among control students is small and imprecise. This suggests that classroom delivery may work in part by strengthening peer effects among those who receive it. The program is also short, modular, and designed to fit within existing educational institutions, and the complete three-session curriculum costs approximately USD 3.43 per student. In the framework of List (2024), the program could be scaled by leveraging existing school infrastructure and preserving facilitated, group-based delivery.

Our study contributes to the literature on self-control problems and interventions designed to address them. A large fraction of social media use is caused by self-control problems exacerbated by habit formation (Allcott, Gentzkow, and Song, 2022). Commitment devices can help people follow through on their intentions. Earlier studies document demand for and impact of commitment devices to manage smartphone and social media use (e.g., Ek and Samahita 2023; Allcott, Gentzkow, and Song 2022; Hoong 2021), while soft commitments can improve classroom focus among university students (Aksoy, Lusher, and Carrell, 2025).[2] Yet the availability of commitment devices does not

[2]Field evidence in other settings documents demand for commitment devices in savings, labor effort, health

ensure their effective use. Individuals may fail to recognize when they need these tools, even after repeated experience, either because they are naive about their own self-control problems (Ericson and Laibson, 2019) or because they are inattentive to certain features of their environment until these are made salient (Hanna, Mullainathan, and Schwartzstein, 2014; Augenblick et al., 2024). These frictions may prevent users from adopting commitment devices. Our intervention provides no new commitment device and instead teaches individuals to recognize tempting moments and use available tools. We show that education increases reported use of these tools and reduces objectively measured social media use.

A growing literature studies classroom interventions to cultivate skills and influence behavior. Prior work shows that school-based curricula can shape grit (Alan, Boneva, and Ertac, 2019), curiosity (Alan and Mumcu, 2024), perspective-taking and social cohesion (Alan et al., 2021), and media literacy (Briole, Cagé, and Prat, 2026).[3] More closely related to self-command are interventions targeting patience (Alan and Ertac, 2018; Chuan, Samek, and Samujjwala, 2024), self-regulation (Schunk et al., 2022), and forward-looking decision-making (Alan, Ertac, and Gümüş, 2021), all relevant to navigating conflicts between immediate temptations and longer-term intentions. Outside economics, two recent studies evaluate high school programs designed to reduce problematic smartphone and internet use, finding mixed effects on self-reported outcomes (Gui et al., 2023; Otsuka et al., 2023). We examine whether a digital self-command curriculum can influence behaviors in adulthood using a broad set of outcomes beyond self-reports. We also show that delivery matters, suggesting that the effects of classroom instruction are not driven by information alone.

Our *Deliberate Use & Discernment* arm builds on a large literature testing ways to counter misinformation (see Zhuravskaya, Petrova, and Enikolopov 2020; Aridor et al. 2024; Kozyreva et al. 2024; Blair et al. 2024 for reviews).[4] Closely related, Amar et al. (2026) find that a four-month classroom curriculum in rural India improved students' ability to discern true from false claims, although evidence from India on digital literacy interventions is mixed (Guess et al., 2020; Badrinathan, 2021). In our study, we find that better evaluation of online information requires targeted instruction on misinformation.

Finally, our findings have direct implications for policies surrounding smartphones and social media. High-speed internet and social media use have been shown to worsen mental health (Donati et al., 2022; McDool et al., 2020; Braghieri, Levy, and Makarin, 2022).[5] Existing policies, such as age limits and phone bans, have largely focused on restricting access.[6] Phone bans have been shown to reduce phone use with mixed effects on well-being and academic outcomes (Abrahamsson,

behaviors, and education (e.g., Ashraf, Karlan, and Yin 2006; Kaur, Kremer, and Mullainathan 2015; Schilbach 2019; Bai et al. 2021; Carrera et al. 2022; Himmler, Jäckle, and Weinschenk 2019).

[3]Other related work uses interventions focused on beliefs about agency, efficacy, and aspirations to shape behavior (Orkin et al., 2025; McKelway, 2025; Kaur, 2025).

[4]Session 3 draws on evidence-based approaches to evaluating online information (List et al., 2024; Guriev et al., 2023; Guess et al., 2020; Athey et al., 2023; Ali and Qazi, 2023; Berger et al., 2023; Bowles et al., 2023).

[5]These effects may be amplified in developing countries, where access to technology is rapidly increasing but resources and support systems for responsible use are lagging behind (World Bank, 2016).

[6]Australia banned social media accounts for children under 16 in December 2025, and many other jurisdictions have adopted or proposed similar measures (OECD, 2026).

2026; Figlio and Özek, 2025; Lichand et al., 2026; Sungu, Choudhury, and Bjerre-Nielsen, 2025; Allcott et al., 2026). Yet restrictions are hard to enforce outside of supervised settings, may have low compliance, as in the case of Australia's social media ban (Bursztyn et al., 2026), and become less feasible as students move into higher education and adulthood. We demonstrate that, without restricting access, education can change objectively measured smartphone and social media use, reduce digital addiction, and lower exam failure. These findings suggest that building individuals' digital self-command may provide a policy lever alongside existing social media regulations.

The paper is organized as follows. Section 2 develops a conceptual framework of digital self-command, and Section 3 describes the institutional context and program design. Section 4 presents the experimental design, data, sample construction, and baseline characteristics, and Section 5 describes the estimation strategy. Section 6 presents the results. Section 7 examines mechanisms, and Section 8 concludes.

2 Self-Command in Digital Environments

We define *digital self-command* as the capacity to exercise informed judgment, to anticipate and prepare for situations in digital environments where the immediately appealing response conflicts with one's intentions, and to follow through on those intentions when such situations arise. We study this capacity in students' use of digital technology and their evaluation of online information.

Our use of the term *self-command* draws on Adam Smith and Thomas Schelling. In *The Theory of Moral Sentiments*, Smith describes self-command as the virtue of governing the passions so that the judgment a person forms in "sober and cool hours" continues to guide conduct when the passions would "mislead" them (Smith, 1759, VI.iii). Schelling uses the term for anticipatory action taken to govern future behavior when future and present preferences are not aligned (Schelling, 1984). Self-command is closely related to, but broader than, self-control. Self-control problems involve conflict between advance intentions and later choices (Laibson, 1997; O'Donoghue and Rabin, 1999).[7] This literature mostly takes intentions as given and distinguishes self-control problems from choices based on incorrect beliefs about costs and benefits (Duckworth, Milkman, and Laibson, 2018). Smith notes that the passions can distort judgment as well as conduct (Smith, 1759, III.4). Following Smith, we use self-command as an integrative term that combines self-control with informed judgment, so that one's intentions reflect an accurate view of what is at stake.

To fix ideas, consider a student deciding whether to continue using a smartphone application or stop and switch to another activity.[8] We summarize the circumstances of this choice by a state s. Before the state arises, the student may adopt a strategy m, with $m = 0$ denoting no strategy.

[7] Self-control can refer to the costly act of resisting an available temptation (Gul and Pesendorfer, 2001), advance strategies that preempt or reshape such conflicts (Thaler and Shefrin, 1981; Duckworth, Gendler, and Gross, 2016; Duckworth et al., 2016; Duckworth, Milkman, and Laibson, 2018), or an individual trait that predicts a range of outcomes (Tangney, Baumeister, and Boone, 2004; Cobb-Clark et al., 2026).

[8] We focus on continued use because it makes the cost of disengaging from smartphone or social media explicit. If use has not begun, the analogous choice is between opening the application and continuing with the alternative activity. The same payoff comparison applies, except that the friction now works in the other direction, making opening harder rather than stopping easier.

A strategy is a tool or plan chosen in advance to help the student act in line with her goal when temptation arises. Continuing yields an ordinary net immediate benefit $b(s)$, such as entertainment or social connection, net of the value of the best available alternative at that moment. Continuing may also carry an additional in-the-moment temptation wedge, $\gamma(m, s) \geq 0$, shaped by cues in the state that increase the pull to continue using and by any strategy already in place (Laibson, 2001; Bernheim and Rangel, 2004). For example, boredom, fatigue, or emotionally charged content may make continuing more attractive in the moment than it appeared in advance.[9] The wedge separates two valuations of the same choice: in advance the student values continuing at $b(s)$, but in the moment she chooses as if it were worth $b(s) + \gamma(m, s)$, and so continues more often than she would have chosen in advance. It shifts the choice without adding to its value, unlike $b(s)$, and we take the advance valuation, not the one she acts on, as the reference when we describe use as conflicting with a student's intentions.

Let $V(s)$ denote the expected discounted continuation payoff after stopping minus the expected discounted continuation payoff after continuing in state s. This difference may reflect effects on downstream outcomes like health or academic outcomes, as well as dynamic effects on future use. Continuing, for example, may make tempting states more frequent (e.g., via algorithmic recommendation).[10]

We also allow for misperception. Before state s arises, the student forecasts its temptation wedge as $\widetilde{\gamma}(m, s)$, which may differ from the true wedge $\gamma(m, s)$, either because she underestimates the pull or because she misperceives the strategy's effectiveness in reducing it. She also perceives the future gain from stopping as $\widetilde{V}(s; \widetilde{\gamma})$, which may differ from $V(s)$. This perceived gain depends partly on her forecast of future temptation because that forecast shapes the choices she expects to make later. Underestimating temptation can therefore make stopping appear less valuable. Even given her forecast $\widetilde{\gamma}$, the perceived gain $\widetilde{V}(s; \widetilde{\gamma})$ may differ from $V(s)$ if she misjudges how continued use affects downstream outcomes or habit formation.

Conditional on reaching state s with both continuing and stopping available, let $c(m, s) \geq 0$ denote the immediate cost of stopping with strategy m in place. This cost captures the effort of disengaging from the application and redirecting attention to another activity. It is higher when use lacks a natural stopping point, as with infinite scroll or autoplay, and lower when a strategy creates one. When the state occurs, the student continues when

$$b(s) + \gamma(m, s) > \widetilde{V}(s; \widetilde{\gamma}) - c(m, s). \tag{1}$$

The left-hand side is the immediate payoff from continuing. The right-hand side is the future gain from stopping rather than continuing, as perceived by the student, net of the immediate cost of doing so. The true wedge $\gamma(m, s)$, rather than the student's earlier forecast $\widetilde{\gamma}(m, s)$, enters

[9] Allcott, Gentzkow, and Song (2022) estimate a temptation parameter for social media use in general. Here we allow it to vary across states, with cues the student can learn to recognize.

[10] Formally, if $V^0(s)$ and $V^1(s)$ denote continuation values after stopping and continuing, respectively, then $V(s) = V^0(s) - V^1(s)$. These continuation values incorporate the future states and choices generated by the current action. To keep notation light, we suppress their dependence on the strategy m already in place.

equation (1) because the choice is made after the state arises. A student may therefore expect in advance to stop but continue if the pull in that state is stronger than she anticipated.

A strategy can be beneficial in three ways. First, for a given state s, it can lower the cost of stopping. An application limit, for example, creates a stopping point, so that $c(m, s) < c(0, s)$. Second, a strategy can weaken the pull the student feels once the state arises, so that $\gamma(m, s) < \gamma(0, s)$. For example, setting the screen to greyscale leaves the state intact but reduces what makes continuing tempting within it. Finally, a strategy can change which states arise or whether a later use decision arises at all. Scheduled *Do Not Disturb*, for example, reduces the likelihood of a notification-triggered state, while keeping the phone elsewhere can prevent a late-night use decision from arising.

The earlier adoption decision depends on how the student expects m to alter these later choice problems. Let $\Pi(m; \widetilde{\gamma}, \widetilde{V})$ denote her perceived net expected gain from adopting m, relative to having no strategy. It includes the expected benefits generated through the three channels above, net of the costs of adopting and maintaining the strategy. The student adopts m when

$$\Pi(m; \widetilde{\gamma}, \widetilde{V}) \geq 0. \tag{2}$$

Equation (2) formalizes the anticipatory decision emphasized by Schelling (1984), whether to put a strategy in place before future states arise. Equation (1) formalizes the in-the-moment decision emphasized by Smith (1759), governing behavior when the student reaches a state in which continuing and stopping both remain available.[11]

2.1 Teaching Digital Self-Command

Education is a plausible lever for building both components of self-command. The first role of education is to improve recognition of tempting states. Students may already believe that they spend too much time on social media and that these applications are tempting. The program targets something more specific by teaching students to recognize which cues are likely to increase the pull to continue in a future state. In the model, this can bring the student's forecast $\widetilde{\gamma}(m, s)$ closer to the true temptation wedge $\gamma(m, s)$. Because this forecast shapes the choices she expects to make later, better anticipation can also affect $\widetilde{V}(s; \widetilde{\gamma})$ and the perceived expected gain from adopting a strategy, $\Pi(m; \widetilde{\gamma}, \widetilde{V})$.

The second role is to reduce misperception about future payoffs. Students may misjudge how continued use affects sleep, attention, study progress, or habit formation. Holding the temptation forecast fixed, education can bring the student's perceived future gain from stopping, $\widetilde{V}(s; \widetilde{\gamma})$, closer

[11] The same structure can be applied to evaluating online content. Consider a student deciding whether to accept or share a post right away or to stop and verify its claims. Engaging right away yields an immediate benefit $b(s)$, and tempting content, such as emotionally charged posts, adds a wedge $\gamma(m, s)$. The gain from verifying, $V(s)$, includes holding more accurate beliefs and avoiding the private and social costs of spreading false information, and $c(m, s)$ is the effort cost of inspecting the source or looking for corroborating evidence. Verification routines learned in advance are strategies m that lower this cost. For example, a rule to pause before sharing creates a stopping point, similar to an app limit. Other strategies, such as unfollowing sources of untrustworthy content, reduce exposure to tempting content.

to the true gain $V(s)$. This need not reduce the ordinary benefit of digital use, $b(s)$. Because the trade-off is state-contingent, the same technology may be valuable in one state and conflict with the student's goals in another. Education can therefore change when and how students use digital technology without necessarily reducing the value they place on it.

The third role is to encourage strategy adoption and support follow-through in the moment. Adoption is more likely when the student knows about a strategy, anticipates the states in which it will be useful, and believes that it can change her behavior or outcomes. Education can bring a strategy into consideration and increase $\Pi(m; \widetilde{\gamma}, \widetilde{V})$ by clarifying when it is useful, increasing its perceived effectiveness and sense of agency, or reducing unfamiliarity and frictions in set-up. Perceived effectiveness is the improvement the student expects the strategy to deliver, such as the gap $\widetilde{\gamma}(0, s) - \widetilde{\gamma}(m, s)$ between the pull she expects to feel with and without it. Perceived agency affects the same comparison: a student who believes that her own actions cannot change outcomes will perceive little net gain from an application limit, rule, or routine. In the moment, a conditional plan can link a recognizable cue to a specified response, weakening the pull that remains once the state occurs (Gollwitzer, 1999).

Taken together, these roles clarify why information alone may be insufficient. Knowing that social media can be distracting is not the same as anticipating a particular tempting state, understanding what is at stake, adopting an appropriate strategy, and carrying it out when the state occurs. Effects may therefore be concentrated in states where use conflicts with a valuable alternative, such as during class, and a strategy can materially alter the later choice.

3 Institutional Context and Program Design

3.1 Setting

India is one of the world's largest digital markets, with approximately 1.03 billion internet users and 500 million active social media users in October 2025 (Kemp, 2026). Digital adoption has been rapid, reaching young people who came of age alongside these platforms, yet the educational and regulatory infrastructure to support healthy technology use remains thin. Fewer than a third of young adults who use digital devices report being able to verify the reliability of online information (Suresh, 2026). The Government of India's *Economic Survey 2025–26* also identifies digital addiction among young people as a policy concern, citing lost study hours and reduced productivity (Government of India, Ministry of Finance, 2026).

Our study takes place in Punjab, a northwestern state of India. Digital access is high relative to the national average: as of March 2026, Punjab had 96 internet subscribers per 100 people, compared with 77 nationally (TRAI, 2026). In 2023, about 90 percent of surveyed rural households with 14–18 year olds had a smartphone at home (ASER, 2023). We collaborate with five higher-education institutions in and around Ludhiana, one of Punjab's largest urban and industrial centers.[12]

[12]To be included in the study, institutions formalized their commitment by signing a Memorandum of Understanding

3.2 Program Contents

Our curriculum combines publicly available practitioner materials with scientific evidence.[13] We adapted the curriculum for delivery in higher-education classrooms in Punjab, since existing digital citizenship and digital-wellness curricula mostly target school-age students in high-income settings.[14] As part of curriculum design, we held focus group discussions with out-of-sample students. We adapted the content to be age-appropriate and contextually relevant by selecting local examples, incorporating familiar digital scenarios, and adding Punjabi translations for key terms. We also revised the classroom materials after feedback from content experts and local practitioners.[15]

The curriculum is structured into three modules. In Session 1, *Digital Media and Our Minds*, students learn how digital-media platform design shapes attention, self-control problems, and habit formation. Using relatable examples, the session introduces the hook model and explains how features such as variable rewards, infinite scroll, autoplay, notifications, and advertisements can keep users engaged and make stopping more difficult. Students connect these features to familiar experiences such as fear of missing out, difficulty disengaging, social comparison, and other thinking traps associated with social media use. The session also explains how naiveté about self-control problems and inattention to habit formation can lead students to mispredict their future use and overconsume digital media. Students then apply these ideas to their own lives by identifying tempting moments, such as checking notifications while studying, scrolling before sleep, or continuing to watch videos after intending to stop.

In Session 2, *Taking Charge of Our Digital Lives*, students learn to translate these ideas into practical strategies. Students learn how commitment devices can help manage self-control problems by moving decisions earlier, before tempting moments arise. The session covers tools already available to most students, including app limits, scheduled *Do Not Disturb*, notification controls, advertisement blockers, keeping the phone away during study or sleep, and setting concrete goals for use. These tools are presented as planning aids rather than signs of weak willpower, and students learn to deploy them in the moments where temptation is most predictable. The session also introduces goal setting as an internal commitment device that can help students follow through on their plans. Students are asked to match strategies to specific moments, such as using app limits during study time, turning off notifications during class, keeping the phone away before sleep, or writing down goals before beginning a work session.

Finally, in Session 3, *What's Fact? What's Fiction?*, students learn how misinformation spreads through social media and why individuals may be vulnerable to inaccurate or misleading content.

with J-PAL South Asia and agreed to facilitate research activities on their premises.

[13]The practitioner materials include resources from Common Sense Media (`https://www.commonsensemedia.org`), the Digital Wellness Institute certificate program (`https://www.digitalwellnessinstitute.com`), the Digital Inquiry Group (`https://inquirygroup.org/`), the University of Iowa Libraries' guide on misinformation (`https://guides.lib.uiowa.edu/c.php?g=849536&p=6077637`), and Boom Live's Ultimate Fact Checking Kickstarter course (`https://workshop.boomlive.in/courses/Ultimate-Fact-Checking-Kickstarter-627b70040cf2cf6d6b6edc61`).

[14]For a detailed review, see Weinstein and James (2022).

[15]We sought and incorporated feedback from experts at the Center for Digital Thriving, Harvard University, and the National Institute of Mental Health and Neurosciences, India. Initiators of Change, a Punjab-based youth development organization, worked with us to localize the materials and prepare them for classroom delivery.

The session explains how engagement-ranked feeds can amplify content that is emotional, identity-confirming, novel, frightening, or surprising, even when it is inaccurate. Students then practice evidence-based verification routines, including lateral reading, checking the source before engaging with a claim, looking for corroborating evidence, seeking alternative perspectives, recognizing emotional manipulation, and pausing before sharing. These routines are presented as rules for discernment rather than as a way to memorize a fixed list of false claims.

A central feature of our curriculum is that it explains *how* and *why* digital platforms affect behavior, rather than simply prescribing what students should do. The curriculum describes how features such as variable rewards, infinite scroll, and algorithmic content ranking shape attention and beliefs. For example, rather than telling students that "social media is addictive" or exhorting them to "reduce their usage," the curriculum explains how removing natural stopping points, such as in infinite scroll, makes continued use the path of least resistance. Similarly, rather than advising students to "fact-check more," the curriculum explains how engagement-based algorithms can amplify emotionally resonant content regardless of accuracy, creating systematic exposure to misinformation.

This mechanism-to-practice structure is motivated by evidence that causal explanations improve learning and generalization: when individuals understand why a phenomenon occurs, they can apply the same logic in new settings (Ahn and Kalish, 2000; Keil, 2006; Lombrozo, 2012; Sankar et al., 2025). This feature is important for digital behavior because platforms change quickly, and app-specific advice may depreciate. A student who understands why autoplay keeps her watching videos can recognize the same structure when another platform introduces a different form of automatic continuation. Knowledge of these mechanisms transfers in ways that app-specific advice may not.

3.3 Program Delivery

The curriculum was delivered through the institutions' existing academic structure. At the highest level, partner institutions are organized into departments, which encompass majors or degree programs. Within each major, students are organized by academic year. In some major-year cohorts, students are further divided into smaller sections that follow the same curriculum but may have different instructors. The unit of delivery, what we refer to as a *batch*, is the institution-major-year if a cohort year is not further divided, and the institution-major-year-section if it is (Appendix Figure A1).

Institutions joined the study on a rolling basis, and we scheduled surveys and sessions around their availability and schedules. Sessions were held in person in classrooms, though some were held in larger halls or auditoria when multiple batches from the same treatment arm had to be scheduled together. When regular class-hour slots were unavailable, sessions were delivered remotely instead. The same slide decks, scripts, worksheets, and polls were used across delivery modes.

Sessions were conducted in a mix of English and Punjabi, the preferred communication mode in universities in Punjab, and were held during regular class slots. Each session followed a common structure: facilitators presented the core material, led structured classroom exercises, and used

worksheets to help students relate the content to their own experiences. Sessions also included web-based polls, through which students answered questions anonymously on their phones and aggregate responses were displayed in real time.

Facilitators were selected in collaboration with faculty at one of the partner institutions. They were chosen for public-speaking ability and experience with young audiences, and included speakers with backgrounds in journalism, teaching, personality development, neuroscience, and counseling. The research team trained facilitators on the curriculum and familiarized them with the terminology used in the materials. Each facilitator received a slide-by-slide script for every session to standardize the concepts, sequence, and activities, while allowing some flexibility to use examples that clarified the material without changing the content.

Field monitors observed sessions using a structured rubric that recorded facilitator delivery and student engagement. After each session, students completed a short feedback form on the relevance of the content, interaction and engagement, in-class exercises, and the facilitator.

4 Experimental Design and Data

4.1 Randomization and Stratification

We evaluate whether digital self-command can be taught using a cluster-randomized trial with college students across five higher-education institutions in Punjab, India. The unit of randomization is a batch, defined in Section 3.3 as either an institution-major-year or, where enrollment is large enough to require sections, an institution-major-year-section.[16] We stratify by institution-major and randomly assign batches to one of three arms: control, T1 *Deliberate Use*, or T2 *Deliberate Use & Discernment*. The experimental design is illustrated in Figure 1.

Students in the control group complete the same surveys and check-ins as treated students, in the same settings, but receive no classroom sessions. Students assigned to T1 receive the *Deliberate Use* sessions, covering Sessions 1 and 2. Students assigned to T2 receive *Deliberate Use & Discernment*, which adds Session 3 on evaluating online information to the same two sessions.[17] This design lets us estimate the impact of assignment to the classroom intervention and also compare the common *Deliberate Use* sessions with the additional *Discernment* module.

We stratify random assignment at the institution-major level.[18] Within each institution-major, we assign an equal number of batches to treatment and control. When there is an odd number of batches within a major, we randomly assign the singleton unit to treatment or control. Within

[16] In institutions with a single section for each academic year within a program, the major-year combination is a batch (e.g., BA Economics Year I and BA Economics Year II represent two separate batches). In institutions with multiple sections, the batch is the major-year-section (e.g., BA Economics Year I Section A, BA Economics Year I Section B, and BSc Physics Year III are three distinct batches).

[17] We did not include a standalone *Discernment* arm. This preserved statistical power for our primary outcomes, which were *ex ante* harder to detect than discernment of false claims. The marginal effect of *Discernment*, however, can be estimated by comparing the two treated arms under a plausible additivity assumption.

[18] For example, University X BA Economics, University X BSc Physics, and University Y BA Economics represent three different strata.

the treatment arm, batches are assigned to one of the two curriculum arms with equal probability. Given that students with different interests and motivations might self-select into different majors, stratification by major ensures balance across treatment and control on these unobservables.[19]

4.2 Data Sources

Our analysis draws on three data sources: survey data, objective usage data collected through a custom Android app, and administrative records from partner institutions. Surveys measure a range of outcomes, including digital literacy, locus of control, commitment device use, self-reported digital behavior, digital addiction, misinformation discernment, sleep, wellbeing, social comparison, and valuation of digital access. The objective-use data measure screen time at the app level for Android users. Administrative records provide class schedules and academic performance.

Surveys. We collected survey data at baseline, during three short check-ins between baseline and endline, at endline, and at follow-up. Baseline data collection began in February 2024, with subsequent waves scheduled around the academic calendars of participating institutions. At baseline, all students enrolled in participating batches were invited to complete a self-administered survey in a monitored classroom setting. Consenting students entered the study sample. The baseline survey collected standard demographics and a range of measures including digital access, smartphone and social media use, demand for commitment, digital addiction, mental health and wellbeing, misinformation discernment, and valuation of social media and smartphones. The survey instrument was available in English, Punjabi, and Hindi.

The endline repeated the main baseline outcomes and added modules such as curriculum knowledge and digital literacy.[20] Students received an INR 100 gift card (approximately USD 1.05) for the baseline survey, INR 250 (USD 2.63) for completing two or more surveys by the endline, and one student per batch was selected by lottery for an additional INR 1,000 prize (USD 10.53). Students also received certificates of participation.

The three check-in surveys measure higher-frequency changes during the intervention period. They included short modules on perceived digital addiction, mental health, and misinformation discernment.[21] Control group students received check-ins on the same schedule as treated students,

[19]Two batches at University A deviated from their original random assignment due to administrative error: one treatment-assigned batch was not scheduled to receive the classroom curriculum, while one control-assigned batch was. Because the error originated in scheduling rather than in any characteristic of the batches or choices by their students, the implemented assignment remains independent of potential outcomes, and we analyze batches by the assignment they received. To retain both batches, we collapse the two affected strata, which share the same university and degree family and differ only by specialization, into a single stratum. Coding by original assignment would leave each affected stratum with a single treatment status, causing original-stratum fixed effects to absorb the relevant treated-control variation entirely. We show in the robustness checks that results are unchanged when these batches are coded by their original randomized assignment and when the affected strata are excluded entirely.

[20]Misinformation items were administered at baseline only in later rollout waves and were administered again at endline.

[21]In treated batches, students completed the check-in survey after that day's session. At check-in 3, students in the *Deliberate Use & Discernment* arm answered the first part of the survey, received Session 3, and then completed the rest of the survey for a measure of discernment immediately after the session.

so that repeated contact with the research team was similar across study arms. We conducted the in-person follow-up survey in Fall 2025 and Winter 2026. It repeated a subset of endline modules to measure persistence and included additional modules on transfer to new digital domains.[22]

Figure 1 summarizes the timing between survey waves and intervention sessions. Check-in 2 occurred an average of 2.81 weeks after check-in 1, check-in 3 followed 2.32 weeks later, and endline followed check-in 3 by 1.69 weeks. Follow-up occurred an average of 43.96 weeks after endline.

Objective smartphone usage data. We collected objective usage data through Dashboard, an Android application built for this study. Android users in both the treatment and control groups were encouraged to install the application. In the user interface, the application displays daily and weekly summaries of screen time by application and includes limit-setting features.

Participants did not need to keep Dashboard installed continuously throughout the study for us to observe data between surveys. Whenever participants initiated a data export at baseline, check-in, endline, or follow-up, Dashboard retrieved the usage-event history available from the device system and transmitted it to our servers. Retention varies across manufacturers and Android versions, so the main analysis uses up to the three most recent valid days before each survey when data was exported.[23] These records contain timestamped usage events that identify when applications enter and exit the foreground.[24] We use these events to construct measures of total phone use and use at the category and application level, along with pickups and session length. These data also allow us to construct use around reported sleep times and scheduled class periods.

Administrative records. We collected data on class schedules from partner universities and combine these records with objective usage data to measure phone and social media use during class hours. We also obtained examination records to study effects on academic performance.

4.3 Threats and Design Responses

We took several steps to limit threats to internal validity from attrition, limited availability of objective usage data, spillovers, and experimenter demand. Two design features are central to interpreting the results. First, we randomized at the batch level. As courses are organized by major and year, treated and control students largely do not share classes. Baseline friendship nominations also let us measure each student's exposure to treated friends in other batches and test for spillovers directly. Second, several of our key outcomes do not rely on self-reports. We collect objective usage data, administrative records on class schedules and academic performance, task-based measures of discernment, and an incentivized valuation of Instagram access elicited with a multiple price list. Appendix C.1 describes the design choices and field protocols in detail.

[22] It also included an embedded survey experiment administered during the survey session discussed in more detail in Section 7.

[23] Among observed user-periods, 82.5 percent contain all three days. This falls to 78.3 percent when we require four days and 63.2 percent when we require seven days.

[24] Dashboard does not capture content viewed within applications, text entered by the user, or other personal information.

4.4 Sample Construction

The study sample consists of students enrolled in five higher-education institutions in Punjab, India.[25] Within participating batches, all enrolled students were invited to complete the baseline survey. Our main analysis sample consists of students first observed at baseline or check-in 1, and excludes students who entered in later waves.

The main analysis sample consists of 3,430 students in 86 randomized batches. Appendix Figure A2 shows how the sample is constructed, and Appendix Figure A3 shows the subsample with objective usage data.[26] Among students in the analysis sample, 2,523 are observed in the endline survey, and 2,441 have non-missing data for the primary endline outcomes. For the follow-up, the analysis sample contains 1,836 students in 86 batches.

4.5 Balance, Attrition, and Take-up

Pre-treatment balance. Table 1 shows that treatment and control students are similar in pre-treatment characteristics and baseline values of the outcome variables. Although a few individual differences are statistically significant, their magnitudes are small. Joint tests of significance do not reject balance for either the pooled treatment comparison ($p = 0.551$) or the discernment-arm comparison across all displayed covariates ($p = 0.482$).[27]

Attrition. Table 2 shows no evidence of differential attrition by treatment assignment at endline. Among students who entered before the intervention, 74.4 percent of control students participated in the endline survey, an attrition rate of 25.6 percent that is typical for this setting.[28] Students in the treatment group are 0.9 percentage points less likely to participate ($p = 0.695$). The same holds when we restrict to baseline attendees or to students who completed at least two interim surveys.

Appendix Table A1 shows that treatment assignment also does not predict participation in the check-in surveys or in the later follow-up survey. At follow-up, 59.4 percent of control students participated, and the difference between treatment and control is -1.7 percentage points ($p = 0.496$).

Observability of smartphone usage. Usage records from Dashboard are available only when students complete a data export. The main screen-time specification is a difference-in-differences with student fixed effects that uses the pre-intervention window before check-in 1 as the omitted

[25] Most students are enrolled in undergraduate programs. In two institutions, the sample also includes master's students.

[26] Across all surveys, we observe 5,385 students in 112 groups. We drop students in 7 pilot batches. In one institution-major, we learned after randomization that group labels reflected English-instructor assignments rather than distinct sections, and we were not able to deliver the program to these groups as intended. We therefore drop 682 students in 19 groups in this institution-major, and show that results are robust to their inclusion. Finally, we drop students who were not observed by check-in 1, were younger than 18 at the time of the survey, or could not be linked to a randomized batch.

[27] We compute the joint-test p-values using randomization inference, following Kerwin, Rostom, and Sterck (2024), who show that conventional omnibus balance tests over-reject the null.

[28] For comparison, Bandiera, Jalal, and Roussille (2025) retain 69 percent of graduating university students in Pakistan at six months and 61 percent at nine months.

period. In the control group, 40.8 percent of students with an Android device contribute data in this window, while students in the treatment group are 8.4 percentage points more likely to do so ($p = 0.012$, Appendix Table A2). This difference is already present before curriculum delivery and remains stable thereafter. Relative to the pre-intervention period, it changes by -2.0 percentage points at endline ($p = 0.584$).

The higher export rate at check-in 1 in the treatment group comes from students exporting for the first time. Treatment and control students who exported at baseline are similarly likely to export again (1.8 percentage points, $p = 0.766$), but among students who had not, students in the treatment group are 14.6 percentage points more likely to export at check-in 1 ($p = 0.003$).[29]

We find no evidence of change in the gap between treatment and control in objective use before the intervention, consistent with the parallel-trends assumption of our difference-in-differences specification. Appendix Table A2 shows that the gap between treatment and control in total phone use is around four percent both at baseline ($p = 0.172$) and in the pre-intervention window ($p = 0.294$).

Intervention take-up and implementation. Sessions were delivered in group settings, with a median size of 23 students. Most sessions were delivered in person, with 22 percent delivered remotely when regular class-hour slots were unavailable. The same curriculum materials and facilitation protocol were used across delivery modes.

We measure intervention take-up using attendance records collected by field staff at the beginning of each session. Among students assigned to either treatment group, 81.66 percent attended at least one intervention session. 30.47 percent attended all sessions to which they were assigned: two sessions in the *Deliberate Use* arm and three sessions in the *Deliberate Use & Discernment* arm. Attendance rates were 62.00 percent for the first self-command session, 50.08 percent for the second self-command session, and 54.43 percent for the misinformation session.[30]

Classroom observations indicate that implementation quality was high. Facilitators used relevant examples, maintained an appropriate pace, and left time for student participation. Most students completed the worksheets, and classroom distractions were limited. Worksheet responses suggest that students engaged in class. In the Session 1 reflection worksheet, students named a frequently used social media application and described its benefits and costs. In Session 2, students stated a broad goal for the year and wrote concrete steps toward it. 56 percent of planning worksheets mention academic or skill goals, while 59 percent of planned steps mention self-discipline and habits and 57 percent mention academic or skill steps. Nearly all students who completed the worksheet wrote a goal and at least one concrete step. The most commonly named screen-time tools were app limits and timers, notification control, offline substitutes, and self-discipline (Appendix Figure A4).

[29]This could reflect longer scheduled check-in sessions for the treatment group that made it easier for students who are unfamiliar or uncomfortable with the process to complete the export.

[30]Students who missed a scheduled session were offered a make-up session, either in person or online. When scheduling constraints made separate make-ups infeasible, material from multiple missed sessions was covered in a single condensed session. Approximately 2.49 percent of treated students attended a make-up session.

4.6 Baseline Digital Environment

Smartphone and social media use is near-universal in the sample. Among baseline respondents, 98.8 percent own a smartphone and 99.8 percent report using social media. Students in the sample are young adults: the mean age is 19.3, and 50.4 percent are female. Around 52.4 percent have at least one parent who attended college, and 73.4 percent of students come from urban areas. 55.9 percent report having Wi-Fi at home, and 71.4 percent acquired their first mobile phone before age 18. Smartphones and social media are therefore already integrated into daily life for most students in the sample.

Objectively measured phone use at baseline is high and driven largely by social media use (Appendix Figure A5). The mean total phone use is 5.7 hours per day and the median is 5.6 hours per day. Social media accounts for most of this use, with a mean of 3.8 hours per day and a median of 3.6 hours per day. The distribution is right-skewed, with a 95th percentile of 9.4 hours per day, and the treatment and control distributions overlap closely at baseline (Appendix Figure A6).

The baseline survey provides self-reported social media use for the entire analysis sample. Students report using social media for approximately 4.4 hours per day on average and 4.0 hours per day at the median. Among students with both self-reports and objective use data, self-reported social media use averages 4.17 hours per day, while objective social media use averages 3.62 hours per day. The two distributions overlap substantially and both have long right tails (Appendix Figure A7).[31]

Some students report using social media more than they would like. 21.7 percent of students say they use social media too much, while 66.5 percent say they use the right amount and 11.7 percent say they use too little (Appendix Figure A8). Even the "right amount" group reports about four hours per day, suggesting that high use is normal in this population.

Most students report at least some interest in setting limits or managing use more intentionally. Only 16.4 percent report no interest in better managing time spent on social media.

5 Estimation Strategy

Survey outcomes. Following our pre-registration report, the main estimates are intent-to-treat effects of assignment to the program.[32] Unless specified otherwise, (i) we report effects on individual outcomes in control-group standard deviation units, and (ii) when a family contains multiple related outcomes, we standardize the components, orient them so that higher values indicate improvement, and aggregate them into an index using inverse-covariance weights following Anderson (2008). We estimate

$$Y_{ijk} = \beta_0 + \beta_1 T_{jk} + X_{ijk}^{c}\gamma + \lambda_k + \varepsilon_{ijk}, \tag{3}$$

[31] For this comparison, we use the seven-day objective-use window to match the survey question's past-week recall period. The results are unchanged using the three-day window.

[32] See Joshi, Kaur, and Song (2024). Any deviations from the pre-registration report are documented in Appendix Section G.

where Y_{ijk} is the outcome for student i in batch j and randomization stratum k. T_{jk} equals one if batch j is assigned to one of the treatment arms and zero if it is assigned to control. λ_k denotes randomization-stratum fixed effects. Standard errors are clustered at the batch level, the unit of randomization. The vector X^c_{ijk} contains gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its batch mean when available. We handle missing covariate values following Zhao and Ding (2024).[33] We also report specifications without baseline outcome controls, with alternative covariate sets, and with controls selected by post-double-selection lasso in robustness checks.[34]

Our parameter of interest is β_1, which captures the intent-to-treat effect of treatment assignment. Because treatment is randomized at the batch level, our main specifications weight each observation by the inverse of its pre-intervention batch size, so that each randomized batch receives approximately equal total weight.[35]

Since the misinformation session is present only in T2, to test the impact on arm-specific outcomes such as ability to discern misinformation, we estimate

$$Y_{ijk} = \beta_0 + \beta_1 T1_{jk} + \beta_2 T2_{jk} + X^c_{ijk}\gamma + \lambda_k + \varepsilon_{ijk}, \tag{4}$$

where $T1_{jk}$ equals one for batches receiving Sessions 1 and 2, and $T2_{jk}$ equals one for those receiving all three sessions.

Academic outcomes. We construct failure rates and grades from partner institutions' administrative records. We estimate intent-to-treat effects by regressing each outcome on treatment assignment with randomization-stratum fixed effects, inverse baseline batch-size weights, the corresponding pre-intervention academic status where available, and standard errors clustered at the batch level.

Screen-time outcomes. Our main specification uses data at the student-day level. For each survey or check-in, the measurement period contains up to the three most recent valid days before the reference date, excluding the survey day itself. A day is valid if it contains a complete record of screen time.[36]

[33] We replace missing covariates with zero, add indicators for missing values, center covariates and indicators at their means, and interact them with treatment, so the treatment coefficient gives the average intent-to-treat effect over that sample. Equation (3) omits these interactions for readability.

[34] As prespecified for the secondary outcomes, we also report sharpened FDR q-values (Appendix Table A3).

[35] Our primary estimand gives equal weight to each randomized classroom batch, reflecting the level at which the program is randomized and delivered. Because batch sizes vary considerably in our sample, an unweighted student-level regression would give larger batches greater influence. With complete outcome data and regressors that are constant within batches, this weighted student-level OLS specification yields the same treatment coefficient as an equal-weighted regression on batch averages (Athey and Imbens, 2017). Outcome-specific missingness and individual-level covariates make this equivalence approximate in our specifications, so we also report collapsed batch-level estimates as a robustness check.

[36] The panel is unbalanced because students do not export at every visit and exports do not always recover every day of the window. Dashboard reads local usage records from the device, and retention varies across manufacturers and Android versions.

With this panel, we estimate a difference-in-differences specification with student fixed effects. Let Y_{ijkmt} denote a screen-time outcome for student i in batch j, randomization stratum k, on recorded day t in measurement period m:

$$Y_{ijkmt} = \alpha_i + \sum_{p \in \mathcal{M}} \beta_p \left(T_{jk} \times \mathbf{1}\{m = p\} \right) + \lambda_{kw} + \delta_m + \mu_r + \phi_{dow} + \varepsilon_{ijkmt}. \tag{5}$$

Here α_i is a student fixed effect and T_{jk} is treatment assignment. The set $\mathcal{M}$ contains the three measurement periods after the program starts: check-in 2, check-in 3, and endline. All β_p are estimated jointly in a single regression, relative to the pre-intervention period. The pre-intervention period uses up to the three most recent valid days before check-in 1, the last measurement window before the first classroom session.[37]

The included fixed effects absorb the main sources of common variation in screen time. Student fixed effects absorb persistent differences in each student's level of use. Randomization-stratum-by-calendar-week fixed effects λ_{kw}, where w is the calendar week containing day t, compare treated and control batches within the same randomization block and calendar week, absorbing shocks such as exams, holidays, and seasonal changes in phone use. Measurement-period fixed effects δ_m, where m indexes the check-ins 1, 2 and 3, or endline, absorb common differences in use across periods.[38] Relative-day fixed effects μ_r, where r records the number of days before the survey or check-in, align the same position within the retrospective measurement period across students and waves. Finally, day-of-week fixed effects ϕ_{dow} absorb systematic differences across days of the week.

Each student-day is weighted by the inverse of the student's baseline batch size and by the inverse of the number of valid days that student contributes.[39] Standard errors are clustered at the batch level.

The coefficients of interest are the β_p, which compare screen-time outcomes in post-intervention measurement period p to the immediate pre-intervention period for treated students, relative to the corresponding change for control students. These wave-specific estimates trace how the effect evolves over the intervention period. We also report estimates from a pooled specification that replaces the summation in equation (5) with a single term $\beta(T_{jk} \times Post_m)$, where $Post_m$ equals one for measurement periods in $\mathcal{M}$, to summarize the average treatment effect after students are exposed to the program.

The main specification uses all valid student records, including those from students without immediate pre-intervention data. The treatment coefficient is identified from residual variation in $T_{jk} \times Post_m$ after absorbing student fixed effects, so data from these students enter mainly through

[37] We do not use the baseline window for two reasons. First, due to scheduling constraints, baseline surveys were often collected well before the intervention, and the time between baseline and check-in 1 varies across batches, with a mean of 18.25 weeks and a standard deviation of 13.79 weeks. Second, as noted in Appendix C.1, we have baseline screen time data for only 59 of the 86 batches in the analysis sample.

[38] The follow-up analysis uses the same specification, with follow-up included as a measurement period.

[39] The first weight matches the survey regressions and gives each randomized batch equal total weight. The second gives each student equal total weight, so students observed on more days do not receive more influence. We also report results under alternative weights in robustness checks.

the estimation of the fixed effects. Daily phone use is nonnegative and right-skewed, with greater dispersion at higher levels of use. We therefore estimate the preferred specification in logs and report effects as percentage changes. We report effects in minutes where absolute magnitudes aid interpretation.

We assess the sensitivity of the survey and objective use estimates to alternative specifications, samples, covariate adjustments, and weighting, and present the results in Appendix D.1.

6 Effects of the Intervention

6.1 Smartphone and Social Media Use

The intervention reduced objectively measured total phone use by 8.9 percent at endline ($p = 0.006$). Using the pre-intervention control mean, this estimate corresponds to 29.4 fewer minutes per day. The specification in minutes gives a similar reduction of 32.0 minutes ($p < 0.001$, Appendix Figure A9). Social media use falls by 10.0 percent ($p = 0.013$), equivalent to 21.7 minutes at the control mean, and by 24.9 minutes ($p < 0.001$) when we use a levels specification.Non-social media phone use also falls, but by a smaller amount, and is not precisely estimated (6.3 percent, $p = 0.358$).

The reduction in phone use appears by check-in 2 and grows through endline (Panel B of Figure 2). Total phone use changes by -3.4 percent at check-in 2 ($p = 0.419$) and -2.8 percent at check-in 3 ($p = 0.346$). Averaging over the three post-curriculum periods, treatment reduces total phone use by 5.2 percent ($p = 0.063$) and social media use by 7.7 percent ($p = 0.041$). Panel A of Figure 2 summarizes the endline effects on total, social media, and non-social media use.

The decline is concentrated in social media, in long sessions, and on weekdays. At endline, social media accounts for the largest component of the decline, and Instagram shows the largest reduction among the apps (Appendix Figure A10). The number of sessions longer than 30 minutes falls while phone pickups rise (Panel A of Figure 3). This suggests that the reduction in use comes from fewer long sessions. Students pick up their phones more often, but very brief sessions of one to two seconds do not increase, so the additional pickups may reflect more deliberate use rather than mindless checking. Treatment reduces total phone use by -10.0 percent from Monday through Friday and -3.9 percent on Saturday and Sunday, though the difference is not statistically significant ($p = 0.144$; Panel B of Figure 3).

6.2 Mental Health

We construct a mental health index that combines digital addiction (Allcott, Gentzkow, and Song, 2022), symptoms of anxiety and depression, and subjective wellbeing. Anxiety and depression are measured with the GAD-7 and PHQ-9 (Spitzer et al., 2006; Kroenke, Spitzer, and Williams, 2001). Subjective wellbeing is measured using the Subjective Happiness Scale (Lyubomirsky and Lepper, 1999), the Satisfaction with Life Scale (Diener et al., 1985), and the three-item UCLA loneliness scale (Hughes et al., 2004), which asks how often students feel left out, isolated, and

lacking companionship. Digital addiction, anxiety, depression, and loneliness are reverse-coded so that higher values indicate better outcomes.

Figure 4 shows an improvement of 0.066 SDs ($p = 0.080$) in the mental health index, largely driven by digital addiction. The program had little impact on symptoms of anxiety and depression, happiness, or life satisfaction. Within subjective wellbeing, only loneliness changes, falling by 0.118 SDs at endline ($p = 0.003$). This suggests that while social media allow people to connect online, students do not feel more lonely with reduced use.

Treatment reduced the digital addiction index by 0.078 SDs at endline ($p = 0.025$). This index captures the perceived loss of control and costs of phone use that the curriculum directly addresses. We adapt the scale in Allcott, Gentzkow, and Song (2022) for the local context and use a battery of items including using the phone longer than intended, unsuccessfully trying to cut back, losing sleep because of late-night use, and allowing phone use to interfere with relationships or other activities.[40] All estimates in the component decomposition of digital addiction are negative (Appendix Figure A11). The largest reductions are in salience, which captures worry about missing online activity and checking the phone immediately after waking, and mood modification, which captures using the phone to distract from problems or negative feelings. Both conflict measures also decline, with treatment students reporting less loss of control and fewer functional and interpersonal costs from phone use. The estimates for withdrawal and relapse are negative but imprecise.

To track changes during the intervention, we measure digital addiction and symptoms of anxiety and depression at the check-ins. Panel B of Figure 4 traces treatment effects over the course of the study. For the addiction index, treatment effects are close to zero at baseline and check-in 1. At check-in 2, after Sessions 1 and 2, treated students score 0.134 SDs higher than control students ($p = 0.002$). The effect attenuates at check-in 3 and reverses by endline, when treated students score 0.115 SDs lower ($p = 0.002$).[41] For anxiety and depression, there is little change over time.[42]

The short-run increase in the addiction index occurs while objective phone use is already falling. We interpret this divergence as an awareness effect. The first two sessions made students more aware of problematic use they had previously normalized, temporarily raising reported addiction even as use declined. By endline, continued reductions in use dominated this initial change in self-assessment, leaving treatment students with lower digital addiction scores.

[40]The scale is based on the components model of addiction (Griffiths, 2005), which specifies salience, tolerance, mood modification, relapse, withdrawal, and conflict. Our scale does not include tolerance, which Charlton and Danforth (2007) treat as a marker of high engagement rather than addiction. Conflict includes items on costs to work and relationships as well as items on subjective loss of control, which Griffiths (2013) subsumes within conflict.

[41]The check-ins use a shorter battery than the baseline and endline to keep the surveys brief. Check-in 2 completion rates do not differ by treatment, and baseline covariates are jointly balanced among respondents ($p = 0.198$). Furthermore, among the same 1,291 students observed at both check-in 2 and endline, the treatment effect is 0.111 SDs higher at check-in 2 ($p = 0.014$) and 0.128 SDs lower at endline ($p = 0.013$).

[42]One channel through which social media may affect mental health is social comparison, as students are regularly exposed to other people's lives through curated content on social media. Using a five-item scale based on Tandoc Jr, Ferrucci, and Duffy (2015), we find no statistically significant effect on social comparison (-0.039 SDs, $p = 0.222$; Panel C of Figure 5), matching the null effects on anxiety and depression.

6.3 Discernment

We measure discernment by asking students to rate the accuracy of claims on a 5-point scale. Each student saw a randomly selected subset of items drawn from a larger item bank, allowing us to cover a wider range of content that people usually see online. Claims appeared as screenshots of social media posts, so students judged each claim in the format they encounter it online (Appendix Figure A12). Because students answered overlapping but different subsets, raw averages would treat students who saw easier and harder items alike. We therefore construct the discernment index using item response theory (IRT), which places students on a common scale that accounts for differences across items.[43] Ratings of true claims are scored in their original direction and ratings of false claims are reverse-coded, so higher values indicate greater accuracy and discernment. Effects are reported in endline control standard deviations.

The *Deliberate Use & Discernment* arm, which received Session 3, increases discernment by 0.183 SDs relative to control ($p < 0.001$, Figure 6). The gain comes from better rejection of false claims, which rises by 0.237 SDs ($p < 0.001$). The effect on recognition of true claims is negative (-0.100 SDs; $p = 0.069$), suggesting that the treatment also led to more caution overall.

The negative estimate for true-claim recognition raises the question of whether the higher discernment index reflects better separation of true from false claims or simply a general tendency to rate every claim as less accurate. The latter would improve performance on false claims but worsen performance on true claims, with no change in the gap between them. To distinguish the two, we estimate separately the average shift in ratings across all claims, which we label generalized caution, and the change in the true–false rating gap. Students in the *Deliberate Use & Discernment* arm rate claims 3.9 percentage points lower on average, indicating more skepticism or caution for accepting online information as accurate. The true–false gap also widens by 3.2 percentage points ($p = 0.028$).

The increase in false-claim rejection also extends beyond claims students may previously have encountered. Treatment increased rejection of novel false claims, designed for this study by the research team and never in circulation, by 0.169 SDs ($p = 0.001$), and we cannot distinguish this effect from the effect on circulating false claims ($p = 0.223$). Because these items had not circulated, the improvement does not rest on recalling particular claims or corrections encountered elsewhere.

Discernment only improves with targeted instruction on misinformation. In principle, Sessions 1 and 2 could have improved discernment indirectly by changing the information students encountered online or how they engaged with it. The estimates for the *Deliberate Use* arm, which received the first two sessions but not Session 3, are smaller and not statistically distinguishable from zero. The two treatment arms have similar effects on the digital-use outcomes targeted in Sessions 1 and 2 (Appendix Table A16). The estimated endline discernment effect is 0.109 SDs larger in the arm that additionally received Session 3 ($p = 0.106$). The timing of the effects provides a complementary check. The arm that receives Session 3 does not show any improvement in discernment before that

[43] See Lord (1980) and Embretson and Reise (2000) for an overview of item response theory. We summarize the construction and estimation details in Appendix C.2.

session, but scores higher immediately afterward, at check-in 3 (Appendix Figure A13).[44]

6.4 Sleep and Academic Outcomes

Sleep The time students free up from screens goes mainly to sleep. Students were asked to allocate 24 hours to describe how much time they spend in a typical day across screen time unrelated to studying, study time, offline activities, and sleep. Panel A of Figure 5 shows an increase in the time assigned to sleep of 17.647 minutes ($p = 0.005$). Study time and offline activities do not change.[45]

Panel B of Figure 5 shows that the effect on the sleep index is positive and imprecise (0.051 SDs, $p = 0.178$). We find no detectable change in objective phone use in the hour before bedtime and the hour after waking, using sleep and wake-up times reported at baseline (Appendix Table A18).

Attention during class We construct a novel proxy for in-class focus and attention from students' phone behavior during scheduled class time by combining objective smartphone usage data with course timetables from partner institutions.[46] Because scheduled class time varies across batches and days, we normalize phone use by the available hours in each class period and express it per class hour. We construct two measures: time spent on the phone during class, which captures the duration of distraction, and phone pickups, which capture its frequency.

Treatment changes the pattern of phone use across class and break periods (Appendix Table A19). At endline, the estimated reduction in social media use is 0.953 minutes per scheduled class hour ($p = 0.160, q = 0.154$), with a smaller estimated reduction during breaks. In an exploratory analysis, we find that the reduction in social media use relative to other app use is 2.2 minutes per hour greater during class than during breaks ($p = 0.038$). This pattern is consistent with students becoming more selective about phone use when it competes with instruction, rather than simply reducing all use during class.

Treatment increases the share of class days with no observed phone pickup during class by 1.3 percentage points, relative to a control mean of 1.4 percent ($p = 0.037$). The estimated shifts between days with at most five and more than five pickups per class hour are small and imprecise (Appendix Table A20).[47]

Academic performance. We find that assignment to treatment reduced the probability that a student failed or had to reappear for end-of-term exams by 5.0 percentage points ($p = 0.033, q =$

[44]The difference between the two treatment arms is -0.036 SDs at check-in 1 ($p = 0.744$) and -0.085 SDs at check-in 2 ($p = 0.311$). Immediately after Session 3, the difference is 0.300 SDs ($p < 0.001$).

[45]The absence of any change in study and offline time also provides evidence against students shifting self-reports toward socially desirable activities.

[46]Phone use during class competes directly with attention to instruction, so time spent on the phone, particularly on social media, is a natural behavioral proxy for divided attention. Unlike self-reports or classroom observation, objective usage data merged with course timetables measure this unobtrusively and at scale, without relying on monitors or students' recall. This class-time attention proxy complements the task-based attention measures we examine at follow-up (Section 7.2).

[47]Our measure matches objective phone use with scheduled class periods and is not conditioned on actual attendance. To the extent that some scheduled periods correspond to student absences, it captures phone behavior during scheduled class time rather than while physically seated in class.

0.086), relative to a control mean of 37.9 percent (Panel B of Figure 5).[48] This outcome comes from administrative records and is therefore not vulnerable to experimenter demand effects. We do not find a corresponding impact on student grades (0.003 SDs, $p = 0.951$; Appendix Table A12). Appendix Table A21 reports robustness checks.

The results suggest that the intervention was effective at helping students cross the threshold required to pass their exams. Some potential explanations are that the program helped students set more specific academic goals or manage their phone use more effectively while studying, or that it increased the amount of sleep, which has been shown to improve academic performance (Giuntella, Saccardo, and Sadoff, 2024). During Session 2, students stated their goals for the year and identified practical strategies for managing phone use while studying. Academic and skill-related goals were the most frequently chosen (56 percent), and treatment students were more likely to report having set a specific academic target for the semester at endline (Appendix Table A22).

6.5 Internal Validity

Experimenter demand and social desirability effects. A natural concern is that students in treatment would infer the goals of the program, by virtue of being treated, and report what they think the research team wants to hear. Such reporting cannot explain the effects on phone use and academic performance, which we measure from device and administrative records rather than self-reports. Appendix C.1 describes additional design features that limit this concern. Other main outcomes are also measured through tasks and standardized scales rather than students' descriptions of their behavior. Moreover, both treatment and control groups completed the same survey waves and check-ins in the same settings and on the same schedule, so survey contact with the research team was common to both groups.

For self-reported outcomes, two checks suggest that favorable reporting does not explain the main results. First, the trajectory of digital addiction is difficult to reconcile with students simply reporting better phone habits. Treated students initially report greater difficulty regulating their phone use, before reporting lower addiction at endline. If students were consistently reporting what they thought the research team wanted to hear, we would expect lower reported addiction throughout instead of the trajectory we see.

Second, we test whether treatment effects are larger among students more prone to providing socially desirable responses. At baseline we administered the five-item scale of Hays, Hayashi, and Stewart (1989) to a subset of 43 batches in strata where both treated and control students received the module. The treatment interactions with baseline social desirability are small and imprecise for the main outcomes, providing no systematic evidence that socially desirable responding accounts for the survey effects (Appendix Table A23).

[48] The high control mean is not unusual in this setting. Universities in Punjab routinely classify students who do not clear an examination as "reappear," and retaking one or more exams is common. Our study also coincided with the transition to the National Education Policy (NEP), during which several participating universities reported lower pass rates (Attri, 2025).

Spillovers and interference. One potential concern is that students in the treatment and control groups may interact, leading to SUTVA (Stable Unit Treatment Value Assumption) violations. Three features of the setting limit this concern. First, the intervention involved facilitated sessions, worksheets about students' own digital habits, classroom discussion, polls, and repeated practice. Casual conversations with treated friends are unlikely to reproduce the full curriculum. Second, treatment was assigned at the batch level, and batches correspond to institution-major-year cells. Since courses are organized largely by year, treated and control students usually did not share common classes. Third, we link baseline friendship nominations to the nominated students' survey records using names and other identifying information. Most friendships stay within batches: 80.1 percent of nominations are within the student's own batch, and 37.0 percent of students have any friend in another batch.

Nevertheless, spillovers may still occur through two channels. The first is information sharing. Treated students could repeat concepts or describe tools to untreated friends. If this channel were important, control students with more treated friends should show greater knowledge or understanding of the curriculum. Among control students, we find no detectable effect of exposure to treated friends in other batches on content knowledge or digital literacy, providing little evidence of curriculum transmission across treatment assignments (Appendix Table A24).

A second channel is behavioral spillover. Because phones and social media facilitate social interaction, treated students could affect untreated friends by changing their availability, responsiveness, or expectations around phone use. Among control students, we find no detectable effects of exposure to treated friends on digital addiction, interest in setting limits, sleep, or wellbeing. The estimated effects on internal locus of control and commitment device use are positive and marginally significant. Thus, we find no broad pattern of behavioral spillovers, although modest effects on perceived agency or tool adoption cannot be ruled out. To the extent that such spillovers move control students in the same direction as treated students, they would attenuate rather than generate the estimated treatment effects.

Selection into smartphone usage measurement. The difference in usage data availability documented in Section 4.5 could bias the difference-in-differences estimate if, absent treatment, the additional treatment group students entering measurement would have followed different screen time trends from the students observed in the control group. Student fixed effects absorb stable differences in levels, so only differences in these counterfactual trends threaten identification.

To check whether differential selection into the objective-use sample drives the results, we reweight students by their predicted probability of appearing in that sample. This leaves the endline estimates essentially unchanged: -8.1 percent for total phone use and -8.9 percent for social media use, against -8.9 and -10.0 percent in the main specification (Appendix Table A25).

Changes in the composition of students contributing data do not appear to explain the endline reductions. We restrict the analysis to the same 632 students observed at check-in 1 and endline, so the estimates cannot be driven by entry into or exit from measurement. Total phone use falls

by 21.4 minutes per day ($p = 0.032$) and social media use by 12.7 minutes ($p = 0.062$). Separating students already observed at baseline from those first contributing at check-in 1 gives similar, less precisely estimated reductions across the two groups (Appendix Table A25).[49]

Together, these checks suggest that the endline reductions are not driven by selection into measurement. At endline, a joint test finds no evidence that measurement availability sorts differently by treatment status across baseline objective use, digital addiction, digital literacy, and interest in setting limits ($p = 0.531$, Appendix Table A2). These exercises cannot rule out selection on unobserved characteristics. Such selection would, however, need to predict different untreated changes in phone use within small randomization strata while surviving the support restrictions and reweighting exercises above. We therefore treat differential measurement availability as a limitation, but the available diagnostics provide little evidence that it explains the estimated reduction in phone use.[50]

7 Mechanisms

The conceptual framework in Section 2 suggests that education can improve digital self-command by changing beliefs and encouraging the adoption of strategies that help students act on those beliefs. We test these channels to shed light on how the program produced the effects in Section 6. We then consider whether digital self-command is a capacity that persists and transfers to domains not covered in the curriculum. To assess whether the program can be delivered to full classrooms, we examine effects across subgroups. Finally, we consider the role of classroom delivery, how information provision and peers matter, and the cost-effectiveness of the program.

7.1 Channels for Digital Self-Command

If education builds self-command, students should become better at recognizing how digital environments influence their choices and identifying ways to respond. We find evidence of both. Figure 7 shows that the program increased students' scores by 0.201 SD on a quiz covering the core concepts taught in the sessions, including platform design, habit formation, commitment devices, and misinformation ($p < 0.001$). This suggests that, weeks after the sessions, students retained the concepts that were covered in the curriculum. Moreover, students learned to apply these concepts to unfamiliar situations. Performance on a vignette-based measure of digital literacy, adapted from the Youth Social Media Literacy Inventory (Purington Drake et al., 2023), increased by 0.099 SD ($p = 0.029$). Students were presented with scenarios such as a peer who feels worse after comparing

[49] We report this diagnostic in minutes because minutes aggregate linearly across students, so changes in use within a fixed group can be separated from changes in sample. Among baseline contributors, the reductions are 22.6 minutes for total use and 9.9 minutes for social media, and among students first contributing at check-in 1, they are 25.8 and 24.1 minutes. Tests of equal effects give $p = 0.881$ and $p = 0.404$.

[50] A sensitivity calculation asks whether selection could explain the reduction among students observed at both surveys. For selection alone to explain the reduction in use, the additional students contributing data under treatment would need to have reduced phone use, without the program, by 275.9 minutes per day more than students whose data would be available regardless of treatment (Appendix Table A25).

herself with others online, or another who keeps checking how many likes a post has received, and asked to recommend a response. Because the sessions did not present these specific scenarios, the result suggests that treated students were better able to diagnose the underlying mechanism and identify an appropriate response.

Panel B of Figure 4 shows that the program initially increases the addiction index before lowering it by endline. This pattern suggests that the program first raises awareness before it changes behavior. In the framework, these results are consistent with more accurate beliefs, as the program brings the student's forecast of temptation, $\widetilde{\gamma}(m, s)$, and her perceived future gain from stopping, $\widetilde{V}(s; \widetilde{\gamma})$, closer to their true values, $\gamma(m, s)$ and $V(s)$.

More accurate beliefs about temptation and future payoffs may have little effect if students lack a sense of agency over what happens in their lives. A student who believes that her own actions cannot change outcomes will perceive little net gain from an application limit, rule, or routine. We examine whether the program strengthened students' sense of agency. We measure this using an index of locus of control following Pearlin and Schooler (1978), which captures how much control students felt they had over their lives during the preceding three weeks. Panel B of Figure 7 shows that treatment increased internal locus of control by 0.081 SD ($p = 0.057$). This is consistent with the emphasis in the program that students' own choices and strategies can shape how digital technologies affect their lives.

Adoption is more likely when the student knows about a strategy, anticipates the states in which it will be useful, and believes that it can change her behavior or outcomes. We next examine whether these beliefs translate into the adoption of strategies. Panel C of Figure 7 shows that students in the treatment group were 3.9 percentage points more likely to report strong interest in limiting or better managing their time on social media ($p = 0.032$, control mean 36.6 percent). They were also 4.6 percentage points more likely to report using a commitment device ($p = 0.016$, control mean 59.7 percent). Because treatment and control students had access to the same built-in phone tools, this difference reflects greater take-up of existing tools rather than differential access.

Importantly, the curriculum does not imply that all smartphone and social media use is harmful, and instead targets how, not whether, students use these technologies. We ask whether the program changes how much students value their access. In an incentivized multiple price list, students reported the payment they would require to deactivate Instagram for two weeks.[51] While Instagram accounts for the largest reduction in use among apps, we do not find evidence that treatment reduces the value students place on Instagram access (Appendix Table A13). The change in valuation is INR 50 ($p = 0.620$), with a control mean of INR 2,632.

We find evidence that, while students still value access, they are more selective about when they use it. At endline, we asked students for the lowest monthly salary they would accept for a standard 9-to-5 job and for the same job if they had to leave their phone in a locker during

[51] Students chose whether they would deactivate Instagram for INR 500, 1,000, 2,000, 3,000, 4,000, or 5,000. Students who rejected all offers entered the lowest amount they would accept, or indicated that they would not deactivate at any price. One student per institution was randomly selected for implementation. For the analysis, reservation prices above the largest listed offer are top-coded at INR 5,000.

working hours, with access only during the lunch break. We define the compensation premium in this hypothetical scenario as the reservation wage for the restricted job minus that for the standard job, coding negative values as zero. The program reduces the compensation premium for giving up phone access by 14.4 percent of the control mean ($p = 0.015$, Appendix Table A13). This suggests the program helps students recognize that the same technology may be valuable in one state and conflict with their goals in another, such as during work hours. Objectively measured phone use shows a similar pattern in the school setting. As discussed in Section 6.4, students in the treatment group shift social media use away from scheduled class time toward breaks and are more likely to have class days with no phone pickups at all. Together, these results suggest that students value and use their smartphones less when phone use conflicts with a valuable alternative.

7.2 Persistence and Transfer

We next ask whether the program builds a capacity that persists over time and extends to technologies the curriculum did not cover. We test both using a follow-up survey conducted months after the program.

Persistence. We report effects at follow-up in Figure 8. Students in the treatment group retain greater digital literacy and perceived control over their phone use, and report better sleep and lower digital addiction.[52] They also score 0.187 standard deviations ($p < 0.001$) higher on digital literacy and 0.079 standard deviations ($p = 0.063$) higher on internal locus of control.[53] They report better sleep, with a 0.114 standard deviation ($p = 0.021$) improvement in the sleep index. Digital addiction remains 0.077 standard deviations lower ($p = 0.021$) compared to the control group. The estimates for digital literacy, sleep, and digital addiction are not statistically different from those at endline. In contrast, we detect no follow-up effects on interest in setting limits or commitment device use.[54]

The estimates for objectively measured total and social media use are negative but too imprecise to distinguish from either zero or the endline estimates. Total phone use changes by -1.4 percent ($p = 0.772$) and social media use by -4.0 percent ($p = 0.707$). The non-social media estimate is positive and imprecise (11.4 percent; $p = 0.261$).

Treatment students, however, continue to report reduced phone use during class hours. They are 4.6 percentage points more likely to report never checking their phone in class ($p = 0.024$), relative to a control mean of 17.0 percent, and 6.2 percentage points more likely to report checking

[52] Follow-up was conducted a median of 11.1 months after Session 2, ranging from 7.5 to 15.6 months across batches. The pre-analysis plan anticipated follow-up approximately three months after endline. In implementation, the median interval after endline was 9.2 months. All 86 batches were surveyed at least three months after endline, and 85 at least six months after.

[53] At follow-up, we measure locus of control using the Internet Locus of Control Scale (Duradoni et al., 2023), which adapts the construct to online behavior.

[54] One interpretation is that the capacity built by the program may outlast continued use of the specific tools it taught.

five times or fewer rather than more often ($p = 0.002$).[55]

A task-based measure of attention corroborates the self-reported reduction in class-time phone use. We use three attention checks embedded at different points in the follow-up survey to construct a task-based measure of sustained attention. Unlike self-reported phone checking, these outcomes record whether students successfully read and followed instructions as the survey progressed.[56] On the first check, treatment raises the pass rate by 2.2 percentage points relative to a control mean of 75.1 percent, although the estimate is imprecise ($p = 0.309$). On the middle and final checks, treated students are 8.5 ($p < 0.001$) and 6.3 ($p = 0.006$) percentage points more likely to pass (Appendix Figure A14). Because these checks measure observed task performance rather than students' descriptions of their attention, treated students' better performance suggests they can focus better even on a digital device, potentially consistent with improved cognitive endurance (Brown et al., 2025).

Together, these results suggest that effects of the program on deliberate use persist almost a year later. In contrast, we do not detect an improvement in discernment at follow-up. Figure 8 shows that the effect of *Deliberate Use & Discernment* on discernment is close to zero approximately 10.5 months after Session 3 (-0.065 SD; $p = 0.296$). Neither true-claim recognition nor false-claim rejection improves at follow-up.

Transfer. We examine students' use of generative AI to assess whether the program built a general digital self-command capacity that extends beyond the technologies covered in the curriculum. Among students who reported using AI tools in the previous 30 days, we find that the treatment raises an index of purposeful AI use by 0.205 SDs ($p < 0.001$, Figure 9).[57] The index combines using AI for exam preparation, using it productively for that purpose, verifying its output, attempting problems before asking, and not substituting it for books and notes. Treatment students are 4.8 percentage points more likely to use AI while preparing for exams ($p = 0.004$) and 7.1 percentage points more likely to use it productively for that purpose ($p = 0.008$). Students also report less substitution of AI for books and notes ($p = 0.048$), while the estimate for verification is imprecise. This suggests that students applied a more deliberate approach to digital tools in a domain not covered in the curriculum.

7.3 Subgroup Effects and Heterogeneity

We examine heterogeneity by pre-specified baseline variables, including gender, urban residence, and age at first smartphone ownership. The null of equal treatment effects cannot be rejected at the 10 percent level in most cases, for either survey outcomes or objective phone use (Appendix

[55] Phone use during class is self-reported at follow-up. Because class schedules were collected only during the intervention period, the timetable-linked objective measure is unavailable at follow-up.

[56] Two checks required reading a short passage and answering a literal question, while the third instructed students to select a specified response. The checks appeared roughly one-third, two-thirds, and three-quarters of the way through the survey.

[57] The estimated effect on beliefs about AI's benefits and students' confidence using it is positive but imprecise (0.142 SDs, $p = 0.316$, Appendix Table A33).

Tables A34 and A35). One exception is that students who acquired their first smartphone at age eighteen or later increase commitment device use by 11.1 percentage points more than earlier adopters ($p = 0.018$). Total phone use also falls more among women than men, while the reductions in social media use are similar.

The discernment effect at endline is also concentrated among women. Assignment to the *Deliberate Use & Discernment* arm increases women's scores by 0.363 SDs relative to control ($p < 0.001$), whereas the estimate for men is -0.051 SDs and statistically indistinguishable from zero ($p = 0.516$). The difference between women and men is 0.414 SDs ($p < 0.001$).[58] Similar gender differences have appeared in prior misinformation interventions, although their direction varies across settings (Altay, De Angelis, and Hoes, 2024; Ali and Qazi, 2023).

Across all outcomes, no subgroup is detectably worse off, suggesting that the program can be delivered to full classrooms rather than targeted to some groups.

7.4 Role of Classroom Delivery

The program bundles content with classroom delivery, including facilitators, classmates, and opportunities to apply ideas through worksheets, polls, and discussions. To provide suggestive evidence on which parts of the bundle matter, we use a cross-randomized microlearning intervention that delivers the content privately, and friendship nominations to test for network effects.

We embedded an information provision experiment, which we refer to as microlearning, in the follow-up survey. This experiment was cross-randomized with the classroom intervention and placed in the latter part of the survey, after eliciting the key follow-up outcomes for the main classroom experiment. We randomly assigned batches to receive short, self-guided lessons on either English grammar (control) or digital self-command. The digital self-command lesson covered key curriculum ideas from the first two sessions, including platform design and commitment devices. Each lesson consisted of a few slides, each followed by a multiple-choice question and a page explaining the correct answer. Appendix E.1 provides additional details.[59]

Information provision improves understanding, but its effects on other outcomes are limited. Figure 10 shows that microlearning raises content knowledge by 0.160 SDs ($p = 0.002$) and digital literacy by 0.114 SDs ($p = 0.019$). Beyond understanding of key concepts, the figure shows little movement in beliefs about social media platforms, intentions to change use, or perceived social pressure, although we see a reduction in fear of disconnecting (-0.147 SDs, $p = 0.012$). While the two interventions differ in dosage as well as delivery, these results suggest that information provision

[58] Two participating institutions enroll only women, so the difference may partly reflect variation across colleges rather than heterogeneity between women and men within the same institutions. Excluding these institutions reduces the estimated difference, but the effect remains positive and statistically significant among women and indistinguishable from zero among men.

[59] This design replaced the text-message intervention described in the original pre-analysis plan, which we did not implement because pilots showed low take-up when materials were delivered through WhatsApp or text messages. The revised design was pre-specified in an amendment to the pre-registration before field deployment. Appendix G documents this change. The survey experiment avoids the take-up problem of text-message delivery and provides a benchmark for the impact of private information provision when students are actually exposed to the materials.

alone is unlikely to account for the broader effects of the classroom program.

We ask next what classroom delivery adds beyond the content itself. While gains in content knowledge and digital literacy are also evident in larger cohorts, point estimates suggest larger benefits for mental health, internal locus of control, discernment, and exam failure in cohorts with fewer than 45 students (Appendix Table A38). One possible explanation is that smaller classrooms allow more scope for discussion, practice, and feedback that reinforce learning. However, because cohort size is correlated with other factors that could affect outcomes, such as student composition, we examine the role of peers more directly using nominations of friends in other randomized batches.[60]

One channel through which peer effects could work is information transmission. If treated students explain the material to their friends in the control group, we would find that control students with greater treated-friend exposure are more likely to acquire knowledge and understanding of the program content. As discussed in Section 6.5, we find little evidence of such spillovers to the control group.

Another channel through which classroom delivery could affect outcomes is peer reinforcement. Social media use can create a product market trap, in which it is more costly to manage use when others do not (Bursztyn et al., 2025). When more friends receive the program, limits on use and slower responses online may become less socially costly and easier to sustain. Appendix Figure A15 shows that, among treated students, a one SD increase in exposure to treated friends is associated with a 0.092 SD increase in internal locus of control ($p = 0.002$).[61] The corresponding gradient in content knowledge is smaller and not statistically different from zero. This suggests that treated friends may help students feel able to act on what they learned, rather than simply passing on information.

To understand how well this program can scale, we perform a cost-effectiveness analysis in Appendix F. The two-session *Deliberate Use* curriculum costs approximately USD 2.29 per student, and the complete three-session curriculum costs USD 3.43. Even doubling these amounts to allow for the opportunity cost of the class or student time used for delivery leaves the complete curriculum at USD 6.86 per student. Translated to key outcomes, this implies around USD 57 per exam failure or reappearance avoided, or USD 0.19 per hour of reduced phone use over 30 days. This suggests that the program is highly cost-effective. It is also short, and the misinformation session can be offered as an optional addition to the two-session *Deliberate Use* curriculum. Viewed through the framework of List (2024), the intervention therefore combines low costs with a delivery model that has already been tested under conditions of real-world implementation. Scaling could leverage the existing infrastructure of educational institutions and focus on preserving the engaged classroom experience.

[60] Students could nominate up to three other students at their institution with whom they regularly interacted online or in person. We match nominations to survey respondents using the nominated student's name, program, and year to identify their batch and treatment assignment (Appendix E.2).

[61] Appendix Table A24 reports results across outcomes.

8 Conclusion

This study shows that education can build digital self-command, helping students exercise informed judgment and follow through on their intentions in digital environments. Students in the treatment group were more likely to adopt strategies for managing their smartphone use. With targeted instruction, they also became better at judging online claims. The benefits extended beyond reductions in screen time to improved mental health and academic performance. Nearly a year after the intervention, we continue to see evidence of students applying this capacity, including to new domains, such as generative AI, that the program did not cover.

Our findings suggest that students learned to anticipate and manage tempting moments. Both treatment and control students had access to the same tools for managing their use, but students in treatment were more likely to recognize predictable cues and adopt corresponding strategies to manage their use. The reduction in smartphone use was concentrated in social media, while students continued to value their access to it, suggesting that they changed how they used technology rather than disengaging from it altogether.

The program is inexpensive and practical to deliver through educational institutions. Our study shows that the program can be incorporated within the regular school schedule and delivered with standardized materials and facilitator training. It does not require individual-level coaching, school-wide enforcement, or introduction of new tools. We also find that classroom-based delivery matters, as a short self-guided version increased knowledge but did not change beliefs or behavioral intentions, so information alone is not sufficient.

Our findings inform public debate around social media regulations. Recent regulations and legislative proposals have largely focused on restricting use. However, these policies do not address what students do once access is available. Education can be a policy lever to consider alongside restrictions. For example, minimum-age rules govern when a person may begin driving, while driver education is designed to prepare people to make decisions once they have access. Similarly, restrictions can delay or limit access to social media for younger populations, while education can help people manage their use once access is available.

An open question is whether the timing of the intervention matters for its effectiveness. In our study, almost all students already owned a smartphone and used social media by the time of the program. It might be more effective to introduce the program before students adopt smartphones or social media, or to deliver it to younger populations, who are in a critical period of development. Reaching students before these habits form could yield larger returns.

Use of digital technologies will continue to be widespread, and new technologies will continue to emerge. People must decide when and how to use technologies that are valuable in some settings but conflict with their intentions in others. As technologies evolve, building digital self-command could help us stay in control.

References

Abrahamsson, Sara. 2026. "Smartphone Bans, Student Outcomes and Mental Health." *Journal of Human Resources* Ahead of print.

Ahn, Woo-kyoung and Charles W. Kalish. 2000. "The Role of Mechanism Beliefs in Causal Reasoning." In *Explanation and Cognition*, edited by Frank C. Keil and Robert A. Wilson. Cambridge, MA: MIT Press, 199–226.

Aksoy, Billur, Lester Lusher, and Scott Carrell. 2025. "From distraction to dedication: Commitment and incentives against phone use in the classroom." *Journal of Economic Behavior & Organization* 236:107082.

Alan, Sule, Ceren Baysan, Mert Gumren, and Elif Kubilay. 2021. "Building Social Cohesion in Ethnically Mixed Schools: An Intervention on Perspective Taking." *The Quarterly Journal of Economics* 136 (4):2147–2194.

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

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

Alan, Sule, Seda Ertac, and İnci Gümüş. 2021. "Does a Forward-Looking Perspective Affect Self-Control and the Demand for Commitment? Results from an Educational Intervention." *Economic Inquiry* 59 (4):1533–1546.

Alan, Sule and Ipek Mumcu. 2024. "Nurturing Childhood Curiosity to Enhance Learning: Evidence from a Randomized Pedagogical Intervention." *American Economic Review* 114 (4):1173–1210.

Ali, Ayesha and Ihsan Ayyub Qazi. 2023. "Countering Misinformation on Social Media Through Educational Interventions: Evidence from a Randomized Experiment in Pakistan." *Journal of Development Economics* 163:103108.

Allcott, Hunt, E. Jason Baron, Thomas Dee, Angela L. Duckworth, Matthew Gentzkow, and Brian Jacob. 2026. "The Effects of School Phone Bans: National Evidence from Lockable Pouches." Working Paper 35132, National Bureau of Economic Research. URL `https://www.nber.org/papers/w35132`.

Allcott, Hunt, Matthew Gentzkow, and Lena Song. 2022. "Digital Addiction." *American Economic Review* 112 (7):2424–2463.

Altay, Sacha, Andrea De Angelis, and Emma Hoes. 2024. "Media literacy tips promoting reliable news improve discernment and enhance trust in traditional media." *Communications Psychology* 2 (1):74.

Amar, Priyadarshi, Sumitra Badrinathan, Simon Chauchard, and Florian Sichart. 2026. "Countering Misinformation Early: Evidence from a Classroom-Based Field Experiment in India." *American Political Science Review* 120 (2):725–745.

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association* 103 (484):1481–1495.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. 2024. "The Economics of Social Media." *Journal of Economic Literature* 62 (4):1422–1474.

ASER. 2023. "Annual Status of Education Report (ASER) 2023: Beyond Basics." URL `https://asercentre.org/Annual-Status-of-Education-Report-ASER-2023-Beyond-Basics`.

Ashraf, Nava, Dean Karlan, and Wesley Yin. 2006. "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines." *The Quarterly Journal of Economics* 121 (2):635–672.

Athey, Susan, Matias Cersosimo, Kristine Koutout, and Zelin Li. 2023. "Emotion-versus Reasoning-based Drivers of Misinformation Sharing: A field experiment using text message courses in Kenya." *Available at SSRN 4489759* .

Athey, Susan and Guido W Imbens. 2017. "The econometrics of randomized experiments." In *Handbook of economic field experiments*, vol. 1. Elsevier, 73–140.

Attri, Akanksha. 2025. "Panjab Varsity Results Spark Concern as Pass Rates Plunge Under NEP." *Hindustan Times* URL `https://www.hindustantimes.com/cities/chandigarh-news/panjab-varsity-results-spark-concern-as-pass-rates-plunge-under-nep-101749671513125.html`.

Augenblick, Ned, B. Kelsey Jack, Supreet Kaur, Felix Masiye, and Nicholas Swanson. 2024. "Retrieval Failures and Consumption Smoothing: A Field Experiment on Seasonal Poverty." Working paper, University of California, Berkeley.

Badrinathan, Sumitra. 2021. "Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India." *American Political Science Review* 115 (4):1325–1341.

Bai, Liang, Benjamin Handel, Edward Miguel, and Gautam Rao. 2021. "Self-Control and Demand for Preventive Health: Evidence from Hypertension in India." *Review of Economics and Statistics* 103 (5):835–856.

Bandiera, Oriana, Amen Jalal, and Nina Roussille. 2025. "The Illusion of Time: Gender Gaps in Job Search and Employment." NBER Working Paper 34051, National Bureau of Economic Research. Revised December 2025.

Berger, Lara Marie, Anna Kerkhof, Felix Mindl, and Johannes Munster. 2023. "Debunking "Fake News" on Social Media: Short-term and Longer-term Effects of Fact Checking and Media Literacy Interventions." .

Bernheim, B. Douglas and Antonio Rangel. 2004. "Addiction and Cue-Triggered Decision Processes." *American Economic Review* 94 (5):1558–1590.

Blair, Robert A., Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J. Stainfield. 2024. "Interventions to Counter Misinformation: Lessons from the Global North and Applications to the Global South." *Current Opinion in Psychology* 55:101732.

Bowles, Jeremy, Kevin Croke, Horacio Larreguy, John Marshall, and Shelley Liu. 2023. "Sustaining Exposure to Fact-checks: Misinformation Discernment, Media Consumption, and its Political Implications." *Media Consumption, and its Political Implications* .

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin. 2022. "Social Media and Mental Health." *American Economic Review* 112 (11):3660–3693.

Briole, Simon, Julia Cagé, and Andrea Prat. 2026. "Making Teenagers Read Newspapers: A Nationwide Experiment in French High Schools." CEPR Discussion Paper DP21706, Centre for Economic Policy Research. URL `https://cepr.org/publications/dp21706`.

Brown, Christina, Supreet Kaur, Geeta Kingdon, and Heather Schofield. 2025. "Cognitive Endurance as Human Capital." *The Quarterly Journal of Economics* 140 (2):943–1002.

Bursztyn, Leonardo, Angela L Duckworth, Rafael Jiménez-Durán, Aaron Leonard, Filip Milojević, Christopher Roth, and Cass R Sunstein. 2026. "Why Bans Fail: Tipping Points and Australia's Social Media Ban." Tech. rep., National Bureau of Economic Research.

Bursztyn, Leonardo, Benjamin R. Handel, Rafael Jiménez-Durán, and Christopher Roth. 2025. "When Product Markets Become Collective Traps: The Case of Social Media." *American Economic Review* 115 (12):4105–4136.

Carrera, Mariana, Heather Royer, Mark Stehr, Justin Sydnor, and Dmitry Taubinsky. 2022. "Who Chooses Commitment? Evidence and Welfare Implications." *The Review of Economic Studies* 89 (3):1205–1244.

Charlton, John P and Ian DW Danforth. 2007. "Distinguishing addiction and high engagement in the context of online game playing." *Computers in human behavior* 23 (3):1531–1548.

Chuan, Amanda, Anya Samek, and Shreemayi Samujjwala. 2024. "Evaluating the Impact of Preschool on Patience, Time Inconsistency and Commitment Demand." Working paper, SSRN. URL `https://ssrn.com/abstract=4704020`.

Cobb-Clark, Deborah A, Juliana Silva-Goncalves, Agnieszka Tymula, and Xueting Wang. 2026. "The Conceptual Foundations of Self-Control and its Link to Impulsivity and Attention." *Available at SSRN 6301058* .

Credé, Marcus, Sylvia G Roch, and Urszula M Kieszczynka. 2010. "Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics." *Review of Educational Research* 80 (2):272–295.

Diener, ED, Robert A Emmons, Randy J Larsen, and Sharon Griffin. 1985. "The Satisfaction with Life Scale." *Journal of Personality Assessment* 49 (1):71–75.

Donati, Dante, Ruben Durante, Francesco Sobbrio, and Dijana Zejcirovic. 2022. "Lost in the Net? Broadband Internet and Youth Mental Health." *Broadband Internet and Youth Mental Health* .

Duckworth, Angela L., Tamar Szabó Gendler, and James J. Gross. 2016. "Situational Strategies for Self-Control." *Perspectives on Psychological Science* 11 (1):35–55.

Duckworth, Angela L., Katherine L. Milkman, and David Laibson. 2018. "Beyond Willpower: Strategies for Reducing Failures of Self-Control." *Psychological Science in the Public Interest* 19 (3):102–129.

Duckworth, Angela L., Rachel E. White, Alyssa J. Matteucci, Annie Shearer, and James J. Gross. 2016. "A Stitch in Time: Strategic Self-Control in High School and College Students." *Journal of Educational Psychology* 108 (3):329–341.

Duradoni, Mirko, Gioele Salvatori, Sara Meacci, Gabriele Panerai, and Andrea Guazzini. 2023. "Development and validation of the Internet Locus of Control Scale (I-LOC)." *New media & society* 25 (6):1227–1247.

Ek, Claes and Margaret Samahita. 2023. "Too Much Commitment? An Online Experiment with Tempting YouTube Content." *Journal of Economic Behavior & Organization* 208:21–38.

Embretson, Susan E. and Steven P. Reise. 2000. *Item Response Theory for Psychologists*. Multivariate Applications Book Series. Mahwah, NJ: Lawrence Erlbaum Associates.

Ericson, Keith Marzilli and David Laibson. 2019. "Intertemporal Choice." In *Handbook of Behavioral Economics: Foundations and Applications 2*, vol. 2, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, chap. 1. Elsevier.

Figlio, David N. and Umut Özek. 2025. "The Impact of Cellphone Bans in Schools on Student Outcomes: Evidence from Florida." Working Paper 34388, National Bureau of Economic Research. URL `https://www.nber.org/papers/w34388`.

Giuntella, Osea, Silvia Saccardo, and Sally Sadoff. 2024. "Sleep: Educational Impact and Habit Formation." Tech. rep., National Bureau of Economic Research.

Gollwitzer, Peter M. 1999. "Implementation Intentions: Strong Effects of Simple Plans." *American Psychologist* 54 (7):493–503.

Government of India, Ministry of Finance. 2026. "Economic Survey 2025–26." URL `https://www.indiabudget.gov.in/economicsurvey/`. Tabled in Parliament, 29 January 2026.

Griffiths, Mark. 2005. "A 'Components' Model of Addiction within a Biopsychosocial Framework." *Journal of Substance Use* 10 (4):191–197.

Griffiths, Mark D. 2013. "Is "loss of control" always a consequence of addiction?" *Frontiers in Psychiatry* 4:36.

Guess, Andrew M, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. "A Digital Media Literacy Intervention Increases Discernment Between Mainstream and False News in the United States and India." *Proceedings of the National Academy of Sciences* 117 (27):15536–15545.

Gui, Marco, Tiziano Gerosa, Gianluca Argentin, and Lucilla Losi. 2023. "Mobile media education as a tool to reduce problematic smartphone use: Results of a randomised impact evaluation." *Computers & Education* 194:104705.

Gul, Faruk and Wolfgang Pesendorfer. 2001. "Temptation and Self-Control." *Econometrica* 69 (6):1403–1435.

Guriev, Sergei, Emeric Henry, Théo Marquis, and Ekaterina Zhuravskaya. 2023. "Curtailing False News, Amplifying Truth." *SSRN* .

Hanna, Rema, Sendhil Mullainathan, and Joshua Schwartzstein. 2014. "Learning Through Noticing: Theory and Evidence from a Field Experiment." *The Quarterly Journal of Economics* 129 (3):1311–1353.

Hays, Ron D, Toshi Hayashi, and Anita L Stewart. 1989. "A five-item measure of socially desirable response set." *Educational and Psychological Measurement* 49 (3):629–636.

Himmler, Oliver, Robert Jäckle, and Philipp Weinschenk. 2019. "Soft commitments, reminders, and academic performance." *American Economic Journal: Applied Economics* 11 (2):114–142.

Hoong, Ruru. 2021. "Self control and smartphone use: An experimental study of soft commitment devices." *European Economic Review* 140:103924.

Hughes, Mary Elizabeth, Linda J Waite, Louise C Hawkley, and John T Cacioppo. 2004. "A Short Scale for Measuring Loneliness in Large Surveys: Results from Two Population-Based Studies." *Research on Aging* 26 (6):655–672.

Joshi, Mridul, Jalnidh Kaur, and Lena Song. 2024. "Digital Empowerment for Youth: Experimental Evidence from India." Registered Report Stage 1: Proposal, Journal of Development Economics manuscript draft (DEVEC-D-24-01036R1).

Kaur, Jalnidh. 2025. "How Much Do I Matter? Teacher Self-Beliefs, Effort, and Student Learning." Tech. rep., Working Paper.

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

Keil, Frank C. 2006. "Explanation and Understanding." *Annual Review of Psychology* 57:227–254.

Kemp, Simon. 2026. "Digital 2026: India." Available at: `https://datareportal.com/reports/digital-2026-india`.

Kerwin, Jason, Nada Rostom, and Olivier Sterck. 2024. "Striking the Right Balance: Why Standard Balance Tests Over-Reject the Null, and How to Fix It." IZA Discussion Paper 17217, Institute of Labor Economics (IZA). URL `https://ssrn.com/abstract=4926535`.

Kozyreva, Anastasia, Philipp Lorenz-Spreen, Stefan M. Herzog, Ullrich K. H. Ecker, Stephan Lewandowsky, Ralph Hertwig et al. 2024. "Toolbox of Individual-Level Interventions against Online Misinformation." *Nature Human Behaviour* 8 (6):1044–1052.

Kroenke, Kurt, Robert L Spitzer, and Janet BW Williams. 2001. "The PHQ-9: Validity of a Brief Depression Severity Measure." *Journal of General Internal Medicine* 16 (9):606–613.

Laibson, David. 1997. "Golden Eggs and Hyperbolic Discounting." *Quarterly Journal of Economics* 112 (2):443–478.

———. 2001. "A Cue-Theory of Consumption." *Quarterly Journal of Economics* 116 (1):81–119.

Lichand, Guilherme, Luca Moreno-Louzada, Thiago da Costa, and Matthew Gentzkow. 2026. "The Educational Impacts of School Phone Bans: Evidence from Brazil." Working paper. URL `https://web.stanford.edu/~gentzkow/research/RioPhoneBans.pdf`.

List, John, Lina Ramirez, Julia Seither, Jamie Unda, and Beatriz Vallejo. 2024. "Toward an Understanding of the Economics of Misinformation: Evidence from a Demand Side Field Experiment on Critical Thinking." Framed Field Experiments 00786, The Field Experiments Website. URL `https://ideas.repec.org/p/feb/framed/00786.html`.

List, John A. 2024. "Optimally Generate Policy-Based Evidence Before Scaling." *Nature* 626:491–499.

Lombrozo, Tania. 2012. "Explanation and Abductive Inference." In *The Oxford Handbook of Thinking and Reasoning*, edited by Keith J. Holyoak and Robert G. Morrison. Oxford: Oxford University Press, 260–276.

Lord, Frederic M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lyubomirsky, Sonja and Heidi S Lepper. 1999. "A Measure of Subjective Happiness: Preliminary Reliability and Construct Validation." *Social Indicators Research* 46:137–155.

McDool, Emily, Philip Powell, Jennifer Roberts, and Karl Taylor. 2020. "The internet and children's psychological wellbeing." *Journal of Health Economics* 69:102274.

McKelway, Madeline. 2025. "Women's Self-Efficacy and Economic Outcomes: Evidence from a Two-Stage Experiment in India." Tech. rep., Working Paper.

O'Donoghue, Ted and Matthew Rabin. 1999. "Doing It Now or Later." *American Economic Review* 89 (1):103–124.

OECD. 2026. "Growing Up in the Social Media Age." OECD Digital Economy Papers 385, OECD Publishing, Paris.

Office of the Surgeon General. 2023. "Protecting Youth Mental Health: The U.S. Surgeon General's Advisory on Social Media and Risks." URL `https://www.hhs.gov/sites/default/files/sg-youth-mental-health-social-media-advisory.pdf`. Accessed: 2024-06-25.

Orkin, Kate, Robert Garlick, Mahreen Mahmud, Richard Sedlmayr, Johannes Haushofer, and Stefan Dercon. 2025. "Aspiring to a Better Future: Can a Simple Psychological Intervention Reduce Poverty?" *The Review of Economic Studies* Forthcoming.

Otsuka, Yuichiro, Yoshitaka Kaneita, Osamu Itani, and Yuuki Matsumoto. 2023. "A school-based program for problematic internet use for adolescents in Japan." *Children* 10 (11):1754.

Parry, Douglas A., Brittany I. Davidson, Craig J. R. Sewall, Jacob T. Fisher, Hannah Mieczkowski, and Daniel S. Quintana. 2021. "A Systematic Review and Meta-Analysis of Discrepancies Between Logged and Self-Reported Digital Media Use." *Nature Human Behaviour* 5 (11):1535–1547.

Pearlin, Leonard I and Carmi Schooler. 1978. "The Structure of Coping." *Journal of Health and Social Behavior* :2–21.

Purington Drake, Amanda, Philipp K Masur, Natalie N Bazarova, Wenting Zou, and Janis Whitlock. 2023. "The Youth Social Media Literacy Inventory: Development and Validation Using Item Response Theory in the US." *Journal of Children and Media* 17 (4):467–487.

Romer, David. 1993. "Do students go to class? Should they?" *Journal of economic perspectives* 7 (3):167–174.

Sankar, Anirudh, Robert Dulin, Benjamin Davies, Vesall Nourani, Jess Rudder, Abraham Salomon, and Godfrey Taulya. 2025. "How Mechanistic Explanations Reshape Learning and Behavior: Evidence from a Fertilizer Choice Experiment in Eastern Uganda." Working Paper.

Schelling, Thomas C. 1984. "Self-command in practice, in policy, and in a theory of rational choice." *The American Economic Review* 74 (2):1–11.

Schilbach, Frank. 2019. "Alcohol and Self-Control: A Field Experiment in India." *American Economic Review* 109 (4):1290–1322.

Schunk, Daniel, Eva M Berger, Henning Hermes, Kirsten Winkel, and Ernst Fehr. 2022. "Teaching self-regulation." *Nature human behaviour* 6 (12):1680–1690.

Smith, Adam. 1759. *The Theory of Moral Sentiments*. London and Edinburgh: A. Millar; and A. Kincaid and J. Bell.

Spitzer, Robert L, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. 2006. "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7." *Archives of Internal Medicine* 166 (10):1092–1097.

Sungu, Alp, Pradeep Kumar Choudhury, and Andreas Bjerre-Nielsen. 2025. "Removing phones from classrooms improves academic performance." *Available at SSRN 5370727* .

Suresh, Nileena. 2026. "Most Indians Say They Cannot Verify If Information Online Is Reliable." Data For India, Data Bytes. URL `https://www.dataforindia.com/data-bytes/most-indians-say-they-cannot-verify-if-information-online-is-reliable/`.

Tandoc Jr, Edson C, Patrick Ferrucci, and Margaret Duffy. 2015. "Facebook Use, Envy, and Depression Among College Students: Is Facebooking Depressing?" *Computers in Human Behavior* 43:139–146.

Tangney, June P, Roy F Baumeister, and Angie Luzio Boone. 2004. "High self-control predicts good adjustment, less pathology, better grades, and interpersonal success." *Journal of personality* 72 (2):271–324.

Thaler, Richard H and Hersh M Shefrin. 1981. "An economic theory of self-control." *Journal of political Economy* 89 (2):392–406.

TRAI. 2026. "Indian Telecom Services Performance Indicator Report January–March 2026." URL `https://www.trai.gov.in/node/18167`. Accessed: 2026-07-31.

Weinstein, Emily and Carrie James. 2022. "School-Based Initiatives Promoting Digital Citizenship and Healthy Digital Media Use." *Handbook of Adolescent Digital Media Use and Mental Health; Cambridge University Press: Cambridge, UK* :365.

World Bank. 2016. *World development report 2016: Digital dividends*. World Bank Publications.

Zhao, Anqi and Peng Ding. 2024. "To Adjust or Not to Adjust? Estimating the Average Treatment Effect in Randomized Experiments with Missing Covariates." *Journal of the American Statistical Association* 119 (545):450–460.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov. 2020. "Political Effects of the Internet and Social Media." *Annual Review of Economics* 12:415–438.

Figures and Tables

Table 1: Balance at baseline

	Control mean (1)	Difference (2)	Observations (3)
Age at baseline	19.157	0.187	2,862
		(0.113)	
Sex is female	0.481	0.038**	2,861
		(0.016)	
Home in urban area	0.726	0.014	2,876
		(0.016)	
Household wealth index (standardized)	0.016	-0.025	2,852
		(0.029)	
Owned smartphone before 18	0.740	-0.036*	2,802
		(0.021)	
Reported social-media use (hours/day)	4.378	-0.135	2,783
		(0.135)	
Digital literacy	0.051	-0.005	2,794
		(0.035)	
Interest in setting limits	0.372	0.025	2,776
		(0.020)	
Commitment-device use	0.406	0.007	2,777
		(0.022)	
Digital addiction	-0.017	-0.049	2,771
		(0.036)	
Sleep index	0.022	-0.067**	2,750
		(0.032)	
Wellbeing index	0.020	0.019	2,755
		(0.041)	
Joint test p-value: pooled treatment		0.551	
Misinformation discernment	0.042	0.026	1,846
		(0.071)	
Joint test p-value: discernment arm		0.482	

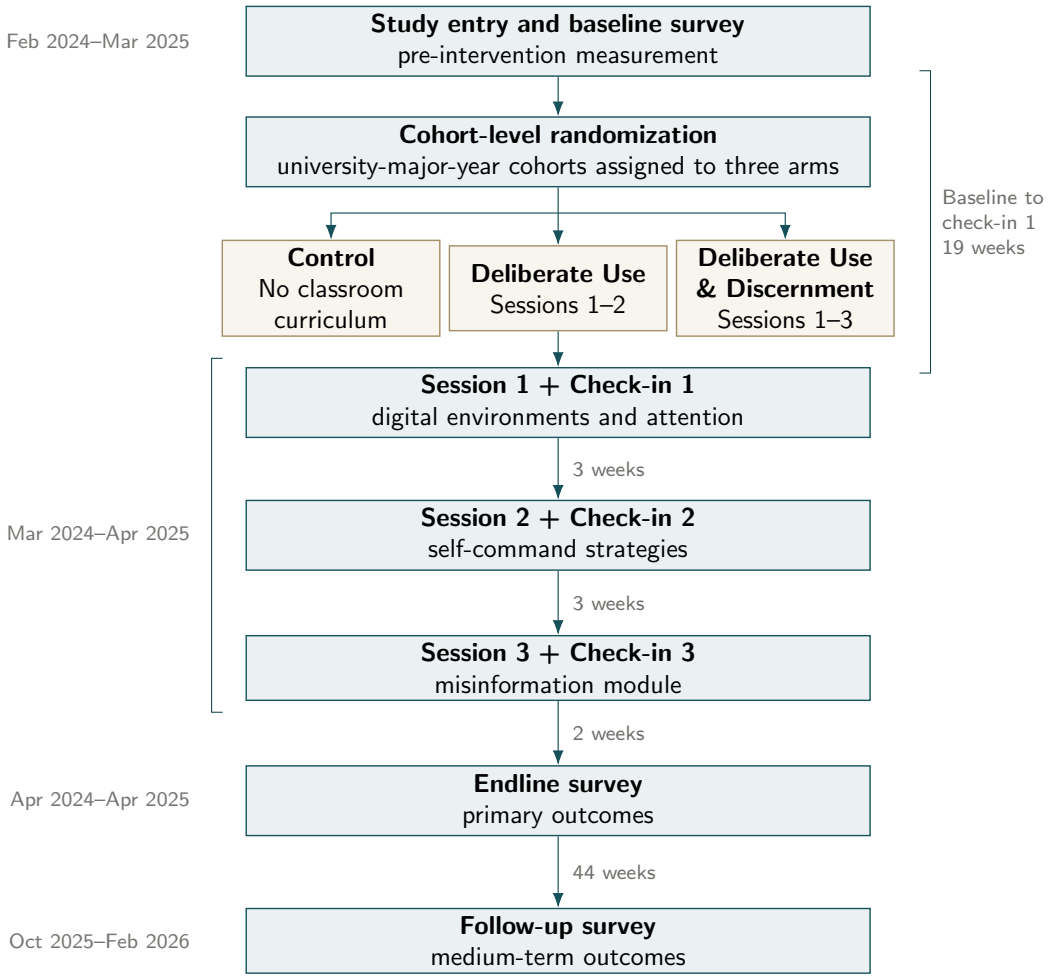
Notes: The unit of observation is a student. Control means and differences use weights inversely proportional to cohort size before the intervention. Regressions include randomization stratum fixed effects and standard errors clustered at the batch level. Each row uses baseline respondents with a nonmissing value of the listed outcome. The misinformation row is available for 60 batches and compares the *Deliberate Use & Discernment* arm with control. Joint tests use observations with complete data for the outcomes in the test. Their p-values use 999 treatment assignments drawn within randomization strata at the batch level.

Table 2: Survey participation at endline

	Observed by check-in 1 (1)	Observed at baseline (2)	Observed in at least two pre-endline surveys (3)
Treatment	-0.009	-0.002	-0.009
	(0.022)	(0.021)	(0.019)
Control mean	0.744	0.761	0.825
Observations	3,430	2,880	2,665

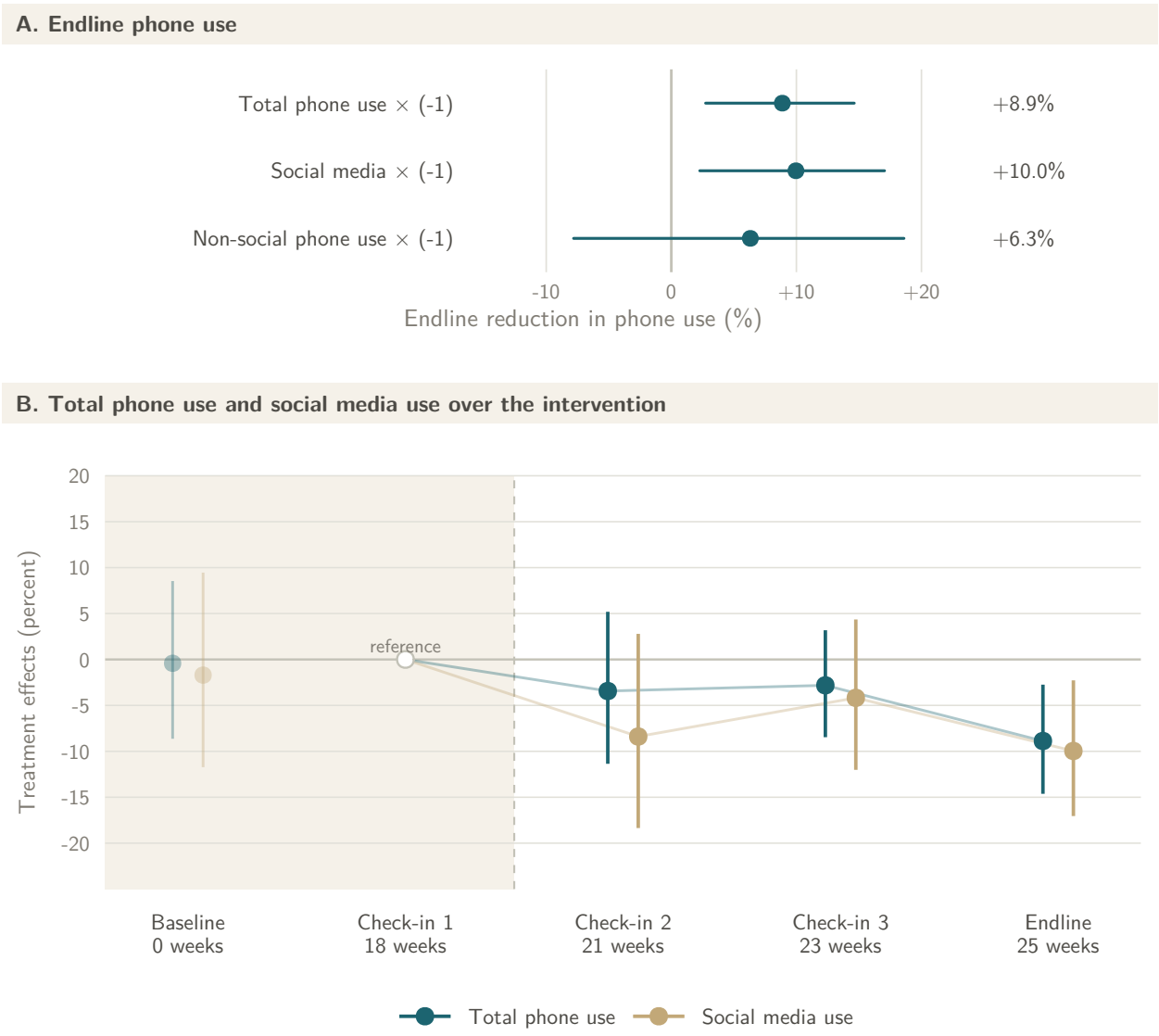
Notes: The dependent variable equals one when a student participates in the endline survey. The unit of observation is a student. Regressions include randomization-stratum fixed effects and are unweighted. Standard errors are clustered at the batch level. Column (1) includes students observed by check-in 1. Column (2) restricts to students first observed at baseline. Column (3) restricts to students observed in at least two of the four pre-endline surveys. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Figure 1: Experimental design and measurement timeline



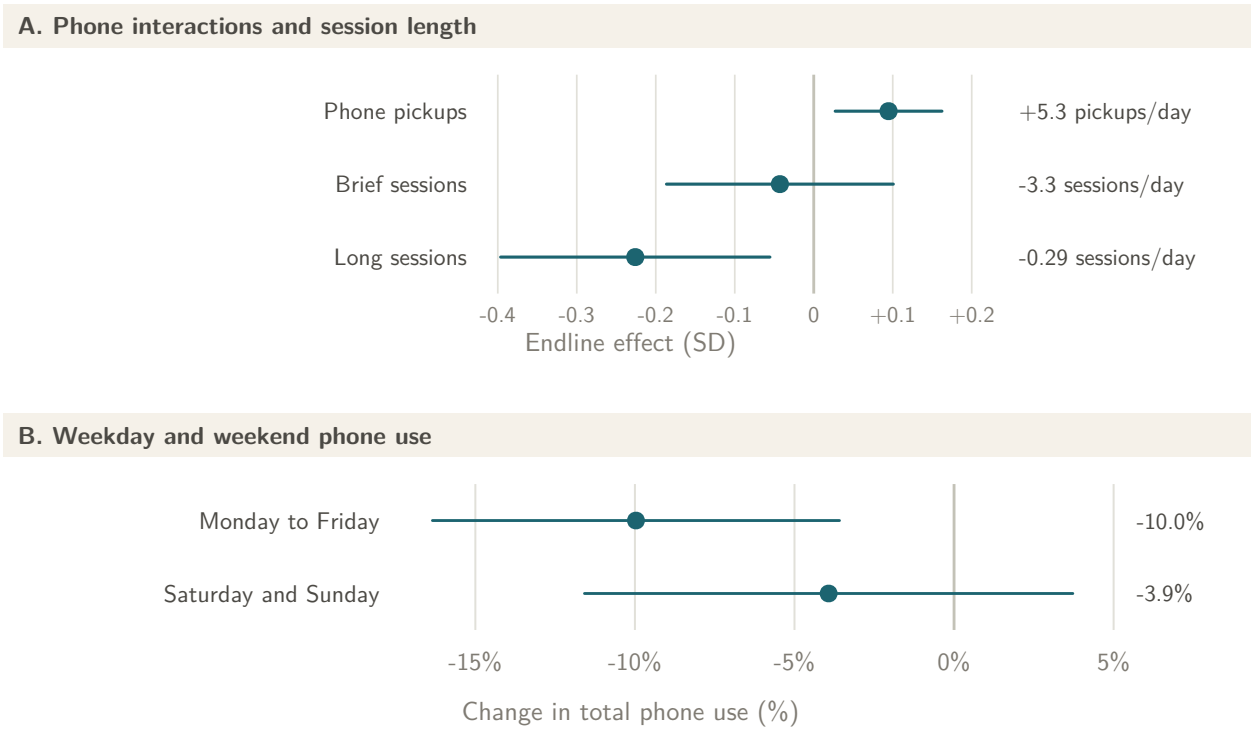
Notes: The figure summarizes the experimental design and measurement timeline. The unit of randomization is a batch: an institution-major-year group, or a section within it where enrollment is large enough to require sections. Each batch was assigned to one of three arms: control, which received no classroom curriculum; *Deliberate Use*, which received Sessions 1 and 2; and *Deliberate Use & Discernment*, which added the Session 3 misinformation module. All arms, including control, completed the same baseline survey, three check-ins during the intervention period, endline, and follow-up on a common schedule. Each treatment session was paired with a check-in survey. Labels on the left report the calendar range over which each stage was completed; these ranges overlap because batches entered the study on different schedules. The bracket reports the mean time from baseline to check-in 1. Values between the later boxes report the mean number of weeks between adjacent stages, measured at the batch level.

Figure 2: Treatment effects on objective phone use



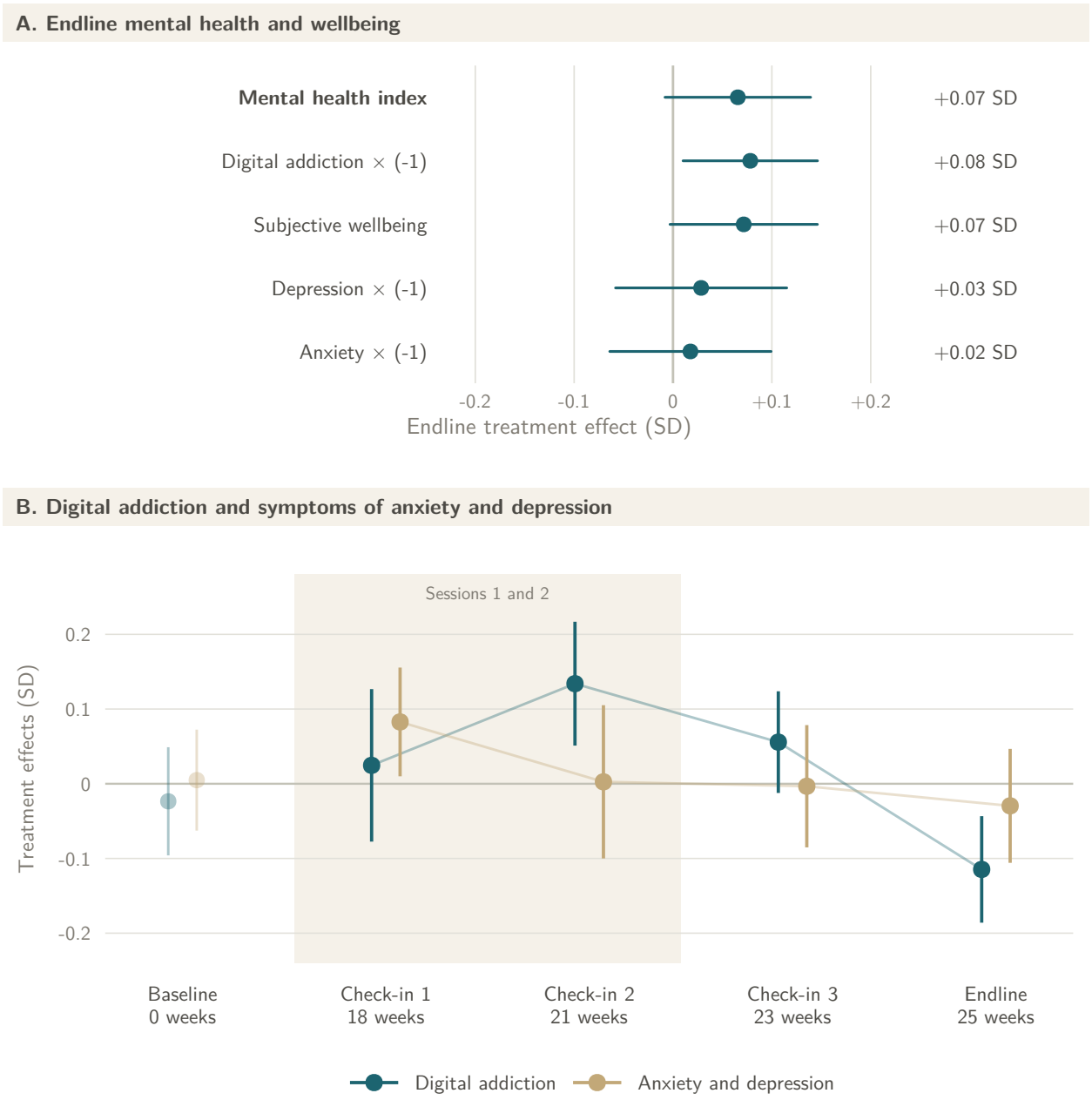
Notes: Panel A plots endline effects on total phone use, social media use, and non-social media phone use. Outcomes in Panel A are reversed for plotting so reductions appear to the right of zero and are marked with $\times(-1)$. The right column reports the corresponding percentage effects. Panel B plots the changes in the treatment–control differences in total phone use and social media use relative to check-in 1 and retains original signs. Estimates use log-transformed outcomes and are expressed as percentage changes. Each window contains up to three valid days before the survey. The baseline points come from separate regressions using all available observations in the baseline and check-in 1 windows. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Panel A of Appendix Table A4.

Figure 3: Treatment effects on phone interactions, session length, and timing



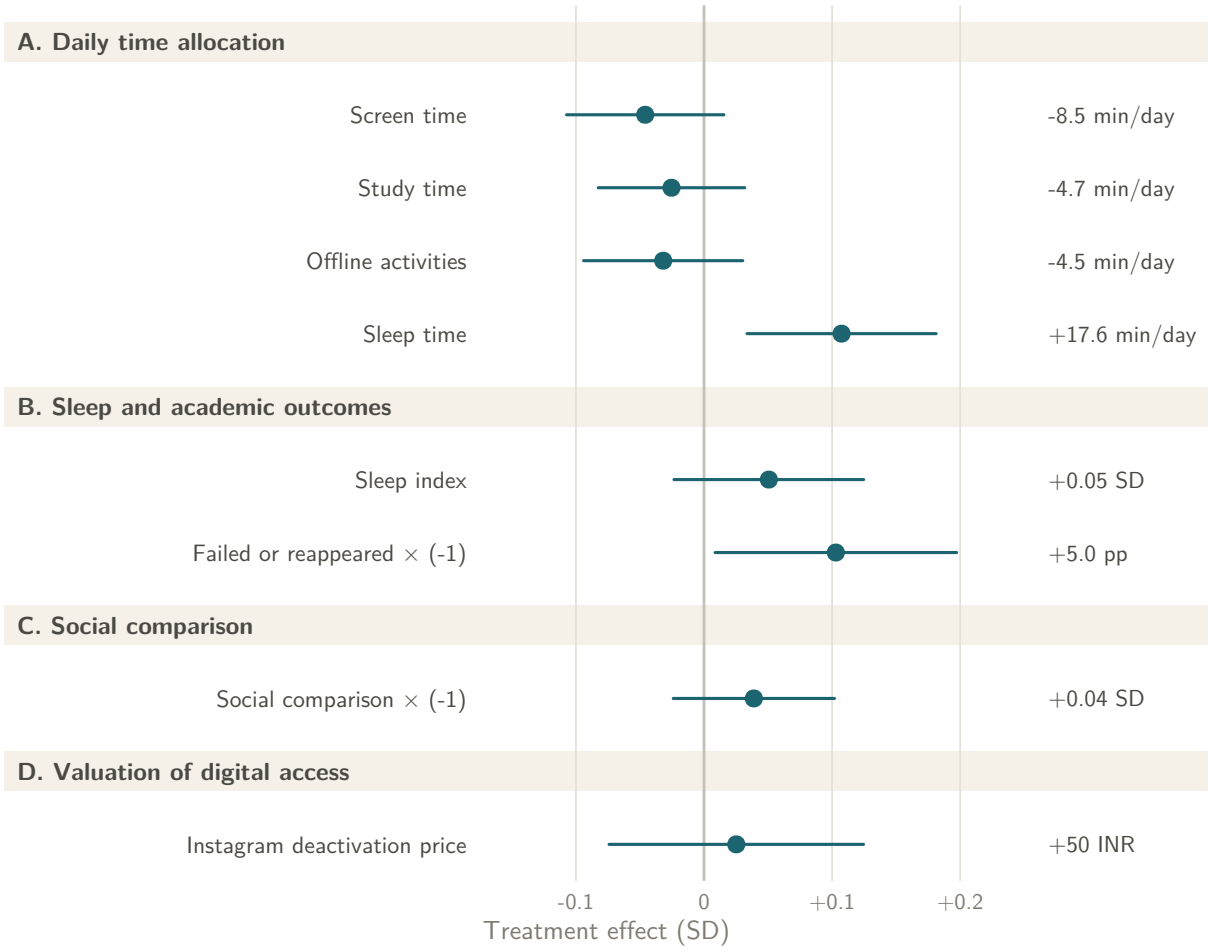
Notes: Panel A plots endline effects on daily phone pickups and brief and long app sessions in control-group standard deviation units. A pickup is a screen activation after at least 30 seconds with the screen off. Brief app sessions last at least one second and less than two seconds. Long app sessions exceed 30 minutes. The two session categories exclude sessions of other lengths. The right column reports pickups per day and sessions per day. Panel B plots endline percentage effects on total phone use from Monday through Friday and on Saturday and Sunday. Panel A uses up to three valid days before each survey. Panel B uses up to seven days so each window can cover a full week. Both panels retain original signs. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Appendix Tables A6 and A7.

Figure 4: Treatment effects on digital addiction and mental health



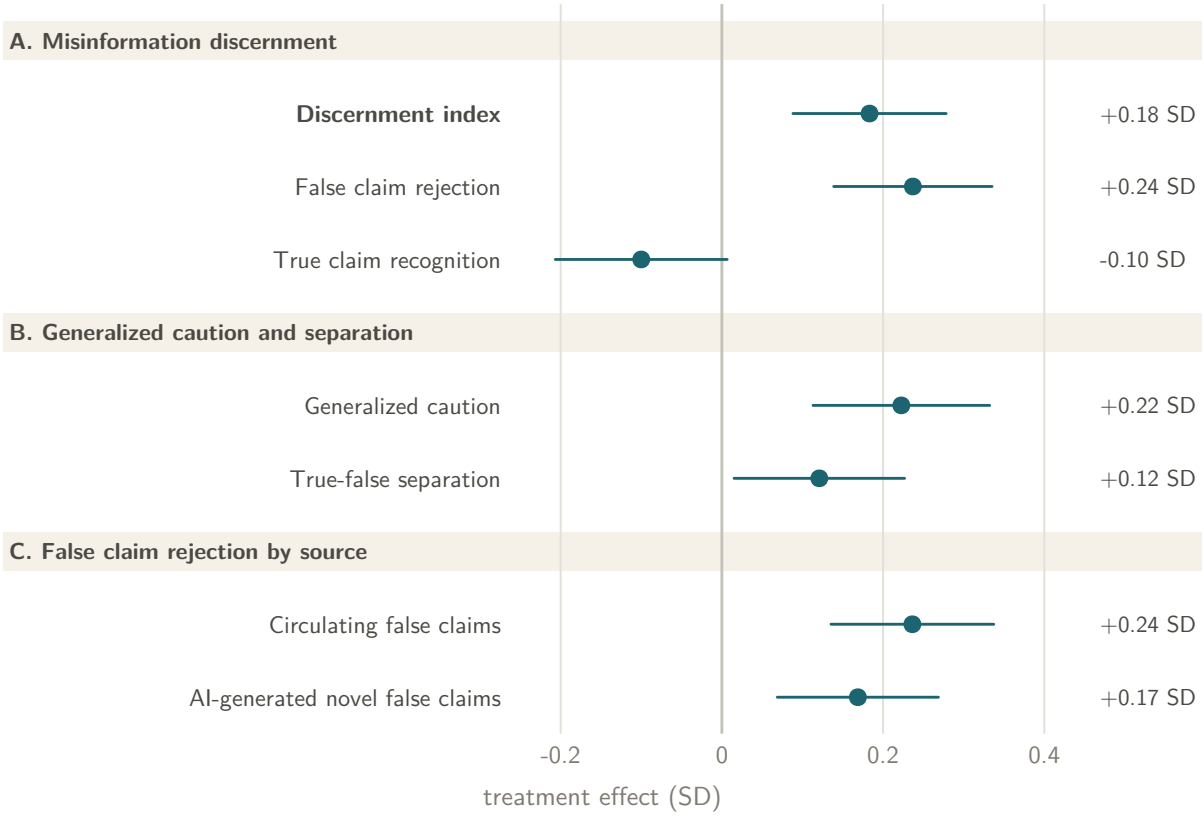
Notes: Panel A plots endline effects on the mental health index, digital addiction using the full battery, subjective wellbeing, depression using the PHQ-9, and anxiety using the GAD-7. Positive effects in Panel A indicate improvement. Components marked with $\times(-1)$ are reversed for plotting. Panel B plots digital addiction using the items common to all five surveys and an Anderson index combining anxiety and depression symptoms from the PHQ-2 and GAD-2. Panel B retains original signs, so negative values indicate reductions. The shaded band marks Sessions 1 and 2. Post-baseline regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. Baseline points are weighted treatment–control differences within randomization strata. Estimates are reported in control-group standard deviation units. Error bars report 95 percent confidence intervals clustered at the batch level. Endline estimates correspond to Appendix Tables A8 and A9.

Figure 5: Treatment effects on secondary and exploratory outcomes



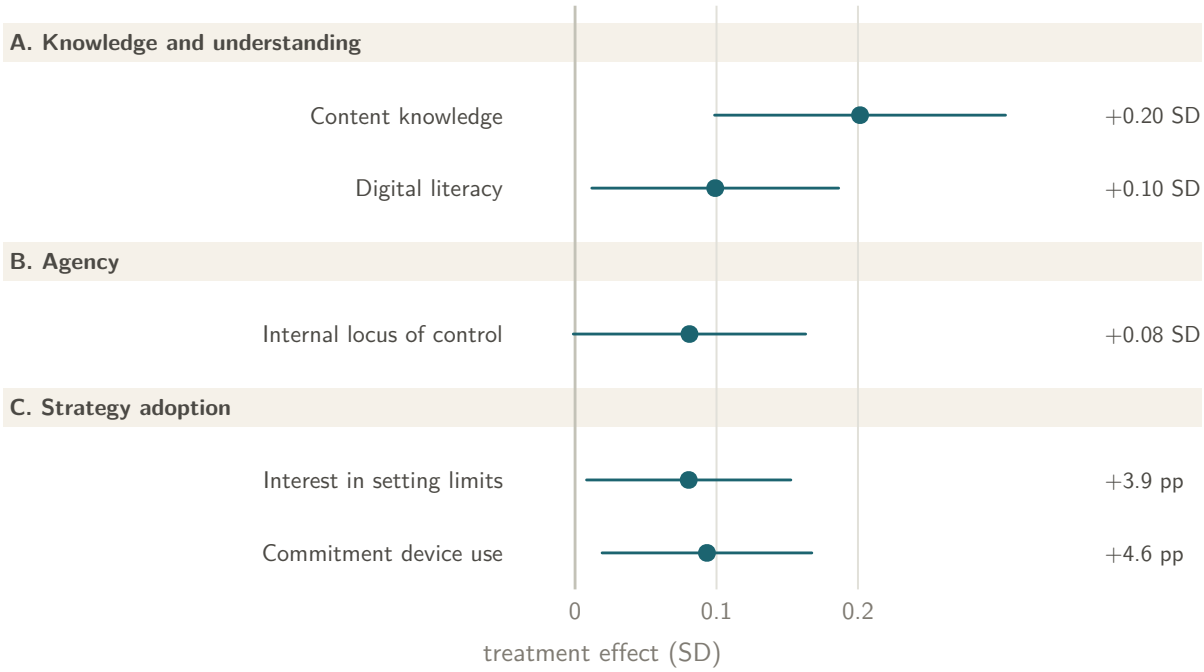
Notes: Panel A plots endline effects on self-reported daily time allocation. Panel B reports the sleep index and failure or reappearance at the first regular semester exam after Session 2 in administrative records. The sleep index combines reported sleep quality, night waking, nighttime sleep duration, and the time between phone use and sleep or waking. Panel C reports social comparison. Panel D reports the reservation price for deactivating Instagram for two weeks, top-coded at INR 5,000. Estimates are plotted in control-group standard deviation units. Outcomes marked with $\times(-1)$ are reversed for plotting, and the right column reports the corresponding estimates in their original units. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Appendix Tables A10, A11, A12, A9, and A13. Appendix Table A3 reports FDR-adjusted q-values for the five pre-specified secondary outcomes.

Figure 6: Treatment effects on discernment



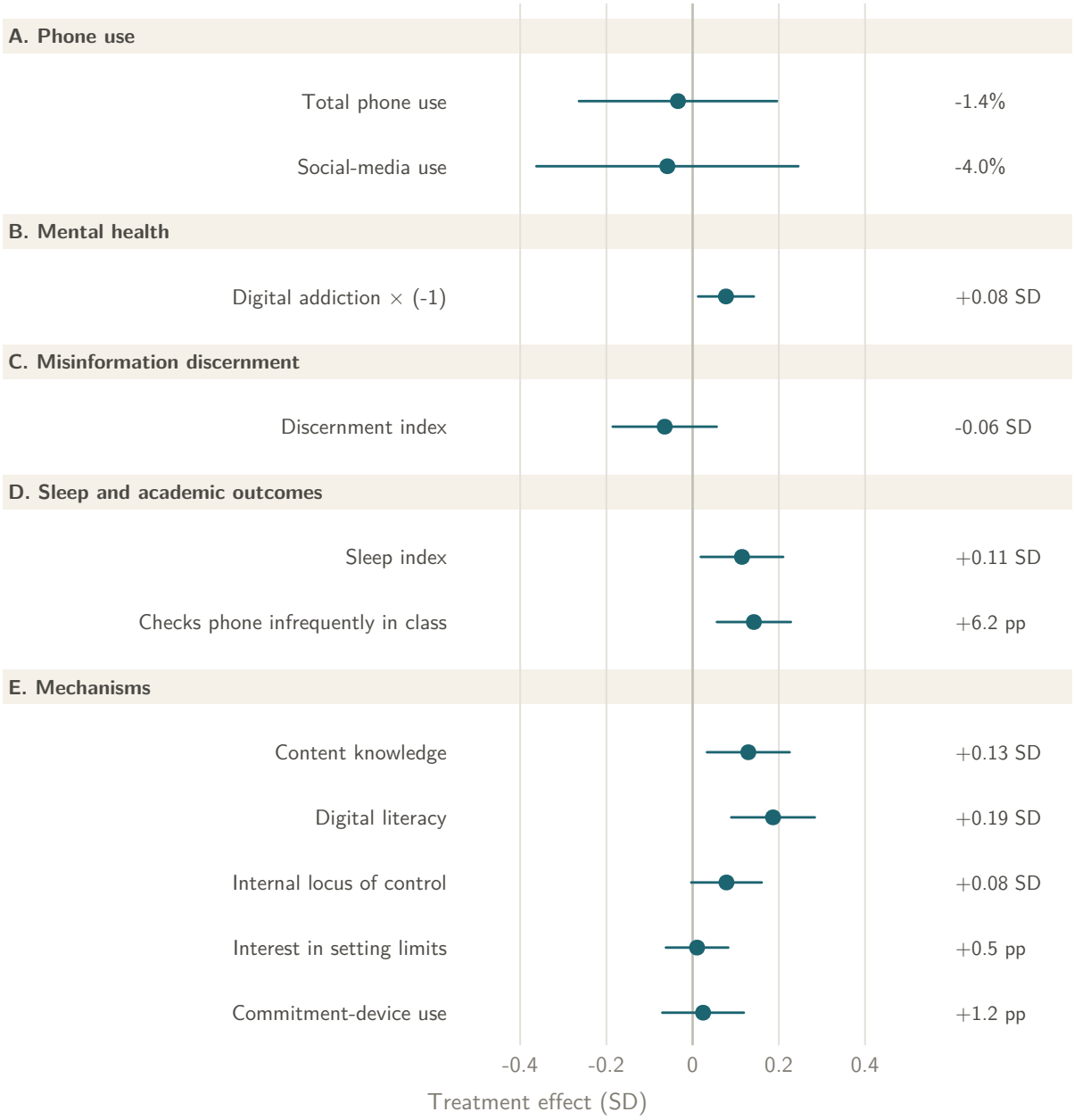
Notes: Panel A plots effects of *Deliberate Use & Discernment* on the discernment index, false rejection, and true recognition. The index combines predictions from separate IRT models, giving each of the two true and four false items in the endline battery equal weight. Panel B uses raw ratings of circulating claims and gives true and false claims equal total weight for each student. The true–false gap includes respondent, claim, and randomization stratum by truth fixed effects. The caution estimate omits respondent fixed effects. Panel C splits false rejection between circulating claims and novel false claims created for the survey. Models in Panels A and C control for both treatment indicators, randomization stratum fixed effects, gender, urban residence, mother's college education, household wealth, and age. Missing covariates are set to zero and flagged with indicators. Controls are centered and interacted with treatment. All estimates use weights inversely proportional to cohort size before the intervention. Effects are reported in endline control standard deviations. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Appendix Tables A14 and A15.

Figure 7: Treatment effects on knowledge, agency, and strategy adoption



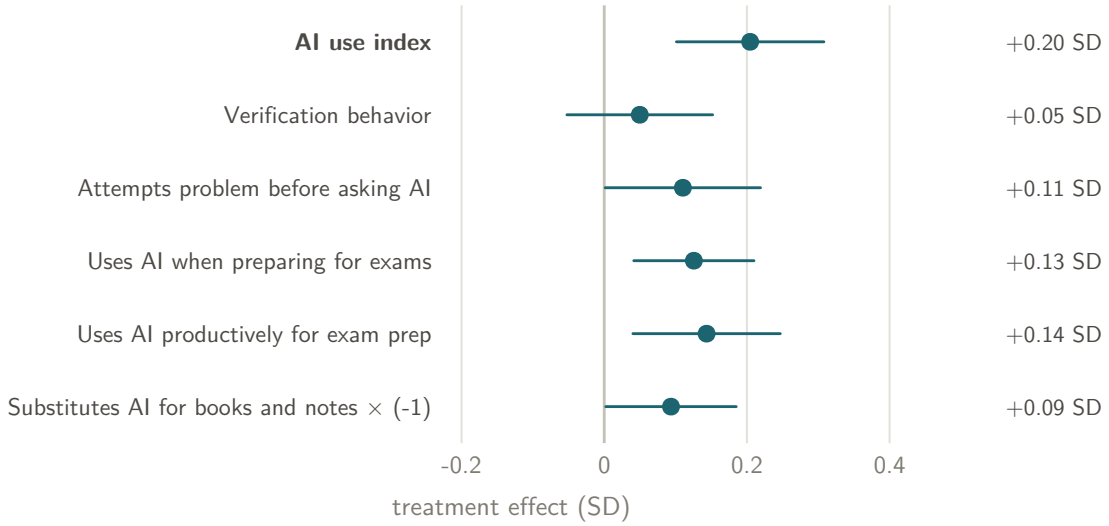
Notes: Panel A plots treatment effects on content knowledge and digital literacy, Panel B on internal locus of control, and Panel C on interest in setting limits and commitment device use. Content knowledge tests material taught in the curriculum. Digital literacy tests its application in six vignettes. Internal locus of control measures perceived control over outcomes. The final two measures capture interest in managing social media use and reported use of tools to limit it. Estimates are reported in control-group standard deviation units. Interest in setting limits and commitment device use are binary, and the right column reports these two effects in percentage points. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Appendix Table A26.

Figure 8: Persistence of classroom-treatment effects at follow-up



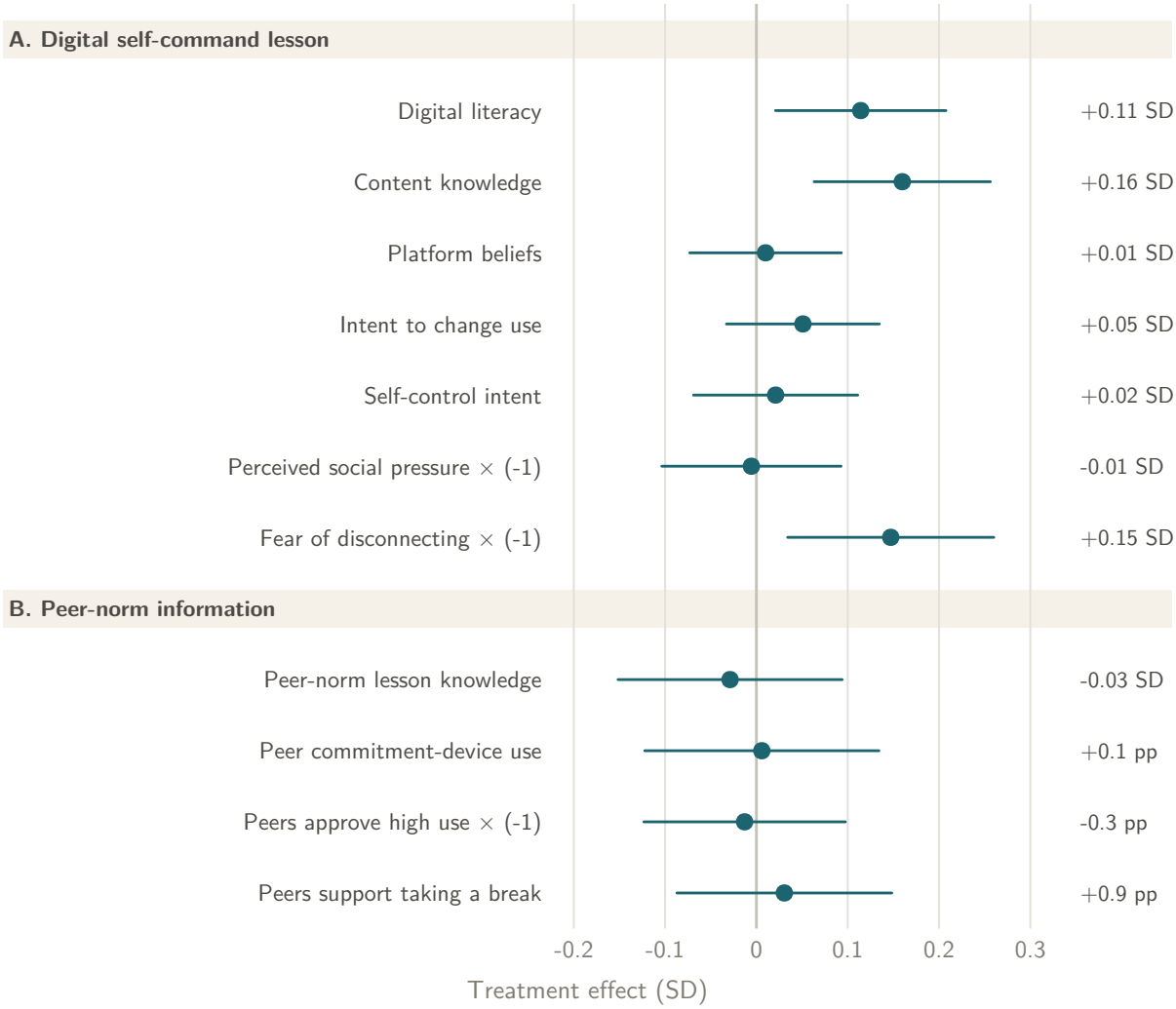
Notes: The figure plots effects of the classroom curriculum measured at follow-up. Panels A, B, D, and E plot the pooled effect of assignment to either treatment arm. Panel C plots the effect of *Deliberate Use & Discernment*, relative to control, on discernment, using the endline control standard deviation. Estimates are reported in control-group standard deviation units. Digital addiction is reversed for plotting, so that positive values indicate lower addiction. Reversed outcomes are marked with $\times(-1)$, and the right column reports the corresponding estimates in their original units. Outcomes measured after the follow-up microlearning module control for microlearning assignment. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Appendix Tables A27, A28, A29, A30, and A31.

Figure 9: Transfer effects on generative-AI use



Notes: The figure plots effects of the classroom curriculum on generative-AI use at follow-up among students who reported using AI in the past 30 days. Estimates are reported in control-group standard deviation units. Outcomes are oriented so positive values indicate more purposeful use. Reversed outcomes are marked with $\times(-1)$, and the right column reports the corresponding estimates. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Appendix Table A32.

Figure 10: Effects of survey-embedded microlearning and peer-norm information

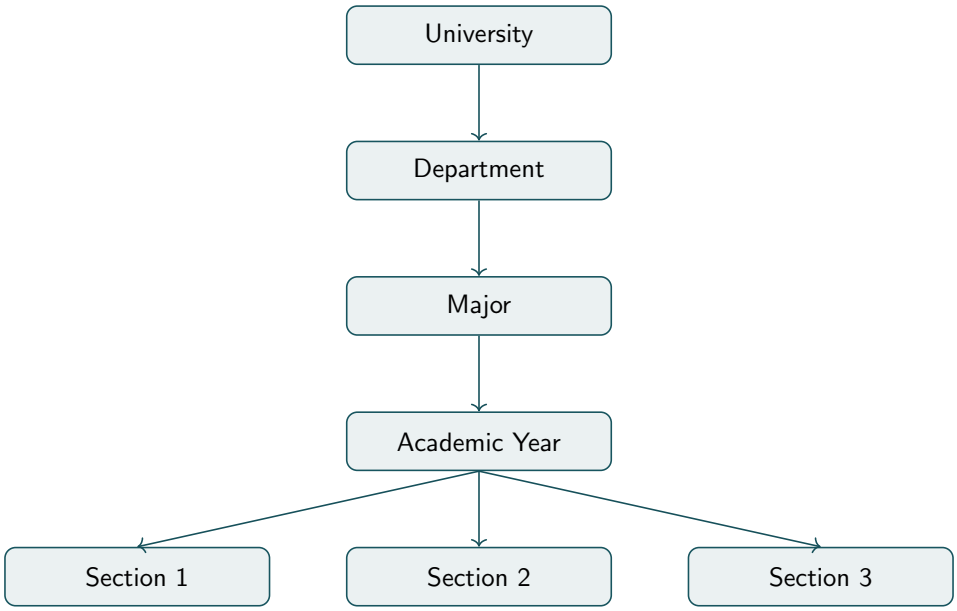


Notes: Panel A plots effects of the digital self-command lesson embedded in the survey, pooling versions with and without peer information, relative to the grammar lesson. Panel B plots effects of adding peer-norm information to that lesson. Estimates are reported in control-group standard deviation units. Outcomes are oriented so that positive values indicate improvement. Reversed outcomes are marked with $\times(-1)$, and the right column reports the corresponding estimates in their original units. Error bars report 95 percent confidence intervals clustered at the batch level. Panel A estimates correspond to Appendix Tables A28 and A36, and Panel B peer-norm estimates correspond to Appendix Table A37.

Appendices

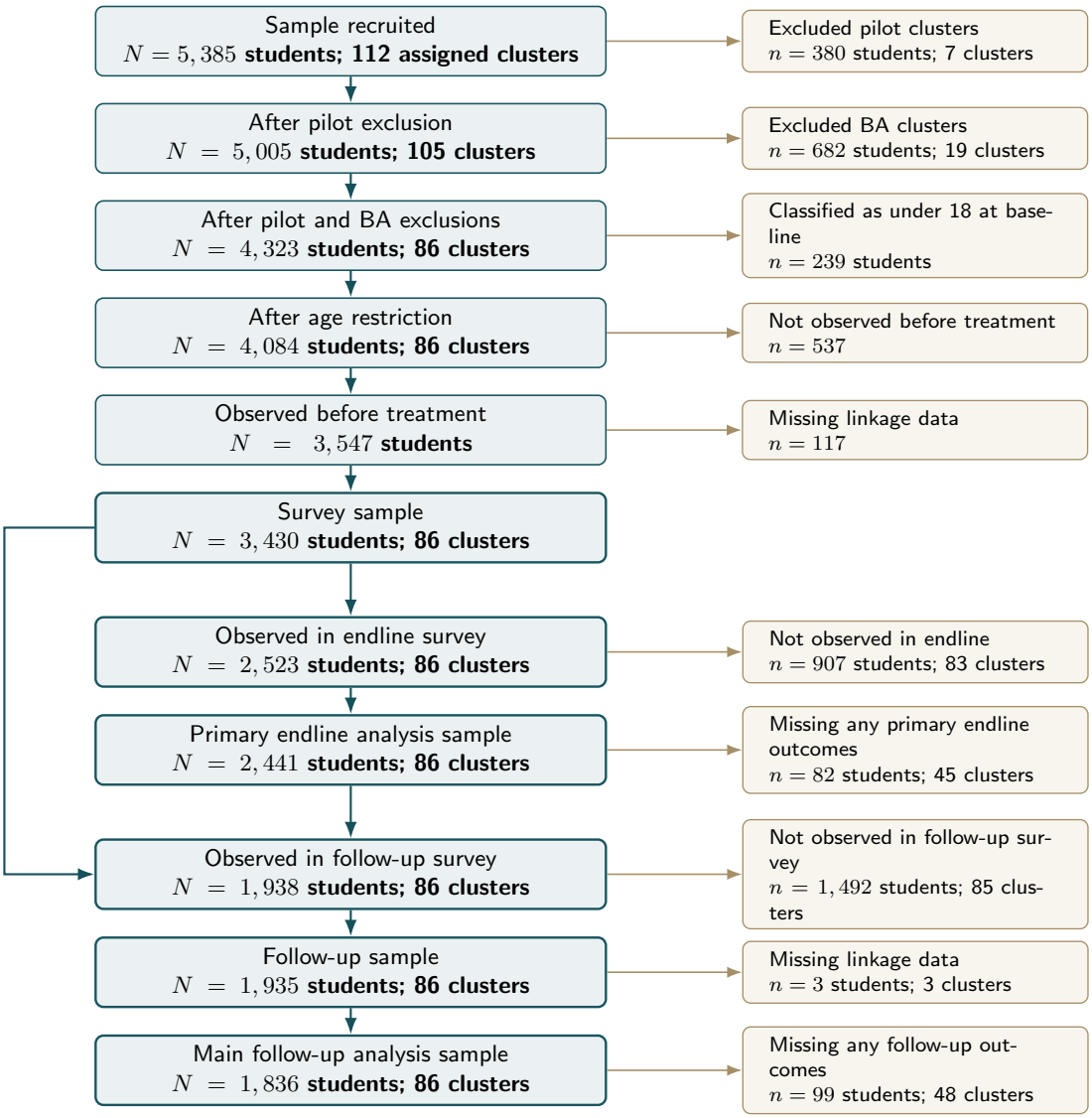
Appendix A Figures

Figure A1: Organizational structure of partner universities



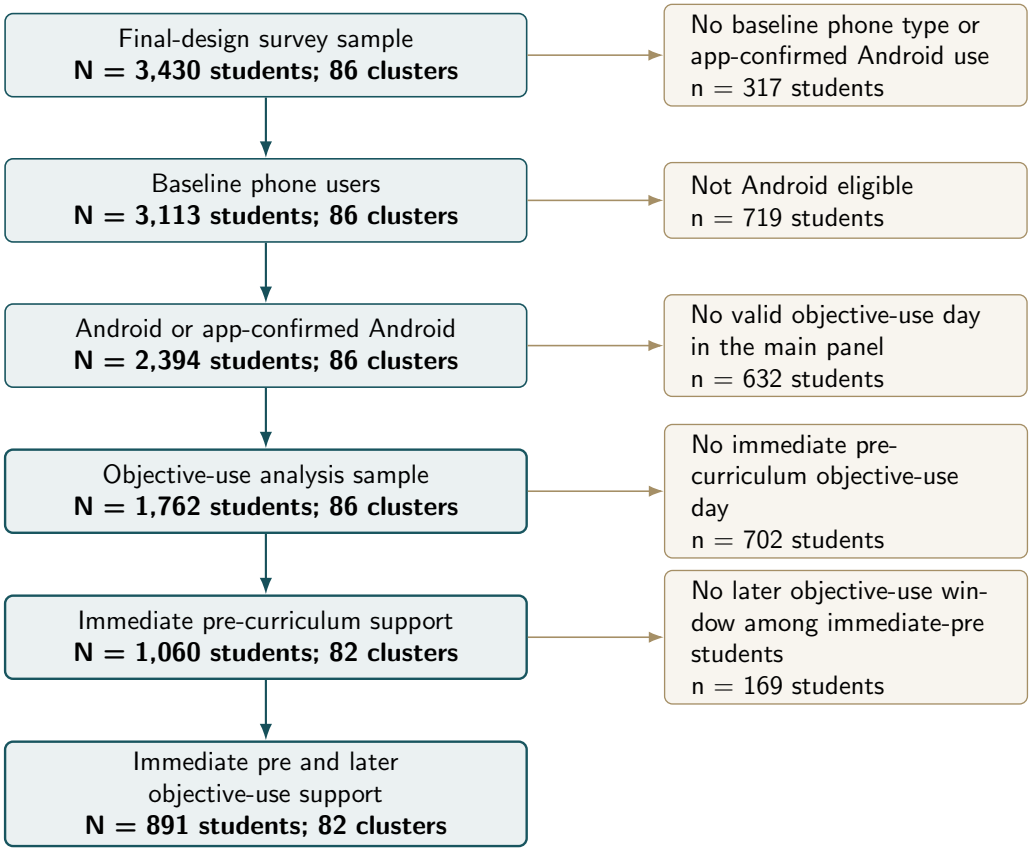
Notes: The figure summarizes the organizational structure of partner institutions. Each department within a college or university houses a major, such as physics or economics. Bachelor's degree programs typically span three years.

Figure A2: Trial sample flow through endline and follow-up survey analyses



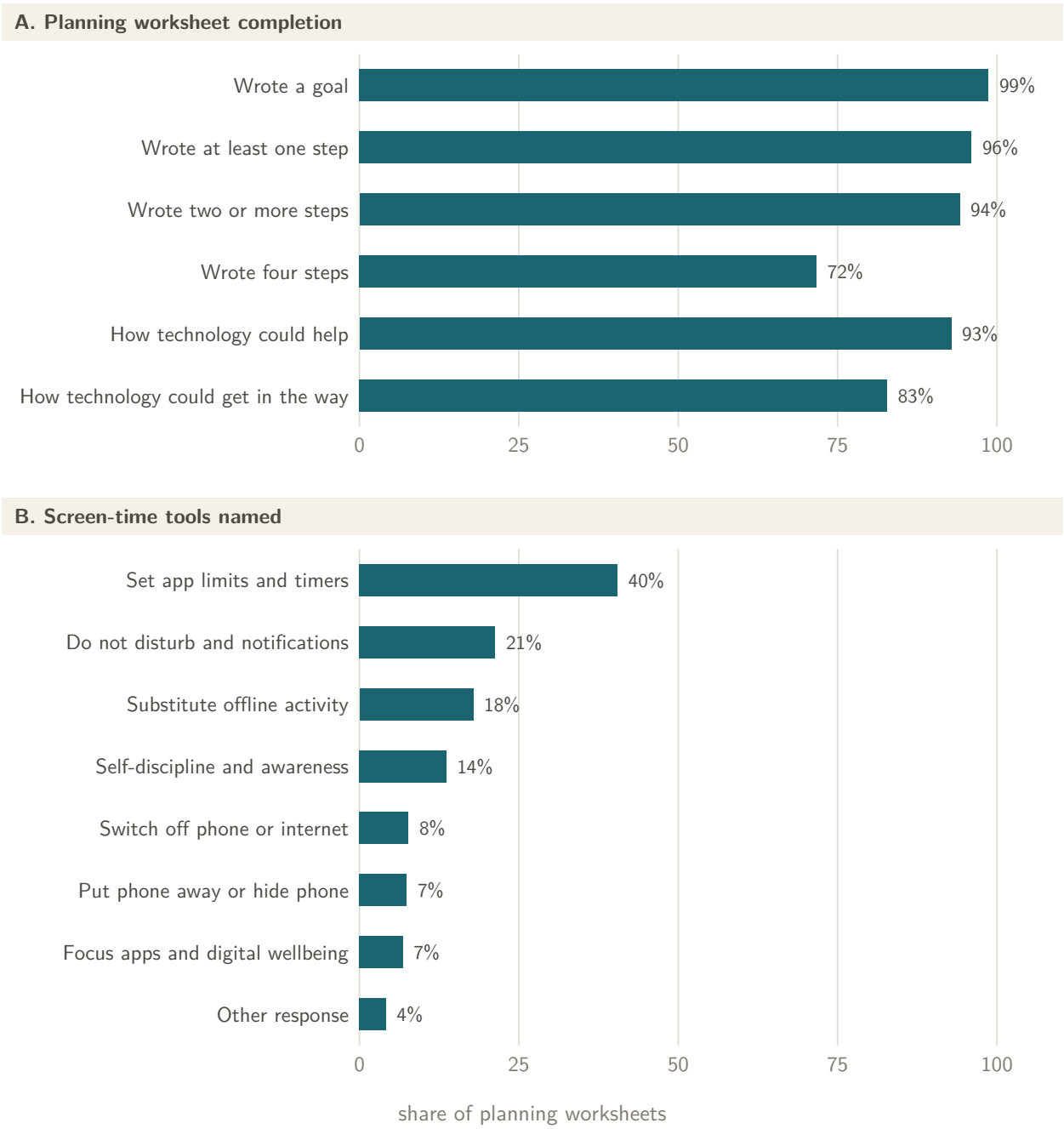
Notes: The figure summarizes how the analysis samples for the main endline and follow-up analyses are built from the survey data. Because this is a cluster-randomized trial, the figure reports both student counts and batch counts where relevant. The endline sample excludes pilot batches and batches with high contamination risk, and requires that a student be observed before treatment, have complete assignment and design information, complete the endline survey, and have non-missing values for the three primary endline outcome families: perceived social media use, mental health, and misinformation discernment. The follow-up branch starts from the same sample and then requires that a student be observed at follow-up, have a recorded microlearning assignment, and have non-missing values on the main follow-up outcomes. Some follow-up analyses use smaller samples because survey modules were shown to different arms and because some students skipped items within a module; these include the peer-information experiment and the AI-use and AI-beliefs modules.

Figure A3: Objective-use sample flow



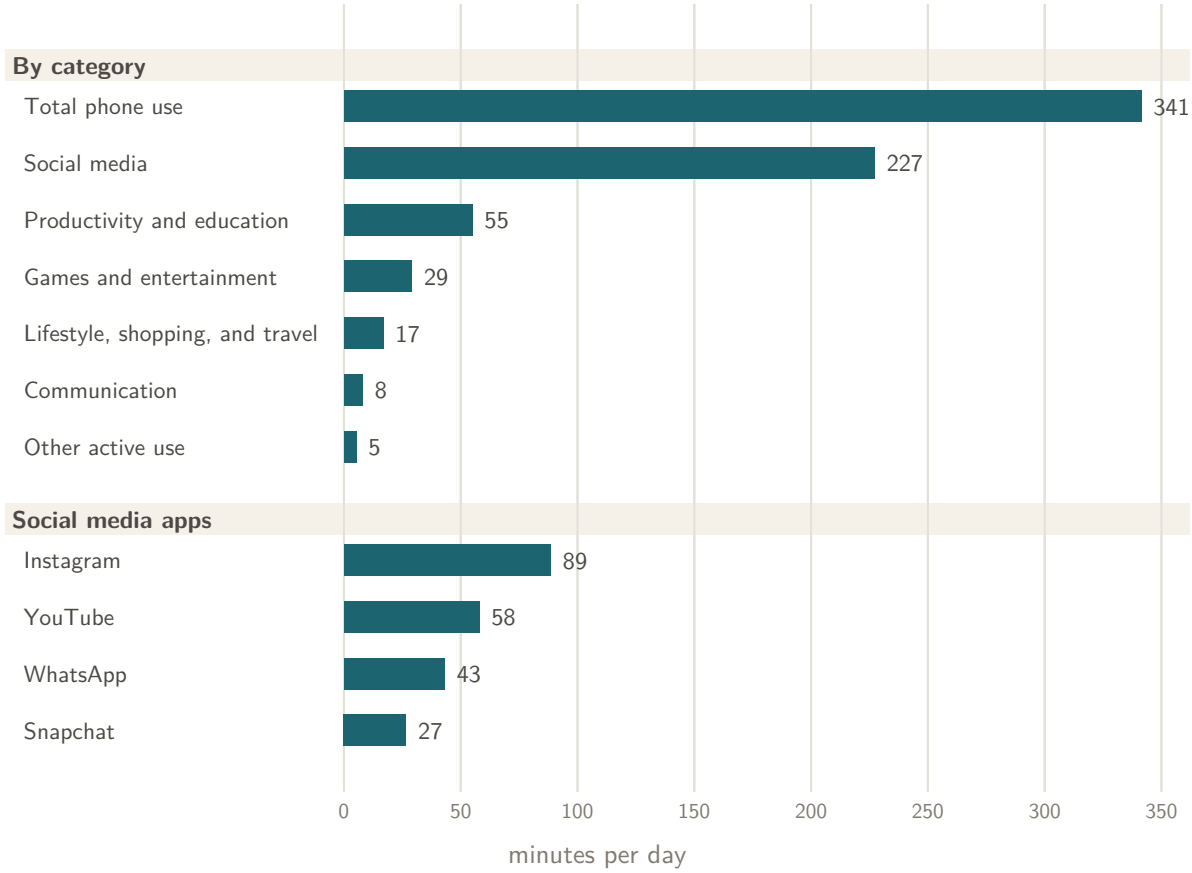
Notes: The figure summarizes how the objective-use sample is built from the survey analysis sample. Dashboard runs only on Android, so a student is eligible for objective-use measurement when Android ownership is reported in the survey or confirmed in the app data. The main objective-use panel uses valid recorded days in the reference, intervention-period, and endline measurement windows, after excluding pilot batches. A day is valid if it contains a complete record of active non-system phone use; partial and invalid days are dropped.

Figure A4: Worksheet engagement during the classroom curriculum



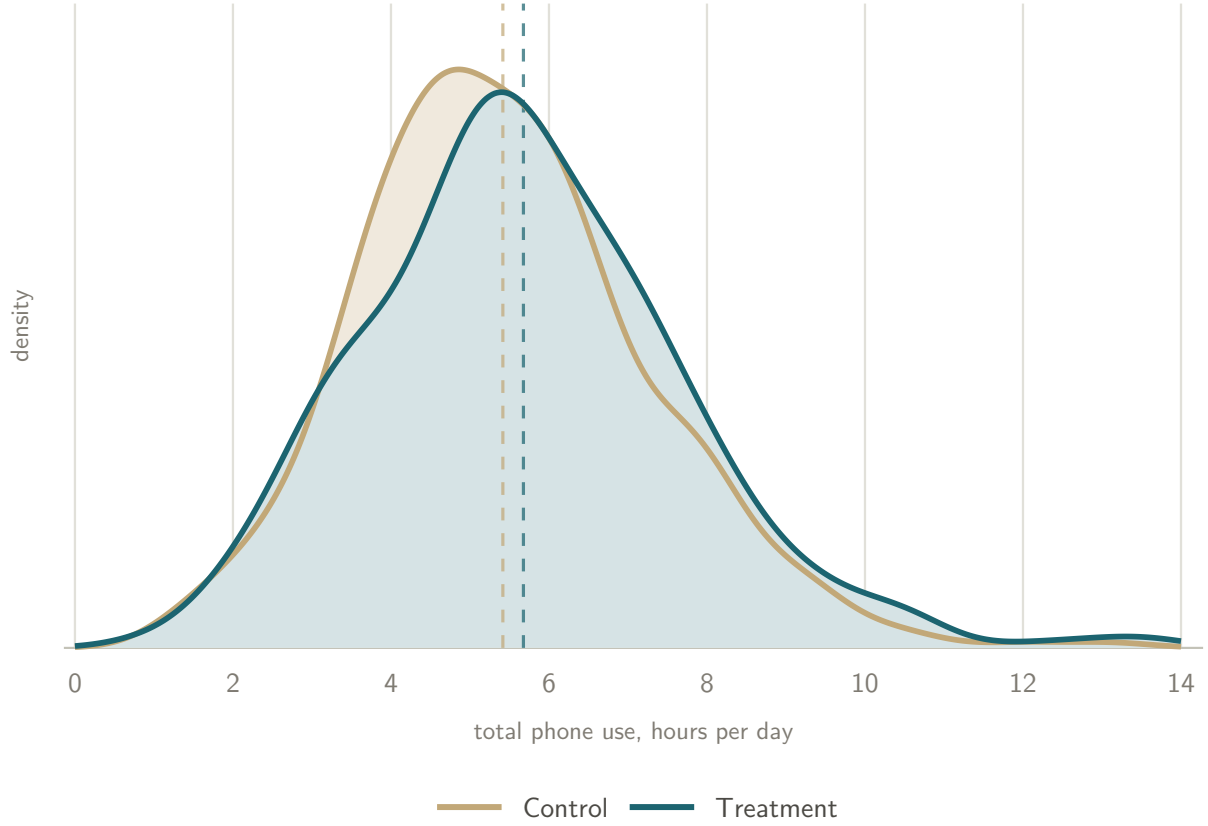
Notes: The figure summarizes descriptive measures from the Session 2 worksheets. Panel A reports the share of worksheets in which students wrote a goal, wrote concrete steps toward that goal, and reflected on how technology could help or get in the way. Panel B reports the tools students named when asked how they would lower time spent on apps they felt they used too much. A worksheet can name more than one tool.

Figure A5: Baseline objective phone use (in minutes) by category



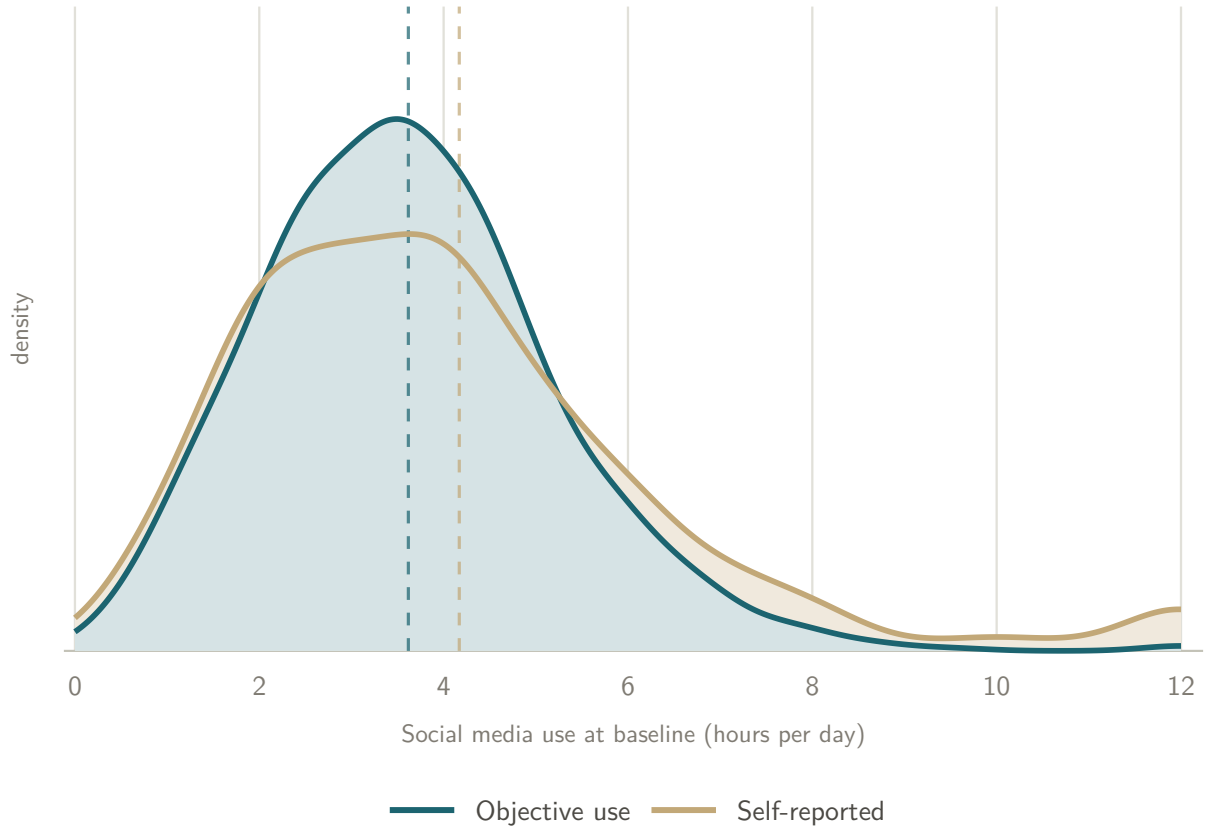
Notes: The figure plots mean daily minutes of total phone use over the three most recent valid days before the baseline survey. Total phone use is active non-system phone use. The category bars are mutually exclusive and sum to total phone use. The individual apps in the lower panel are the largest components of social media use.

Figure A6: Baseline distribution of objective total phone use



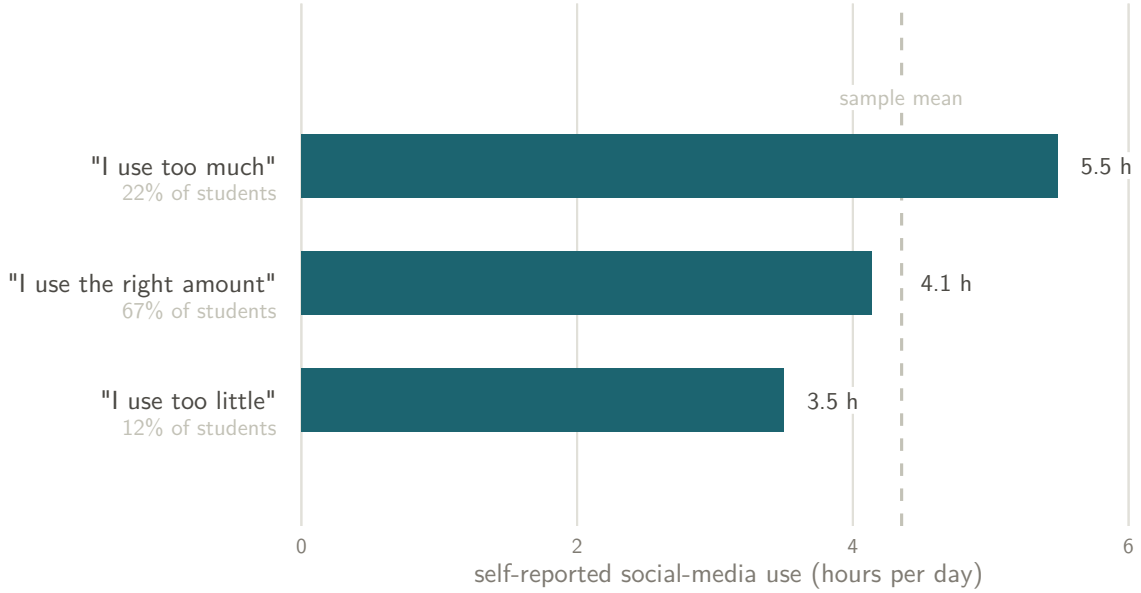
Notes: The figure plots kernel densities of average daily total phone use over the three most recent valid days before the baseline survey, excluding the survey day itself. Total phone use is active non-system phone use. Each student contributes one average, in hours per day. Dashed lines mark group means. Values above the displayed maximum are top-coded for display only.

Figure A7: Baseline social media use in self-reports and objective-use records



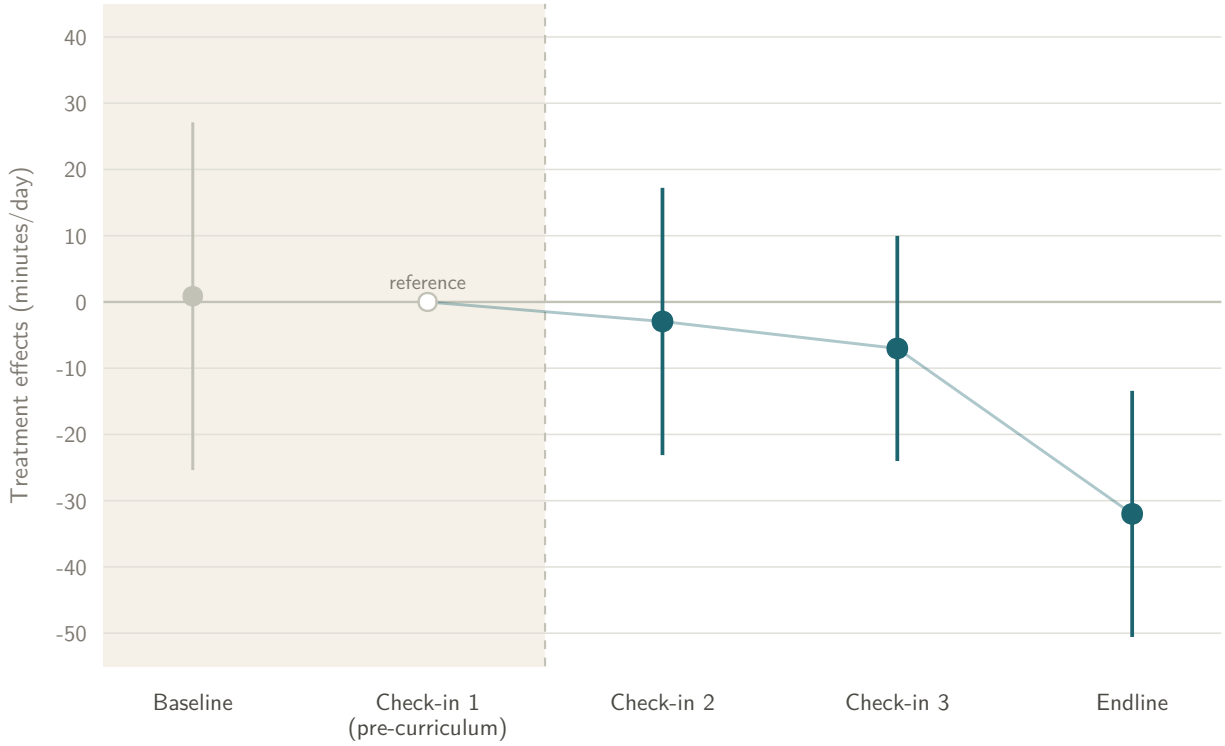
Notes: The figure compares average daily social media use in baseline self-reports with objective social media use, for students observed in both sources. The self-report asks students to think about the past week and report their average daily social media use on the phone. The objective measure uses up to the seven most recent valid days before the baseline survey, excluding the survey day itself. Dashed lines mark group means. Values above 12 hours per day are top-coded for display only.

Figure A8: Baseline self-assessment of social media use



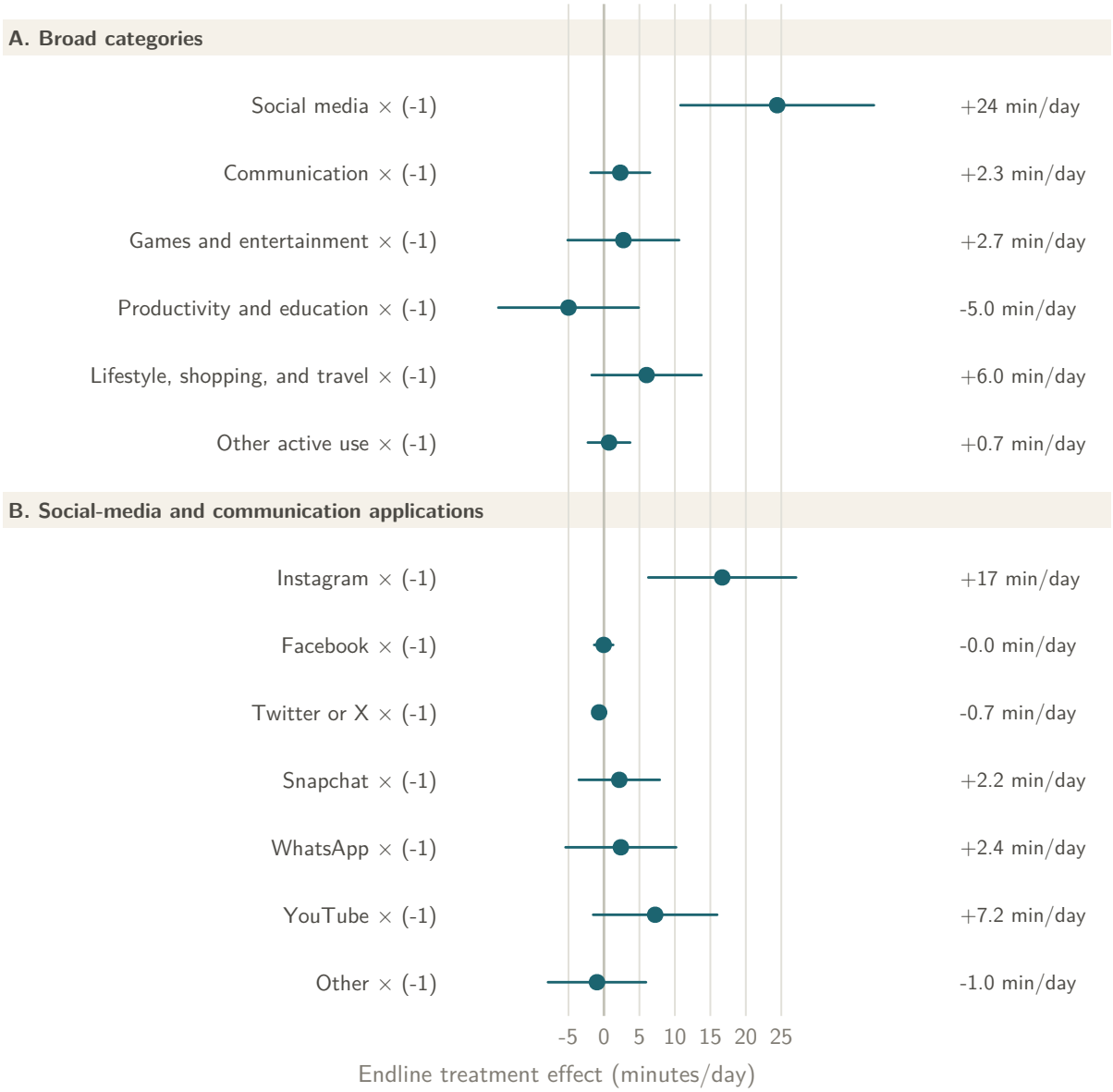
Notes: The figure plots mean self-reported daily social media use, in hours per day, by how students said they felt about their overall social media use at baseline. The left labels report the share of students giving each response. The dashed vertical line marks the sample mean.

Figure A9: Treatment effects on objective total phone use in minutes



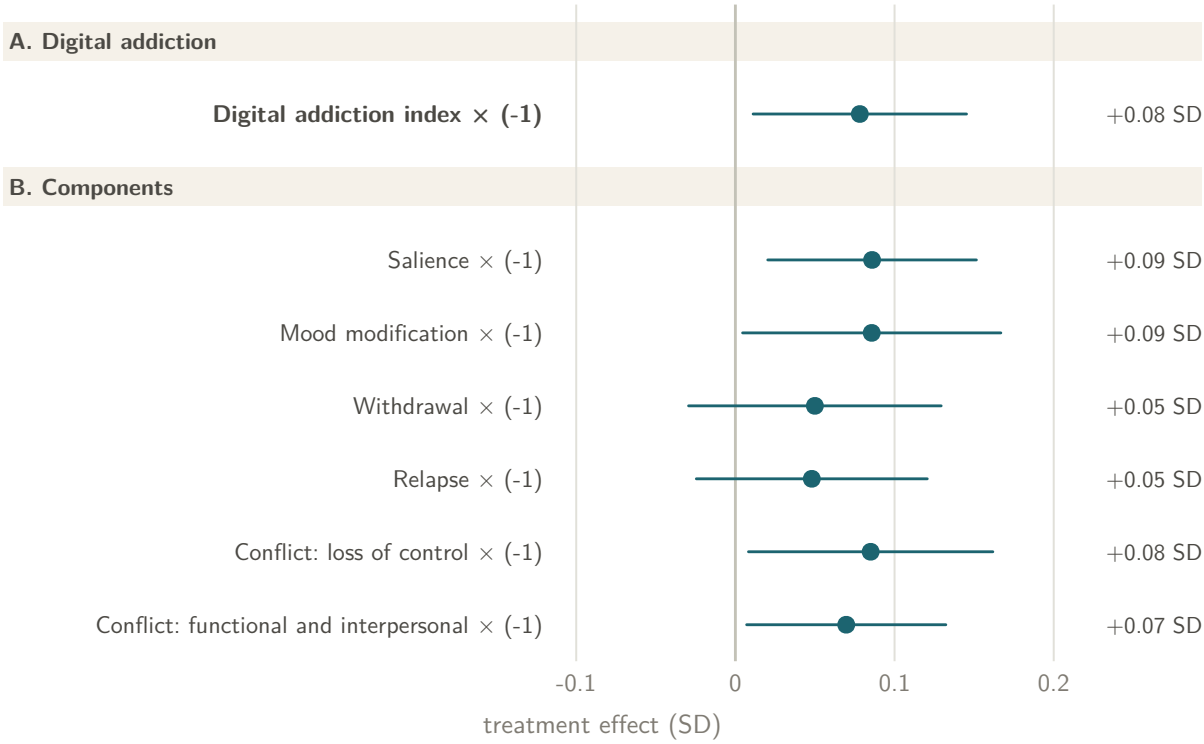
Notes: The figure plots how the treated–control difference in daily phone use, measured in minutes per day, changes relative to the pre-intervention period. The pre-intervention period is the three most recent valid days before check-in 1, the last measurement window before the first classroom session. Total phone use is active non-system phone use. The baseline estimate comes from a separate regression that uses every available observation in the baseline and pre-intervention-period windows, including students seen in only one of the two. Error bars report 95 percent confidence intervals clustered at the batch level. The estimates correspond to Panel B of Appendix Table A4.

Figure A10: Treatment effects on objective phone use by category and application



Notes: Panel A splits the endline effect into mutually exclusive application categories. Panel B splits social media and communication use into the main social media applications. Estimates are reported in minutes per day. Each outcome is reversed for plotting, so that reductions appear to the right of zero. Reversed outcomes are marked with $\times(-1)$, and the right column reports the corresponding estimates. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Appendix Table A5.

Figure A11: Treatment effects on digital addiction and its components



Notes: Panel A plots the treatment effect on the digital addiction index. Panel B plots effects on the components of the addiction scale. Estimates are reported in control-group standard deviation units. All outcomes are reversed for plotting, so that positive values indicate lower addiction. Reversed outcomes are marked with $\times(-1)$, and the right column reports the corresponding estimates. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Appendix Table A8.

Figure A12: Example posts from the misinformation discernment module

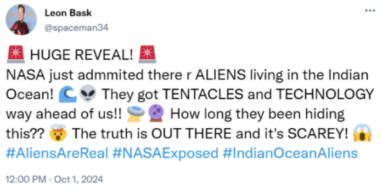
Panel A. True claim



Panel B. Circulated false claim

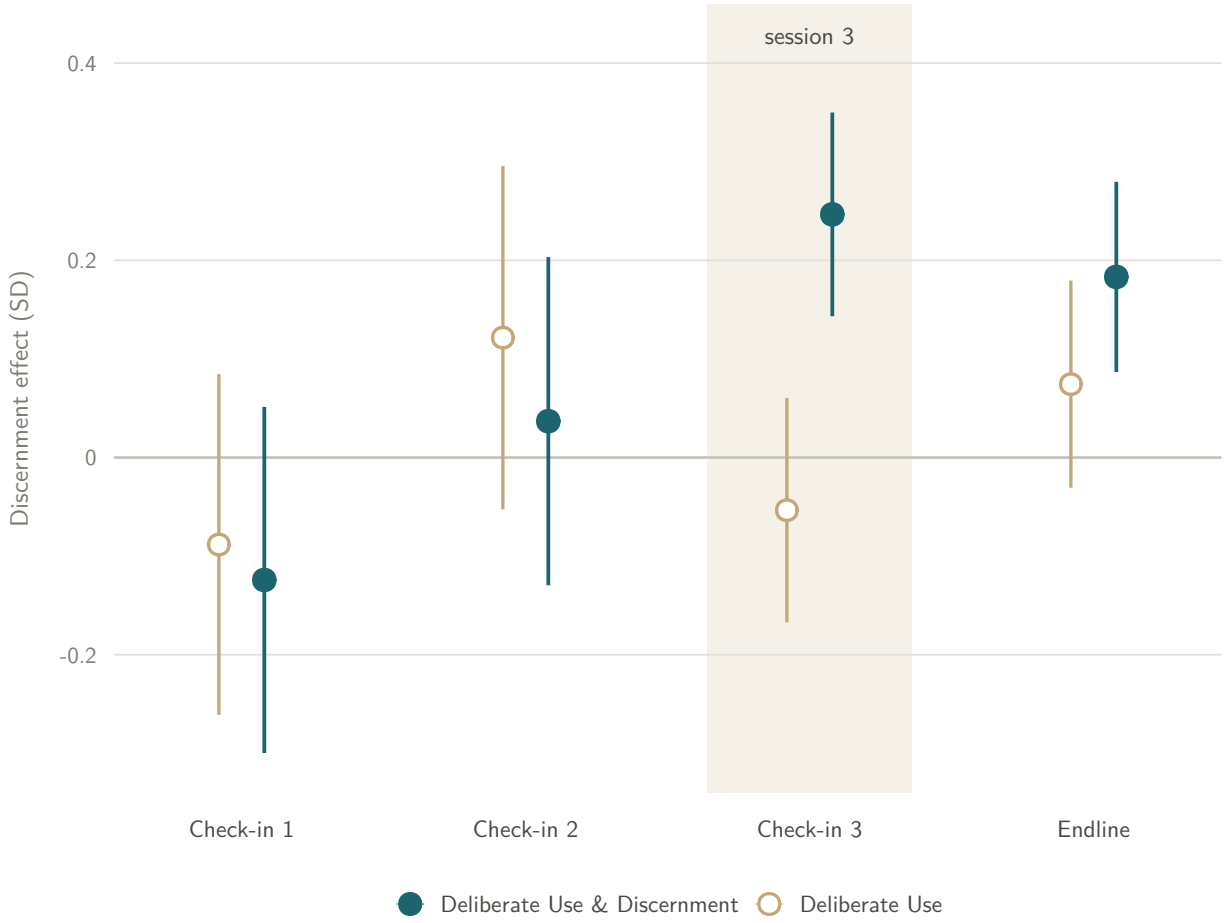


Panel C. Novel false claim



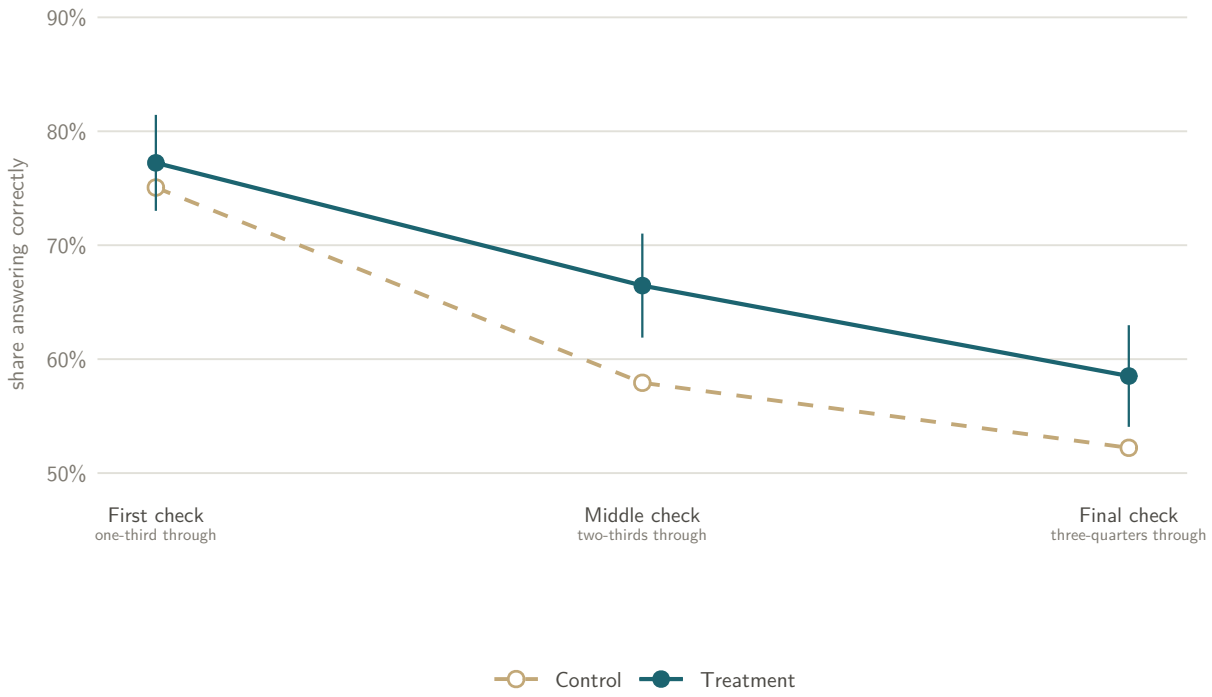
Notes: The figure shows one post of each claim type from the technology and science topic area. Panel A shows a true claim. Panel B shows a false claim that had circulated online. Panel C shows a novel false claim created by the research team for the survey. For each post, students answered "To the best of your knowledge, how accurate is the claim in the above post?" on a five-point scale from "not at all accurate" to "extremely accurate."

Figure A13: Discernment over the intervention



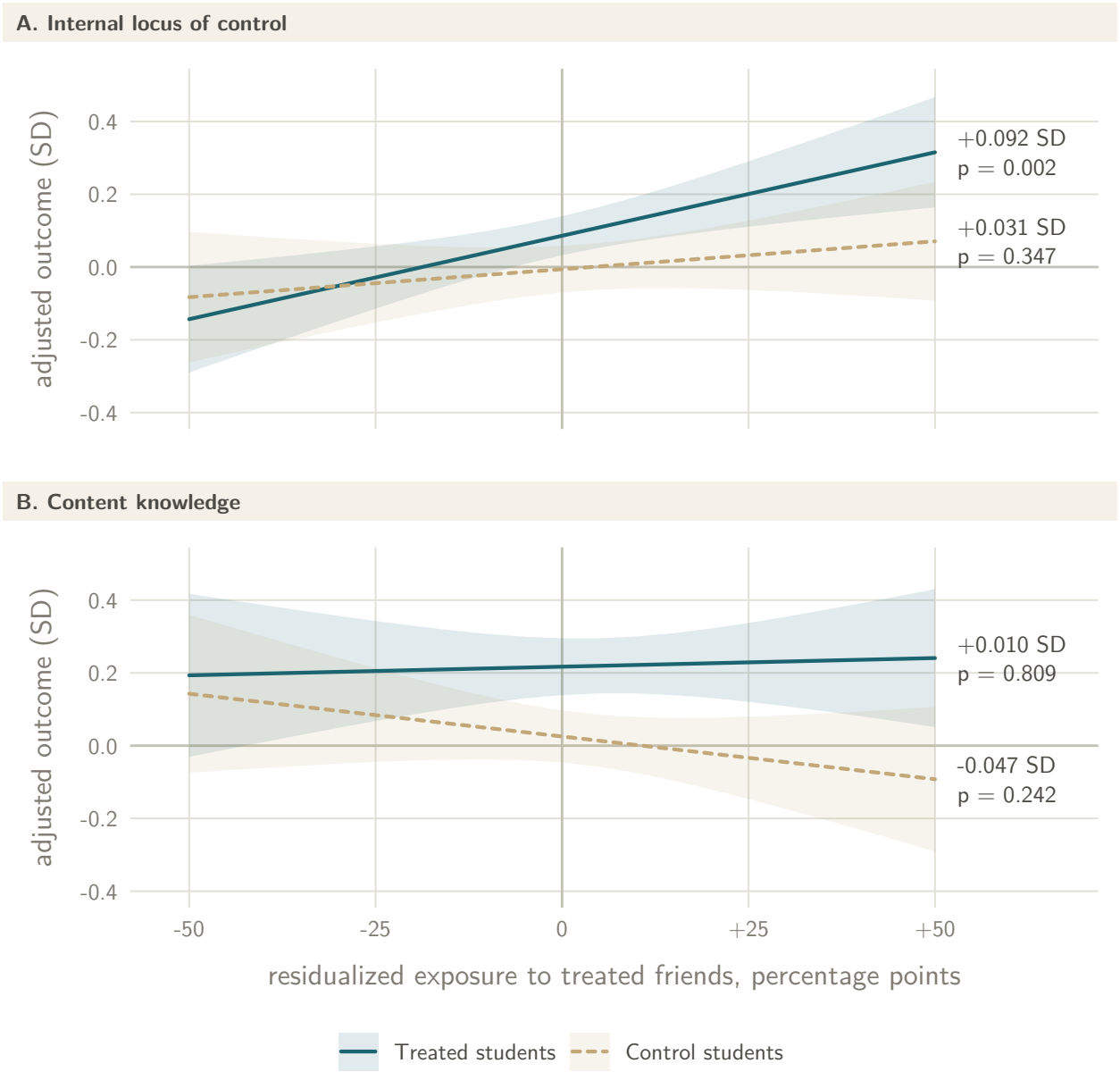
Notes: The figure plots treatment effects on the discernment index at the three check-ins and endline. The solid points report the *Deliberate Use & Discernment* arm and the hollow points report the *Deliberate Use* arm, each relative to control. The shaded band marks Session 3, delivered on the check-in 3 survey day. Separate IRT models for true recognition and false rejection use the same relative weights at every survey, matching the two true and four false items in the endline battery. Models include both treatment indicators, randomization stratum fixed effects, gender, urban residence, mother's college education, household wealth, and age. Missing covariates are set to zero and flagged with indicators. Controls are centered and interacted with treatment. No baseline misinformation measure is included. Estimates use weights inversely proportional to cohort size before the intervention and are reported in endline control standard deviations. March 24 and 25, 2025 endline misinformation responses are excluded. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Appendix Table A17.

Figure A14: Treatment effects on task-based attention at follow-up



Notes: The figure plots pass rates for three attention checks placed at different points in the follow-up survey. The dashed line reports the control-group mean and points report that mean plus the residualized treatment effect. Error bars report 95 percent confidence intervals clustered at the batch level. Estimates correspond to Appendix Table A30.

Figure A15: Peer-exposure gradients in internal locus of control and content knowledge



Notes: The figure plots fitted values of internal locus of control and content knowledge against exposure to treated friends, separately for treated and control students. Exposure uses friends nominated before any classroom content was delivered, who could be matched to students in other batches. It is the share of those friends assigned to treatment, minus the share expected by chance under the randomization. Shaded areas report 95 percent confidence intervals clustered at the batch level. Annotations report the fitted slope per one standard deviation of exposure. Estimates correspond to Appendix Table A24.

Appendix B Tables

Table A1: Survey participation across waves

Survey wave	Control (1)	Treatment (2)	Adjusted difference (3)	SE (4)	p-value (5)
Check-in 1	0.616	0.611	0.005	0.023	0.837
Check-in 2	0.476	0.520	0.034	0.032	0.292
Check-in 3	0.513	0.553	0.021	0.031	0.505
Endline	0.755	0.750	-0.005	0.023	0.821
Follow-up	0.594	0.570	-0.017	0.025	0.496

Notes: The unit of observation is a student. Control and treatment columns report weighted shares completing each survey wave. Every row uses the same 3,430 students in the 86 randomized batches who were observed before treatment. The difference column reports the treatment-control gap after adjusting for randomization stratum fixed effects and weights inversely proportional to cohort size before the intervention. Standard errors are clustered at the batch level.

Table A2: Objective-use measurement and pre-period diagnostics

	Control mean (1)	Difference (2)	SE (3)	p-value (4)	Observations (5)
Panel A: Measurement availability (share)					
Baseline	0.432	-0.002	0.036	0.951	2,395
Check-in 1	0.408	0.084	0.033	0.012	2,395
Check-in 2	0.372	0.009	0.035	0.788	2,395
Check-in 3	0.426	0.069	0.035	0.054	2,395
Endline	0.460	0.064	0.028	0.023	2,395
Pooled post-content windows	0.419	0.048	0.026	0.069	7,185
Endline minus Check-in 1	–	-0.020	0.037	0.584	4,790
Pooled post-content minus Check-in 1	–	-0.037	0.035	0.290	9,580
Panel B: Entry and retention (share)					
Baseline contributors observed at check-in 1	0.698	0.018	0.059	0.766	668
Students without baseline data entering by Check-in 1	0.525	0.146	0.048	0.003	804
Baseline contributors observed at endline	0.695	0.044	0.036	0.230	844
Change in baseline-contributor availability gap	–	0.006	0.072	0.931	1,156
Panel C: Total phone use before the curriculum (percent)					
Baseline	–	3.991	2.945	0.172	1,050
Check-in 1	–	4.000	3.863	0.294	1,060
Change from baseline to check-in 1	–	0.542	4.397	0.902	1,608

Notes: The unit of observation is a student and survey in Panels A and B and a student-day in Panel C. Panels A and B report shares. Panel A reports treatment-control differences in data availability. The pooled row combines check-ins 2 and 3 and endline. Panel B distinguishes baseline contributors from later entrants. Panel C reports percentage differences in total phone use from log-transformed outcomes. Both pre-intervention windows use up to three valid days. Regressions include randomization-stratum fixed effects and inverse pre-intervention cohort-size weights. Standard errors are clustered at the batch level.

Table A3: Sharpened FDR q-values for secondary outcomes

Outcome	Effect (1)	p-value (2)	Sharpened q-value (3)
Commitment device use (pp)	4.573	0.016	0.086
	(1.855)		
Fail or reappear (pp)	-4.992	0.033	0.086
	(2.294)		
Social media use during class (min/hour)	-0.953	0.160	0.154
	(0.673)		
Sleep index (SD)	0.051	0.178	0.154
	(0.037)		
Social comparison index (SD)	-0.039	0.222	0.154
	(0.032)		

Notes: The table reports endline estimates and the first regular exam after Session 2 for the academic outcome. Sharpened FDR q-values are adjusted jointly across the five secondary outcomes. Standard errors clustered at the batch level are in parentheses. The unit of observation is a student, except for classroom use, which uses student-days. pp denotes percentage points. SD denotes control group standard deviations.

Table A4: Treatment effects on objective phone use

	Total phone use (1)	Social media (2)	Non-social phone use (3)
Panel A: Proportional effects (percent)			
Treatment × Pooled post	-5.165*	-7.743**	1.006
	(2.667)	(3.586)	(5.542)
Treatment × Check-in 2	-3.437	-8.391	5.074
	(4.156)	(5.305)	(7.774)
Treatment × Check-in 3	-2.815	-4.179	5.046
	(2.927)	(4.112)	(6.243)
Treatment × Endline	-8.878***	-9.964**	-6.321
	(2.986)	(3.711)	(6.621)
Panel B: Effects in minutes per day			
Treatment × Pooled post	-14.097*	-12.389*	-1.707
	(7.231)	(6.395)	(6.110)
Treatment × Check-in 2	-2.948	-3.633	0.685
	(10.145)	(9.462)	(7.345)
Treatment × Check-in 3	-7.019	-8.709	1.690
	(8.546)	(7.508)	(6.732)
Treatment × Endline	-31.996***	-24.865***	-7.130
	(9.343)	(6.896)	(8.056)
Control mean	330.389	217.073	113.315
Observations	11,627	11,627	11,627

Notes: The unit of observation is a student-day. Each coefficient is a difference-in-differences ITT estimate relative to check-in 1. All survey periods use the three most recent valid days. Panel A reports percentage changes from regressions with log-transformed outcomes. Panel B reports effects in minutes per day. Control means are measured at check-in 1 and reported in minutes per day. Regressions include fixed effects for randomization stratum by calendar week, survey period, relative day, day of week, and student. They use weights inversely proportional to cohort size before the intervention and to the number of observed days. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A5: Treatment effects on the composition of objective phone use

	Treatment × Pooled post (1)	Treatment × Check-in 2 (2)	Treatment × Check-in 3 (3)	Treatment × Endline (4)	Control mean (5)
Panel A: Broad app categories					
Social media	-11.790*	-2.242	-8.841	-24.423***	216.918
	(6.407)	(9.407)	(7.564)	(6.957)	
Communication	-2.186	-3.118**	-0.870	-2.305	7.072
	(1.500)	(1.321)	(1.462)	(2.136)	
Games and entertainment	-0.861	-0.595	1.035	-2.741	31.365
	(3.455)	(3.518)	(4.539)	(4.007)	
Productivity and education	4.685	8.519*	-0.515	4.995	52.327
	(3.479)	(4.838)	(4.063)	(5.046)	
Lifestyle, shopping, and travel	-3.324	-3.151	-0.353	-6.012	16.145
	(3.006)	(2.863)	(2.744)	(3.948)	
Other active use	0.520	0.162	2.433*	-0.713	6.445
	(1.071)	(1.433)	(1.277)	(1.521)	
Panel B: Social-media and communication apps					
Instagram	-13.174***	-11.799**	-10.759***	-16.670***	87.175
	(3.848)	(5.807)	(3.773)	(5.314)	
Facebook	-0.460	0.241	-1.935***	0.038	3.186
	(0.515)	(0.705)	(0.679)	(0.684)	
Twitter or X	0.438*	0.182	0.467**	0.687**	0.596
	(0.227)	(0.188)	(0.214)	(0.329)	
Snapchat	0.292	2.196	0.803	-2.163	22.433
	(1.791)	(1.915)	(2.255)	(2.910)	
WhatsApp	0.209	3.166	-0.445	-2.382	35.600
	(3.703)	(4.244)	(4.335)	(3.977)	
YouTube	-1.463	2.088	0.890	-7.220	61.803
	(4.047)	(5.582)	(4.724)	(4.460)	
Other social and communication apps	0.182	-1.433	1.268	0.982	13.197
	(2.564)	(2.145)	(2.752)	(3.523)	
Observations	11,616				

Notes: Each observation is a category or app total for one student on one day. Each cell is a difference-in-differences ITT estimate relative to check-in 1. All survey periods use the three most recent valid days, and outcomes are measured in minutes per day. Category and app rows are estimated jointly. Regressions include fixed effects for randomization stratum by calendar week, survey period, relative day, day of week, category or app, and student. They use weights inversely proportional to cohort size before the intervention and to the number of observed days. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A6: Treatment effects on phone pickups and session length

	Phone pickups	Session length	
	Pickups per day	Brief sessions per day	Long sessions per day
	(1)	(2)	(3)
Treatment × Endline	5.254***	-3.267	-0.285**
	(1.912)	(5.586)	(0.110)
Control mean	99.474	110.436	0.869
Observations	10,683	10,683	10,683

Notes: The unit of observation is a student-day. The reported row is the endline difference-in-differences ITT estimate relative to check-in 1. All survey periods use the three most recent valid days. A phone pickup is a screen-on event following at least 30 seconds with the screen off. Outcomes use the preferred day-level panel specification with timing and student fixed effects, weights inversely proportional to cohort size before the intervention and to the number of observed days, and standard errors clustered at the batch level. Brief sessions last at least one second and less than two seconds. Long sessions last more than 30 minutes. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A7: Percent effects on total phone use at endline by day type

	Monday–Friday	Saturday–Sunday	Weekend minus weekday
	(1)	(2)	(3)
Treatment × Endline	-9.968***	-3.929	6.038
	(3.254)	(3.904)	(4.133)
Control mean	314.465	356.677	42.212
Observations	24,992	24,992	24,992

Notes: The unit of observation is a student-day. Entries are percentage changes from regressions with log-transformed outcomes. Column (3) is the Saturday–Sunday effect minus the Monday–Friday effect. Its control mean is the corresponding difference in minutes per day at check-in 1. This analysis uses up to seven days before each survey so each window can represent a full calendar week. Regressions include fixed effects for student, randomization stratum by calendar week, survey period, relative day, and day of week. These fixed effects are interacted with day type where appropriate. Regressions use weights inversely proportional to cohort size before the intervention and to the number of observed days. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A8: Treatment effects on digital addiction and its components

	Digital addiction	Components					
	Digital addiction	Salience	Mood modification	Withdrawal	Relapse	Loss of control	Functional/interpersonal conflict
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	-0.078**	-0.086**	-0.086**	-0.050	-0.048	-0.085**	-0.070**
	(0.034)	(0.033)	(0.041)	(0.040)	(0.037)	(0.039)	(0.032)
Control mean	0.025	0.021	0.056	0.033	0.053	0.049	0.032
Observations	2,441	2,441	2,441	2,441	2,441	2,441	2,441

Notes: The unit of observation is a student. The component columns decompose the 16-item digital-addiction battery using the components model. Tolerance is not measured, and conflict is shown in two parts. All outcomes are standardized and oriented so lower values indicate less addiction. Regressions include randomization-stratum fixed effects, weights inversely proportional to cohort size before the intervention, and baseline values when available. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A9: Treatment effects on wellbeing and social outcomes

	Mental health and wellbeing						Social outcomes
	Mental health index	Social connectedness	Happiness	Life satisfaction	Anxiety symptoms	Depression symptoms	Social comparison index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	-0.066*	0.118***	0.009	-0.013	-0.018	-0.028	-0.039
	(0.037)	(0.039)	(0.039)	(0.039)	(0.041)	(0.044)	(0.032)
Adjusted q-value		0.018	1.000	1.000	1.000	1.000	
Control mean	0.016	-0.026	-0.011	0.002	0.016	-0.008	-0.049
Observations	2,432	2,432	2,432	2,432	2,432	2,432	2,432

Notes: The unit of observation is a student. The mental health index applies Anderson weights to three standardized domains, digital addiction, combined anxiety and depression symptoms, and subjective wellbeing. Higher values indicate worse mental health. Outcomes are standardized on the control group. Higher anxiety and depression scores indicate more symptoms. Higher happiness and life-satisfaction scores indicate worse wellbeing. Higher social connectedness indicates improvement. Regressions include randomization-stratum fixed effects and weights inversely proportional to cohort size before the intervention. Adjusted q-values are Anderson sharpened q-values that control the false discovery rate across the five separately reported wellbeing outcomes. The mental health index and the separately reported social-comparison index are not included in this component-family adjustment. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A10: Treatment effects on daily time allocation

	Screen time (1)	Study time (2)	Offline activities (3)	Sleep (4)
Treatment	-0.142	-0.078	-0.075	0.294***
	(0.095)	(0.088)	(0.073)	(0.102)
Control mean	4.731	8.842	3.907	6.520
Observations	2459	2459	2459	2459

Notes: Outcomes are self-reported hours on a typical day from the endline activity-allocation battery. Screen time combines social media and other screen time; study time combines studying alone, studying with others, and class time; offline activities combine offline solo activities and in-person activities with others. The four columns partition a 24-hour day by construction and are estimated on a common sample with complete responses to all eight underlying time-use items. To preserve the full endline activity-module sample, specifications do not control for baseline values of these items, which were available only for a smaller subset. Regressions include stratum fixed effects and inverse cohort-size weights. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. Standard errors are clustered at the cohort level. ***, **, and * denote significance at 1%, 5%, and 10% levels respectively.

Table A11: Treatment effects on sleep quality and sleep hygiene

	Overall	Subindices		Components				
	Sleep index	Sleep-quality index	Sleep-hygiene index	Poor sleep	Night waking	Duration (hours)	Before sleep (minutes)	After waking (minutes)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment	0.051	0.058	-0.002	-0.022	-0.011	0.017	0.206	1.071
	(0.037)	(0.038)	(0.031)	(0.014)	(0.017)	(0.042)	(1.040)	(1.806)
Adjusted q-value				1.000	1.000	1.000	1.000	1.000
Control mean	0.003	0.000	0.003	0.137	0.377	7.700	23.549	27.720
Observations	2,430	2,430	2,430	2,430	2,430	2,430	2,430	2,430

Notes: The unit of observation is a student. The overall index combines the standardized sleep-quality and sleep-hygiene subindices. Sleep quality uses the first three component columns. Sleep hygiene uses the final two. Components are shown in original units. The indices are standardized and oriented so higher values indicate improvement. Poor sleep and night waking are indicators. Adjusted q-values are Anderson sharpened q-values that control the false discovery rate across the five displayed component outcomes. The omnibus index and its subindices are not included in this component-family adjustment. Regressions include randomization-stratum fixed effects, weights inversely proportional to cohort size before the intervention, and baseline values when available. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A12: Treatment effects on administrative academic outcomes

	Failure or reappearance (pp) (1)	Numerical grades (SD) (2)
Treatment	-4.992**	0.003
	(2.294)	(0.043)
Control mean	37.855	0.000
Observations	2,633	2,386
Batches	74	69
Baseline failure status	Yes	Yes
Baseline numerical grades	No	Yes

Notes: The unit of observation is a student. Both columns use the first scheduled regular semester exam after Session 2. Makeup and reappear sittings are excluded. Official corrections to the regular result are retained. Column (1) reports effects in percentage points. Absence counts as failure where institutional rules require it. Column (2) combines semester GPA and percentage marks, standardized separately using fixed untreated reference distributions for the same program or a broader reference within the same university. Failed and absent courses carry zero grade points under the university rules. Cumulative degree marks are excluded. The samples differ because some records report status without a usable numerical grade. Missing outcome grades are not imputed. All regressions include randomization-stratum fixed effects and weights inversely proportional to batch size before the intervention, except where unweighted estimates are reported. Missing baseline values are set to zero and flagged with indicators. Baseline exams precede check-in 1. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A13: Treatment effects on the valuation of digital access

	Instagram deactivation price (INR) (1)	Value of phone access at work (INR) (2)	Accepts no-phone job at equal wage (share) (3)
Treatment	50	-1,801**	0.033*
	(100)	(725)	(0.019)
Control mean	2,632	12,509	0.559
Observations	1,422	2,161	2,161

Notes: The unit of observation is a student. Column 1 reports the stated reservation price for deactivating Instagram for two weeks, capped at the highest listed offer of INR 5,000. Column 2 is the nonnegative difference between the wage required for the no-phone and standard jobs. All columns are estimated by OLS and include randomization-stratum fixed effects and weights inversely proportional to cohort size before the intervention. Column 1 also controls for its baseline value. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A14: Treatment effects on misinformation outcomes

	Discernment index (1)	True-claim recognition (2)	False-claim rejection (3)	Novel false claims (4)
Deliberate Use	0.074	0.019	0.061	0.050
	(0.053)	(0.046)	(0.060)	(0.061)
Deliberate Use & Discernment	0.183***	-0.100*	0.237***	0.169***
	(0.048)	(0.054)	(0.050)	(0.051)
Control mean	0.004	-0.012	0.011	-0.006
Observations	2,377	2,377	2,377	2,377

Notes: The unit of observation is a student. Outcomes summarize predicted accuracy ratings from separate IRT models for true recognition and false rejection. The discernment index gives each of the two true and four false items in the endline battery equal weight. Effects are divided by the corresponding weighted control standard deviation at endline. Control means are predicted under control assignment. Models include separate indicators for both classroom arms, randomization stratum fixed effects, gender, urban residence, mother's college education, household wealth, and age. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. No baseline outcome is included. Weights are inversely proportional to cohort size before the intervention. Misinformation responses from March 24 and 25, 2025 are excluded because images did not display correctly. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A15: False-claim rejection and the interpretation of discernment

	False-claim rejection by source		Interpretive diagnostic	
	Circulating false claims	AI-generated novel false claims	Generalized caution	True–false separation
	(1)	(2)	(3)	(4)
Deliberate Use	0.067	0.050	0.579	1.690
	(0.058)	(0.061)	(0.905)	(1.303)
Deliberate Use & Discernment	0.236***	0.169***	3.861***	3.214**
	(0.052)	(0.051)	(0.970)	(1.435)
Control mean	0.013	-0.006	56.431	4.574
Observations	2,377	2,377	9,508	9,508

Notes: The unit of observation is a student. The unit of observation in columns (3) and (4) is a student-claim. Columns (1) and (2) use separate IRT models for rejection of circulating and novel false claims and report endline control standard deviations. These models use the same controls, missingness adjustment, and weights as the discernment index. Columns (3) and (4) use original ratings on a 0 to 1 scale and report effects in percentage points. They exclude novel claims and give true and false claims equal total weight for each student. Generalized caution includes claim and randomization stratum by truth fixed effects. Its sign is reversed so positive values indicate lower accuracy ratings. The true–false gap additionally includes respondent fixed effects. Neither diagnostic includes demographic or baseline controls. Both treatment indicators enter separately. Weights are inversely proportional to cohort size before the intervention. All columns include current affairs and exclude endline misinformation responses from March 24 and 25, 2025. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A16: Separating deliberate-use and misinformation effects

	Deliberate use index (1)	Discernment index (2)
Deliberate Use	0.120***	0.074
	(0.043)	(0.053)
Deliberate Use & Discernment	0.132***	0.183***
	(0.048)	(0.048)
Difference: Discernment arm − Deliberate Use arm	0.012	0.109
	(0.056)	(0.066)
Observations	2,432	2,377

Notes: The unit of observation is a student. The table compares the *Deliberate Use* arm with the *Deliberate Use & Discernment* arm. The deliberate-use index is an inverse-covariance-weighted index of less perceived social-media use, lower digital addiction, lower social comparison, greater interest in setting limits, and more commitment-device use. The discernment index combines predictions from positive IRT models for true recognition and false rejection, with weights matching the endline battery. No baseline score is included for this outcome. Coefficients are reported in control standard deviations. For discernment, demographic controls include gender, urban residence, mother's college education, household wealth, and age. Missing values are set to zero and flagged with indicators. Controls are centered and interacted with both treatment indicators. The difference row reports the contrast between the two treatment arms. Regressions include stratum fixed effects, weights inversely proportional to cohort size before the intervention, and standard errors clustered at the batch level. The deliberate-use index additionally controls for the baseline analogue with the standard missing-baseline adjustment. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. Both models control for gender, urban residence, mother's college education, household wealth, and age. Only the deliberate-use index includes its baseline value and cohort mean. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment.

Table A17: Discernment during the intervention and at follow-up

	Deliberate Use (1)	Deliberate Use & Discernment (2)	Difference (3)	Observations (4)
Check-in 1	-0.088	-0.124	-0.036	1,918
	(0.087)	(0.088)	(0.110)	
Check-in 2	0.122	0.037	-0.085	1,576
	(0.088)	(0.084)	(0.083)	
Check-in 3	-0.053	0.247***	0.300***	1,700
	(0.057)	(0.052)	(0.068)	
Endline	0.074	0.183***	0.109	2,377
	(0.053)	(0.048)	(0.066)	
Follow-up	-0.041	-0.065	-0.024	1,837
	(0.080)	(0.062)	(0.074)	

Notes: The unit of observation is a student. The discernment index combines predictions from separate positive IRT models for true recognition and false rejection, with weights matching the two true and four false items in the endline battery. Effects are reported in endline control standard deviations. Column (3) is column (2) minus column (1). Models include both treatment indicators, randomization stratum fixed effects, gender, urban residence, mother's college education, household wealth, and age. Missing covariates are set to zero and flagged with indicators. Demographic controls are centered and interacted with treatment. No baseline misinformation measure is included. The follow-up model controls for microlearning assignment. Weights are inversely proportional to cohort size before the intervention. March 24 and 25, 2025 endline misinformation responses are excluded. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A18: Objective phone use around sleep

	Hour before bedtime (1)	Hour after waking (2)
Panel A: Phone minutes per night		
Treatment	0.430	-0.382
	(1.117)	(0.667)
Control mean	23.857	6.173
Panel B: Any phone use		
Treatment	-0.016	0.024
	(0.017)	(0.035)
Control mean	0.944	0.734
Observations	949	949

Notes: The unit of observation is a student. Outcomes are average minutes per observed night, or the share of observed nights with any phone use, in the three nights before endline. Phone minutes include zero-use nights and therefore measure unconditional use in each window. The windows use bedtime and wake-up times reported at baseline, so their placement is determined before treatment. Only complete days with observed hourly use are included. Regressions include randomization-stratum fixed effects, weights inversely proportional to cohort size before the intervention, and standard errors clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A19: Treatment effects on phone use during scheduled class and breaks

	Total phone use (1)	Social media (2)	Non-social phone use (3)	Social media minus non-social use (4)
Scheduled class periods	-0.833	-0.953	0.131	-1.000
	(0.931)	(0.673)	(0.561)	(0.809)
Breaks and free periods	-2.387**	-0.625	-1.773***	1.246
	(1.133)	(0.844)	(0.619)	(0.953)
Scheduled class minus breaks	1.554	-0.328	1.904***	-2.246**
	(1.144)	(0.852)	(0.702)	(1.065)
Control mean: breaks	16.757	10.782	5.975	4.807
Control mean: scheduled class	13.364	7.369	5.998	1.372
Observations	15,009	15,009	15,009	15,009
Students	1,486	1,486	1,486	1,486

Notes: The unit of observation is a student-day. The table reports endline difference-in-differences effects on minutes of objective phone use per available timetable hour. Breaks and free periods are non-class time inside the regular timetable window, including scheduled breaks and time outside recorded class blocks. Column (4) is social-media use minus non-social phone use. Its class-minus-breaks cell is the triple-difference contrast testing whether the class-versus-break treatment effect differs between the two use types. The table uses objective records linked to course timetables. Regressions include fixed effects for student, survey period, relative day, day of week, and randomization stratum by calendar week. They use weights inversely proportional to cohort size before the intervention and to the number of observed days. Standard errors are clustered at the batch level. The sample requires positive scheduled class and non-class time on the same Monday–Saturday student-day and therefore contains 85 batches. This analysis uses up to seven days before each survey so that windows can cover a full calendar week. Calendar weeks use Stata's calendar-week convention. The measure is not conditioned on verified class attendance. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A20: Treatment effects on phone pickups during scheduled class

	No class-time pickup (1)	At most five pickups per class hour (2)	More than five pickups per class hour (3)
Treatment × Endline	0.013**	-0.005	0.005
	(0.006)	(0.028)	(0.028)
Control mean	0.014	0.386	0.614
Observations	7,710	7,710	7,710

Notes: The unit of observation is a student-day. Each column reports an endline difference-in-differences ITT estimate for objective phone pickups during scheduled class periods. A pickup is a screen-on event following at least 30 seconds with the screen off. The check-in 1 and endline windows use up to seven valid days before the corresponding survey. Columns (2) and (3) scale pickups by scheduled class hours. Regressions include fixed effects for randomization stratum by calendar week, survey period, relative day, day of week, and student. Weights inversely proportional to cohort size before the intervention and to the number of observed days. And standard errors clustered at the batch level. Days with incomplete or conflicting records are excluded. The measure does not condition on verified class attendance. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A21: Academic outcomes and robustness

	Control mean (percent) (1)	Effect (pp) (2)	Observations (3)
Panel A: Failure or reappearance			
First regular exam after Session 2	37.855	-4.992**	2,633
		(2.294)	
Any regular exam after Session 2	46.821	-4.729**	2,773
		(2.119)	
No manual or fuzzy matches	37.678	-4.907**	2,508
		(2.314)	
Unweighted estimates	37.209	-4.009*	2,633
		(2.090)	
Panel B: Numerical grades			
		Effect (SD)	
Main numerical grade estimate		0.003	2,386
		(0.043)	
Exclude grades depending on absent courses		0.004	2,168
		(0.043)	
Panel C: Outcome availability and baseline balance			
	Control mean (percent)	Difference (pp)	Observations
Failure outcome available	84.944	0.260	3,111
		(1.658)	
Numerical grade available	73.113	-0.877	3,111
		(3.289)	
Baseline failure or reappearance	32.300	-2.247	1,023
		(4.451)	

Notes: The unit of observation is a student. Panel A uses the first regular semester exam after Session 2 unless indicated otherwise. The row using any regular exam records failure or reappearance in at least one regular exam after Session 2. All adjust for baseline failure status. Panel B also adjusts for baseline numerical grades. Its second row excludes reconstructed grades requiring zero points for absence. Panel C reports outcome availability among eligible students at institutions supplying grade records. Its baseline row uses the failure sample with prior records. All regressions include randomization-stratum fixed effects and weights inversely proportional to batch size before the intervention, except where unweighted estimates are reported. Missing baseline values are set to zero and flagged with indicators. Baseline exams precede check-in 1. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A22: Treatment effects on academic goal setting

	Academic performance target
	(1)
Treatment	0.067*
	(0.040)
Control mean	0.000
Observations	2,422

Notes: The unit of observation is a student. Academic performance target indicates whether the student set a GPA or percentage target for the current semester. Regressions include randomization-stratum fixed effects. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A23: Heterogeneity by baseline social desirability: survey outcomes

	Treatment × social desirability (1)	Observations (2)
Content knowledge	0.064	729
	(0.078)	
Digital literacy	-0.031	872
	(0.067)	
Internal locus of control	-0.186*	872
	(0.109)	
Interest in setting limits	0.019	872
	(0.030)	
Commitment device use	-0.001	872
	(0.037)	
Digital addiction	0.036	872
	(0.076)	
Sleep index	-0.012	872
	(0.080)	
Wellbeing index	-0.040	872
	(0.060)	
Discernment (T2)	0.002	850
	(0.090)	

Notes: The unit of observation is a student. The Hays five-item social-desirability module was added partway through baseline fielding. The table uses 43 cohorts in strata with observed social-desirability responses in both treatment and control. Content knowledge was administered to a smaller sample than the other outcomes. Social desirability is standardized on the control group in the estimation sample. The interaction reports the change in the treatment effect for a one-standard-deviation increase in social desirability. The discernment row reports the change in the predicted T2 effect per standard deviation of social desirability, with T1 entered separately. Regressions use each outcome's main baseline adjustment and randomization-stratum fixed effects. They use weights inversely proportional to cohort size before the intervention and standard errors clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A24: Peer-exposure gradients across deliberate-use outcomes

	Slope among controls (1)	Difference in slopes (2)	Slope among treated students (3)	Adjusted q-value (4)	Observations (5)
Agency and knowledge					
Internal locus of control	0.031	0.061	0.092***	0.015	2,090
	(0.033)	(0.043)	(0.029)		
Content knowledge	-0.047	0.057	0.010	1.000	2,090
	(0.040)	(0.056)	(0.039)		
Other deliberate-use outcomes					
Digital literacy	-0.055	0.068	0.013	1.000	2,090
	(0.037)	(0.052)	(0.034)		
Interest in setting limits	-0.018	0.036	0.018	1.000	2,090
	(0.015)	(0.023)	(0.018)		
Commitment-device use	0.023	-0.022	0.002	1.000	2,090
	(0.016)	(0.024)	(0.018)		
Digital addiction	-0.008	0.006	-0.002	1.000	2,090
	(0.029)	(0.037)	(0.024)		
Sleep index	0.007	-0.030	-0.022	1.000	2,090
	(0.032)	(0.046)	(0.031)		
Wellbeing	0.017	-0.025	-0.008	1.000	2,090
	(0.028)	(0.040)	(0.028)		

Notes: Slopes correspond to a one-standard-deviation increase in residualized treated-friend exposure. Index outcomes are in standard-deviation units. Binary outcomes are in probability units. Exposure uses nominations made at baseline that could be matched to friends in other batches and nets out each friend's treatment probability under the randomization. Students with no matched friend in another batch receive zero exposure. Models control for the number of resolved nominations. Specifications include randomization-stratum fixed effects, weights inversely proportional to cohort size before the intervention, baseline values when available, no demographic controls, and standard errors clustered at the batch level. The unit of observation is the student. Adjusted q-values apply the Anderson sharpened false-discovery-rate adjustment to the eight treated-slope tests. These are exploratory mechanism estimates. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A25: Robustness of endline objective-use effects

	Total phone use	Social media	Non-social phone use	Observations
		Percent change		
	(1)	(2)	(3)	(4)
Panel A: Selection adjustments				
IPW for objective-use sample inclusion	-8.079**	-8.926**	-6.176	11,627
	(3.079)	(3.856)	(6.735)	
Reweight for entry at check-in 1 and retention to Endline	-3.406	-9.306	2.473	631
	(4.308)	(5.346)	(9.631)	
Panel B: Alternative weighting and aggregation				
Equal-student weights only	-7.823***	-10.620***	-4.926	11,627
	(2.720)	(3.320)	(6.547)	
Weights based only on cohort size	-7.092**	-9.493***	-4.462	11,627
	(2.698)	(3.355)	(6.550)	
Unweighted student-days	-5.984**	-10.291***	-3.023	11,627
	(2.564)	(3.184)	(6.500)	
Averaged by student and survey	-7.590**	-8.093*	-6.600	3,753
	(3.078)	(4.520)	(6.694)	
Panel C: Alternative baseline adjustment				
Control for baseline use	-8.850***	-9.878**	-6.172	11,627
	(3.001)	(3.718)	(6.649)	
ANCOVA controlling for check-in 1 use	-3.674	-9.668**	0.216	612
	(3.172)	(4.521)	(4.656)	
Panel D: Students observed at check-in 1 and endline				
		Minutes per day		Students
All students observed at both surveys	-21.369**	-12.693*	–	632
	(9.762)	(6.712)	–	
Also observed at baseline	-22.598	-9.870	–	346
	(15.159)	(9.822)	–	
First observed at check-in 1	-25.759*	-24.104**	–	286
	(13.310)	(11.992)	–	
Equality across baseline groups (p-value)	0.881	0.404	–	

Notes: The unit of observation is a student-day unless a row states that the data are averaged by student and survey or by student. Panels A–C report percentage effects from endline regressions with log-transformed outcomes. Standard errors appear below the estimates. Panel A changes the weights used to adjust for data availability. The first row reweights students for inclusion in the objective-use sample. The second reweights students first for entry at check-in 1 and then for retention through endline. Its standard errors come from 999 stratified bootstrap samples of batches that refit both models. Panel B changes the observation weights or level of aggregation. Panel C changes the pre-treatment adjustment. Panel D reports effects in minutes per day among the same students observed at check-in 1 and endline. Its last column counts students rather than student-days. The subgroup models allow fixed effects to differ by baseline contribution. Unless stated otherwise, specifications use the preferred three-day difference-in-differences design. They include fixed effects for randomization stratum by calendar week, survey period, relative day, day of week, and student. They use weights inversely proportional to cohort size before the intervention and to the number of observed days. Standard errors are clustered at the batch level. The ANCOVA row uses one observation per student and controls for check-in 1 use. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A26: Treatment effects on knowledge, agency, and strategy adoption

	Knowledge and understanding		Agency	Strategy adoption	
	Content knowledge	Digital literacy	Internal locus of control	Interest in setting limits	Commitment-device use
	(1)	(2)	(3)	(4)	(5)
Treatment	0.201***	0.099**	0.081*	0.039**	0.046**
	(0.052)	(0.045)	(0.042)	(0.018)	(0.019)
Control mean	-0.001	0.102	-0.021	0.366	0.597
Observations	2,090	2,090	2,432	2,432	2,432

Notes: The unit of observation is a student. Knowledge and agency outcomes are in control-group standard-deviation units. Strategy adoption outcomes are indicators. All regressions include randomization-stratum fixed effects and weights inversely proportional to cohort size before the intervention. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A27: Persistence of effects at follow-up

	Digital addiction (1)	Sleep index (2)	Internal locus of control (3)	Interest in setting limits (4)	Commitment-device use (5)
Class treat	-0.077**	0.114**	0.079*	0.005	0.012
	(0.033)	(0.049)	(0.042)	(0.017)	(0.024)
Control mean	0.020	-0.040	-0.050	0.328	0.482
Observations	1,838	1,838	1,838	1,838	1,838

Notes: The unit of observation is a student. Each column reports an ITT estimate of classroom treatment at follow-up. The outcomes were collected before the follow-up microlearning module, so the regressions do not condition on later microlearning assignment. Regressions control for the baseline value of the corresponding outcome, when a baseline analog exists. No baseline analog exists for the follow-up locus-of-control item. All specifications include randomization-stratum fixed effects. Specifications use weights inversely proportional to cohort size before the intervention and standard errors clustered at the batch level. Digital addiction, the sleep index, and locus of control are in control-group standard deviations. Interest in limits and commitment-device use are binary outcomes. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment.

Table A28: Digital literacy and content knowledge at follow-up

	Digital literacy (1)	Content knowledge (all) (2)	Content knowledge (self-control) (3)
Class treat	0.187***	0.129***	0.225***
	(0.049)	(0.049)	(0.054)
Microlearning	0.115**	0.160***	0.276***
	(0.048)	(0.050)	(0.057)
Control mean	0.014	0.051	0.008
Observations	1,841	1,841	1,841

Notes: The unit of observation is a student. Digital literacy and content knowledge were measured after the follow-up microlearning module. The regressions therefore include indicators for classroom treatment and follow-up microlearning assignment. The digital-literacy specification controls for baseline digital literacy. The content-knowledge outcomes have no baseline value. All specifications include randomization-stratum fixed effects. They use weights inversely proportional to cohort size before the intervention and standard errors clustered at the batch level. Outcomes are in control-group standard deviations. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment.

Table A29: Effects on objective phone use at endline and follow-up

	Total phone use (1)	Social media (2)	Non-social phone use (3)
Treatment × Endline	-8.878***	-9.964**	-6.321
	(2.992)	(3.717)	(6.636)
Treatment × follow-up	-1.435	-3.977	11.415
	(4.913)	(10.337)	(10.596)
Equality p-value	0.145	0.522	0.091
Students at endline	1,762	1,762	1,762
Students at follow-up	629	629	629

Notes: The unit of observation is a student-day. The table reports difference-in-differences ITT estimates from log-transformed daily objective use as percentage changes. The omitted period is check-in 1. All periods use the three most recent valid days. The model includes fixed effects for randomization stratum by calendar week, survey period, relative day, day of week, and student. Weights inversely proportional to cohort size before the intervention and to the number of observed days. And standard errors clustered at the batch level. The endline and follow-up rows use all eligible students observed in the corresponding panel and are not restricted to the same students. The equality p-value tests whether the endline and follow-up effects are equal. All specifications contain 86 batches. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A30: Persistence of attention and focus at follow-up

Panel A: Self-reported phone checking during class			
	Never checks phone	Checks at least twice	Checks at least five times
	(1)	(2)	(3)
Treatment	0.046**	-0.086***	-0.062***
	(0.020)	(0.020)	(0.019)
Control mean	0.166	0.645	0.258
Observations	1,804	1,804	1,804
Panel B: Task-based attention checks			
	First attention check	Middle attention check	Final attention check
	(1)	(2)	(3)
Treatment	0.022	0.085***	0.063***
	(0.021)	(0.023)	(0.022)
Control mean	0.751	0.579	0.522
Observations	1,935	1,935	1,935

Notes: The unit of observation is a student. Panel A reports self-reported phone checking during class. Panel B reports accuracy on three attention checks embedded at different points in the follow-up survey. The phone-checking thresholds are cumulative. Regressions include randomization-stratum fixed effects, weights inversely proportional to cohort size before the intervention, and the standard follow-up demographic controls. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A31: Misinformation discernment at follow-up

	Discernment index (1)	True claim recognition (2)	False claim rejection (3)
Deliberate Use	-0.041	0.068	-0.080
	(0.080)	(0.058)	(0.080)
Deliberate Use & Discernment	-0.065	0.002	-0.064
	(0.062)	(0.062)	(0.050)
Control mean	-0.075	-0.046	-0.046
Observations	1,837	1,837	1,837

Notes: The unit of observation is a student. The table reports effects from separate IRT models for true recognition and false rejection. The discernment index retains the relative weights on true and false claims used at endline. All effects use the corresponding endline control standard deviation. Control means are predicted under control assignment. Models include both classroom treatment indicators, follow-up microlearning assignment, randomization stratum fixed effects, gender, urban residence, mother's college education, household wealth, and age. Missing covariates are set to zero and flagged with indicators. All demographic controls are centered and interacted with classroom assignment. Weights are inversely proportional to cohort size before the intervention. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A32: Transfer effects on generative-AI use at follow-up

	Purposeful AI use	Components				
	Purposeful AI-use index	Verifies AI output	Tries problem first	Uses AI for exam prep	Productive exam-prep use	Substitutes for books or notes
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.205***	0.050	0.110*	0.126***	0.143***	0.094**
	(0.053)	(0.052)	(0.056)	(0.043)	(0.053)	(0.047)
Control mean	0.010	-0.008	0.017	0.034	-0.060	0.016
Observations	1,312	1,312	1,312	1,312	1,312	1,312

Notes: The unit of observation is a student. The table is estimated among respondents who reported using generative AI in the past 30 days. Outcomes are in control-group standard deviations and oriented so higher values indicate more purposeful use. The index combines columns (2)–(6). Regressions control for the later randomized microlearning assignment and standard demographic covariates, include randomization-stratum fixed effects and weights inversely proportional to cohort size before the intervention, and use standard errors clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment.

Table A33: Transfer effects on beliefs about generative AI at follow-up

	AI beliefs	Components					
	AI beliefs index	Academic benefits (3-item index)	Improves skills	Trusts accuracy	No skill erosion	Creates opportunities	Responsible-use confidence
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment	0.142	0.094	0.136	0.130	-0.006	0.090	0.140
	(0.140)	(0.115)	(0.119)	(0.133)	(0.107)	(0.128)	(0.109)
Control mean	0.020	0.020	0.006	0.004	0.005	0.023	0.010
Observations	651	651	651	651	651	651	651

Notes: The unit of observation is a student. These questions were administered to the smaller survey branch without the later microlearning module. All outcomes are in control group standard deviations. The AI beliefs index combines all components in columns (2)–(7), with concern about skill erosion reversed. The academic benefits outcome combines three survey items, so the table covers all eight items in the instrument. Regressions include standard demographic controls and randomization stratum fixed effects, weights inversely proportional to cohort size before the intervention, and standard errors clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment.

Table A34: Treatment effects by subgroup

	Digital addiction		Commitment-device use		Digital literacy		Discernment	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Sex is female	Male	Female	Male	Female	Male	Female	Male	Female
Treatment	-0.045	-0.152**	0.010	0.051	0.159***	0.116*	-0.051	0.363***
	(0.055)	(0.058)	(0.022)	(0.037)	(0.056)	(0.069)	(0.078)	(0.064)
Control mean	0.059	0.025	0.643	0.555	-0.146	0.186	0.060	-0.028
Observations	1,148	1,094	1,148	1,094	1,148	1,094	1,114	1,070
Equality p-value	0.187		0.377		0.610		< 0.001	
Home in urban area	Rural	Urban	Rural	Urban	Rural	Urban	Rural	Urban
Treatment	-0.091	-0.102*	0.077**	0.014	0.147*	0.133**	0.200*	0.163***
	(0.080)	(0.053)	(0.038)	(0.026)	(0.084)	(0.066)	(0.108)	(0.057)
Control mean	-0.006	0.061	0.592	0.602	-0.019	0.036	-0.024	0.026
Observations	594	1,652	594	1,652	594	1,652	573	1,615
Equality p-value	0.914		0.185		0.911		0.762	
Owned smartphone before 18	18 or after	Before 18	18 or after	Before 18	18 or after	Before 18	18 or after	Before 18
Treatment	-0.138	-0.085*	0.112***	0.002	0.087	0.152***	0.197*	0.154***
	(0.084)	(0.046)	(0.041)	(0.022)	(0.105)	(0.051)	(0.107)	(0.058)
Control mean	-0.074	0.079	0.557	0.613	0.015	0.023	0.018	0.014
Observations	590	1,633	590	1,633	590	1,633	572	1,594
Equality p-value	0.579		0.018		0.572		0.734	

Notes: The unit of observation is a student. Each panel reports subgroup-specific intent-to-treat estimates from a pooled regression that includes treatment, the subgroup indicator, and their interaction. Digital addiction, commitment-device use, and digital literacy report the effect of assignment to either treatment arm. Discernment reports the effect of the *Deliberate Use & Discernment* arm relative to control, from a regression that includes both arms and their interactions with the subgroup. The reported p-value tests equality of the two subgroup effects. Discernment uses the main IRT models with treatment interactions and averages the predicted treatment effect within each subgroup. Regressions include randomization-stratum fixed effects, weights inversely proportional to cohort size before the intervention, standard demographic controls (excluding the subgroup variable itself), and the baseline value of the outcome where available. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A35: Treatment effects on objective phone use by subgroup

	Total phone use		Social-media use	
	(1)	(2)	(3)	(4)
Sex is female	Male	Female	Male	Female
Treatment × Endline	-2.363	-59.734***	-25.088**	-23.177**
	(15.757)	(13.321)	(11.270)	(10.342)
Control mean	356.848	352.671	212.223	250.010
Observations	4,819	5,626	4,819	5,626
Equality p-value	0.009		0.904	
Age at baseline	Above median age	Median age or younger	Above median age	Median age or younger
Treatment × Endline	-20.103	-45.441***	-24.507*	-26.844***
	(17.645)	(12.581)	(12.431)	(9.499)
Control mean	372.579	348.494	252.885	223.188
Observations	3,053	7,398	3,053	7,398
Equality p-value	0.227		0.885	
Home in urban area	Rural	Urban	Rural	Urban
Treatment × Endline	-22.253	-39.992***	-9.413	-29.274***
	(21.316)	(11.748)	(17.082)	(8.446)
Control mean	342.925	359.267	225.159	233.125
Observations	2,622	7,842	2,622	7,842
Equality p-value	0.447		0.299	
Owned smartphone before 18	18 or after	Before 18	18 or after	Before 18
Treatment × Endline	-43.733**	-34.550***	-14.391	-30.296***
	(20.908)	(11.547)	(17.214)	(10.167)
Control mean	362.068	351.321	244.856	225.652
Observations	2,991	7,393	2,991	7,393
Equality p-value	0.698		0.485	

Notes: The unit of observation is a student-day. Each panel reports subgroup-specific endline difference-in-differences estimates in minutes per day. The omitted period is check-in 1. All periods use the three most recent valid days. Regressions include fixed effects for student, survey period, relative day, day of week, and randomization stratum by calendar week. They use weights inversely proportional to cohort size before the intervention and to the number of observed days. Standard errors are clustered at the batch level. Control means are endline means in the corresponding subgroup. Baseline age is split at the control-group median of 19 years. The reported p-value tests equality of the two subgroup effects. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A36: Effects of the microlearning lesson

	Platform beliefs (1)	Change intent (2)	Self-control intent (3)	Social pressure (4)	Fear of disconnecting (5)
Microlearning	0.010	0.051	0.021	0.005	-0.147**
	(0.042)	(0.043)	(0.046)	(0.050)	(0.058)
Control mean	-0.016	-0.055	0.002	-0.024	0.128
Observations	1,842	1,842	1,842	1,842	1,842

Notes: The unit of observation is a student. The row reports the effect of assignment to the digital self-command lesson, pooling versions with and without peer information, relative to the grammar lesson. Regressions control for classroom-treatment assignment, randomization-stratum fixed effects, weights inversely proportional to cohort size before the intervention, and the standard follow-up covariates. Regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A37: Incremental effects of peer-norm information on perceived peer norms

	Peers use commitment devices (1)	Peers approve high use (2)	Peers support a social-media break (3)
Peer-norm add-on	0.001	0.003	0.009
	(0.016)	(0.015)	(0.017)
Control mean	0.296	0.339	0.389
Observations	951	951	951

Notes: The unit of observation is a student. The sample is restricted to students assigned to a digital-self-command microlearning lesson. Coefficients compare the peer-norm add-on with the core lesson alone. Regressions include cohort fixed effects and control for gender, urban residence, mother's college education, household wealth, and age. Missing covariates are set to zero and flagged with indicators. All controls are centered and interacted with treatment. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A38: Treatment effects by cohort size: unweighted estimates

	Cohort size			
	Fewer than 45 (1)	45 or more (2)	Equal effects p (3)	Observations (4)
Mental health index (SD)	-0.168***	0.065	0.002	2,432
	(0.059)	(0.041)		
Content knowledge (SD)	0.218**	0.157**	0.592	2,090
	(0.094)	(0.063)		
Digital literacy (SD)	0.056	0.150**	0.374	2,090
	(0.074)	(0.061)		
Internal locus of control (SD)	0.155***	0.008	0.069	2,432
	(0.055)	(0.050)		
Interest in setting limits (pp)	3.138	3.886	0.859	2,432
	(2.986)	(2.445)		
Commitment device use (pp)	5.951*	1.696	0.268	2,432
	(3.050)	(1.933)		
Sleep index (SD)	0.044	0.031	0.873	2,430
	(0.064)	(0.044)		
Social comparison (SD)	-0.045	-0.087**	0.548	2,432
	(0.057)	(0.033)		
Discernment (SD)	0.271***	0.113*	0.129	2,377
	(0.083)	(0.057)		
Total phone use (%)	-14.138**	-4.175	0.175	11,627
	(5.980)	(2.782)		
Social media use (%)	-12.486*	-9.356**	0.709	11,627
	(6.826)	(3.573)		
Non-social phone use (%)	-18.846*	3.364	0.127	11,627
	(9.991)	(8.739)		
Failure or reappearance (pp)	-10.567***	-0.612	0.043	2,633
	(3.895)	(2.470)		

Notes: The unit of observation is a student for survey and academic outcomes and a student-day for objective phone use. Cohorts are split at the median pre-intervention size of 44.5 students. Columns (1) and (2) report endline treatment effects, except for academic failure, which uses the first regular semester exam after Session 2. Column (3) tests equality of these effects. Column (4) counts observations across both groups. SD denotes standard deviations and pp denotes percentage points. Lower values of the mental health index indicate better mental health. The misinformation estimate compares Deliberate Use & Discernment with control, including the other treatment arm separately. Survey regressions control for gender, urban residence, mother's college education, household wealth, age, and the baseline outcome and its cohort mean where available. Missing covariates are set to zero and flagged with indicators. Controls are centered and interacted with treatment. The discernment row averages IRT predictions within each size group. Phone-use estimates compare endline with check-in 1 and report implied percentage changes in daily use. They include student, survey-period, relative-day, day-of-week, and stratum-by-calendar-week fixed effects. Academic estimates control for baseline failure status and its missingness indicator. All models retain randomization-stratum adjustment and standard errors clustered at the batch level. No model uses inverse cohort-size weights. Phone-use regressions retain inverse observed-day weights. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A39: Robustness of main endline survey effects

	Knowledge, agency, and strategy adoption					Other outcomes			
	Content knowledge (1)	Digital literacy (2)	Internal locus of control (3)	Interest in setting limits (4)	Commitment-device use (5)	Digital addiction (6)	Sleep index (7)	Wellbeing index (8)	Discernment (T2) (9)
Main specification	0.201***	0.099**	0.081*	0.039**	0.046**	-0.078**	0.051	0.044	0.183***
	(0.052)	(0.045)	(0.042)	(0.018)	(0.019)	(0.034)	(0.037)	(0.043)	(0.048)
All students observed at endline	0.188***	0.154***	0.079*	0.038**	0.046**	-0.075**	0.046	0.041	0.183***
	(0.053)	(0.043)	(0.042)	(0.018)	(0.018)	(0.034)	(0.038)	(0.042)	(0.048)
English baseline survey only	0.181***	0.099**	0.093*	0.027	0.020	-0.090**	0.054	0.051	0.203***
	(0.055)	(0.048)	(0.050)	(0.020)	(0.018)	(0.037)	(0.037)	(0.046)	(0.053)
With demographic controls	0.201***	0.099**	0.081*	0.039**	0.046**	-0.078**	0.051	0.044	0.183***
	(0.052)	(0.045)	(0.042)	(0.018)	(0.019)	(0.034)	(0.037)	(0.043)	(0.048)
PDS lasso controls	0.195***	0.106**	0.077*	0.037**	0.047**	-0.078*	0.025	0.041	–
	(0.053)	(0.046)	(0.042)	(0.018)	(0.018)	(0.041)	(0.039)	(0.043)	
Excluding largest cohorts	0.195***	0.065	0.067	0.034	0.057**	-0.164***	0.080	0.085*	0.253***
	(0.061)	(0.053)	(0.057)	(0.024)	(0.024)	(0.040)	(0.051)	(0.050)	(0.075)
Excluding largest-cohort outliers	0.199***	0.091**	0.071*	0.040**	0.048**	-0.082**	0.052	0.042	0.181***
	(0.054)	(0.045)	(0.043)	(0.018)	(0.019)	(0.035)	(0.038)	(0.044)	(0.051)
Collapsed cohort means	0.176***	0.070	0.071	0.039*	0.047*	-0.061	0.046	0.024	–
	(0.066)	(0.064)	(0.047)	(0.022)	(0.024)	(0.045)	(0.046)	(0.052)	
Including pilot clusters	0.188***	0.162***	0.079*	0.036**	0.050***	-0.072**	0.053	0.031	0.183***
	(0.053)	(0.041)	(0.042)	(0.017)	(0.017)	(0.032)	(0.036)	(0.040)	(0.048)
Including BA clusters	0.196***	0.173***	0.072*	0.028	0.045**	-0.058*	0.059	0.038	0.160***
	(0.051)	(0.041)	(0.040)	(0.018)	(0.018)	(0.033)	(0.037)	(0.040)	(0.047)
No baseline imputation	0.201***	0.087*	0.081*	0.038**	0.031*	-0.109***	0.061	0.054	0.183***
	(0.052)	(0.047)	(0.042)	(0.018)	(0.018)	(0.035)	(0.039)	(0.045)	(0.048)
No baseline analog controls	0.201***	0.104**	0.081*	0.039**	0.046**	-0.117***	0.028	0.050	0.183***
	(0.052)	(0.048)	(0.042)	(0.018)	(0.019)	(0.043)	(0.042)	(0.047)	(0.048)
Original strata, supported sample	0.211***	0.105**	0.091**	0.039**	0.046**	-0.070*	0.049	0.027	0.204***
	(0.054)	(0.046)	(0.044)	(0.019)	(0.019)	(0.035)	(0.038)	(0.043)	(0.051)
Original randomized assignment	0.203***	0.107**	0.089**	0.040**	0.042**	-0.066*	0.030	0.019	0.170***
	(0.052)	(0.044)	(0.042)	(0.018)	(0.019)	(0.034)	(0.039)	(0.041)	(0.048)

Notes: The unit of observation is a student, except for the row based on cohort means. Each cell reports the treatment coefficient, with its standard error below. Treatment pools the two classroom arms except in the misinformation-discernment column, which reports the T2 effect relative to control while controlling for T1. The main specification uses weights inversely proportional to cohort size before the intervention, harmonized randomization-stratum fixed effects, baseline analogues with missing-value adjustment when available, and standard errors clustered at the batch level. In columns (1)–(8), the main specification includes centered demographic controls and treatment interactions, so the demographic-control row repeats it. Broader sample rows omit demographics. The cohort-mean and lasso rows use additive baseline controls. Excluding largest cohorts drops the top cohort-size tercile. Excluding largest-cohort outliers drops the two cohorts with 81 and 71 students. The collapsed row estimates cohort means, the English-only row restricts the baseline survey language, and the final two rows use the original strata or original assignment. Discernment uses the positive IRT models without baseline outcome controls. Lasso and collapsed cohort variants are not reported for this measure. Sample sizes vary across outcomes and rows. Content knowledge was administered to a smaller sample. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A40: Treatment effects under cohort and student weighting

Outcome	Cohort-weighted (1)	Student-weighted (2)	Difference (3)	Observations (4)
Content knowledge	0.201***	0.176***	-0.025	2,090
	(0.052)	(0.055)	(0.017)	
Digital literacy	0.099**	0.114**	0.015	2,090
	(0.045)	(0.045)	(0.017)	
Internal locus of control	0.081*	0.061*	-0.020	2,432
	(0.042)	(0.036)	(0.021)	
Interest in setting limits	0.039**	0.036**	-0.003	2,432
	(0.018)	(0.017)	(0.007)	
Commitment-device use	0.046**	0.033**	-0.013	2,432
	(0.019)	(0.016)	(0.008)	
Digital addiction	-0.078**	-0.014	0.065***	2,441
	(0.034)	(0.035)	(0.012)	
Sleep quality and hygiene	0.051	0.036	-0.014	2,430
	(0.037)	(0.035)	(0.014)	
Wellbeing	0.044	0.017	-0.026	2,432
	(0.043)	(0.039)	(0.016)	
Discernment (T2)	0.183***	0.163***	-0.020	2,377
	(0.048)	(0.047)	(0.023)	

Notes: The unit of observation is a student. Column (1) uses the preferred weights inversely proportional to cohort size before the intervention, while column (2) gives each student equal weight. Column (3) is the student-weighted estimate minus the cohort-weighted estimate and uses joint cluster inference for the two estimates. Standard errors are reported in parentheses on the row below each estimate. For all outcomes except misinformation discernment, treatment pools the two classroom arms. The discernment row reports the effect of *Deliberate Use & Discernment* (T2) relative to control from a regression that also includes the *Deliberate Use* (T1) indicator. All specifications use the same outcome-specific samples, baseline adjustment, harmonized randomization-stratum fixed effects, and standard errors clustered at the batch level as the corresponding main tables. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A41: Objective-use estimates across measurement windows

Window	Total phone use (1)	Social media (2)	Observations (3)
Panel A: Log-transformed outcomes (percent)			
Up to 2 days	-10.570***	-12.128***	7,836
	(3.662)	(4.285)	
Up to 3 days	-8.878***	-9.964**	11,627
	(2.986)	(3.711)	
Up to 4 days	-8.286**	-10.159***	15,281
	(3.049)	(3.595)	
Up to 5 days	-8.204**	-10.497***	18,688
	(3.133)	(3.650)	
Up to 6 days	-8.993***	-11.566***	21,931
	(2.964)	(3.496)	
Up to 7 days	-8.848***	-9.975***	25,029
	(2.831)	(3.502)	
Panel B: Outcomes in minutes per day			
Up to 2 days	-34.840***	-28.432***	7,836
	(10.917)	(7.737)	
Up to 3 days	-31.996***	-24.865***	11,627
	(9.343)	(6.896)	
Up to 4 days	-24.947***	-20.331***	15,281
	(8.821)	(6.567)	
Up to 5 days	-19.499**	-15.710**	18,688
	(9.052)	(6.910)	
Up to 6 days	-21.159**	-16.407**	21,931
	(8.597)	(6.571)	
Up to 7 days	-22.445***	-15.179**	25,029
	(8.187)	(6.639)	

Notes: The unit of observation is a student-day. Each row varies the maximum number of valid days included before each survey and uses all available observations for that window. Panel A uses log-transformed outcomes and reports the implied percentage change. Panel B reports effects in minutes per day. The omitted period is check-in 1, and the reported coefficient is the endline difference-in-differences estimate. All specifications include fixed effects for randomization stratum by calendar week, survey period, relative day, day of week, and student. They use weights inversely proportional to cohort size before the intervention and to the number of observed days. Standard errors are clustered at the batch level. ***, **, and * denote significance at the 1%, 5%, and 10% levels, respectively.

Table A42: Cost-effectiveness and external benchmarks

	(1)	(2)
Panel A: Resource cost per student place		
		Cost (USD)
Sessions 1–2		2.29
Session 3 add-on		1.14
Full Sessions 1–3		3.43
Design-average package		2.86
Full package plus one hypothetical refresher		4.58
Panel B: Partial benefit–cost calculations		
		Estimate (USD)
Cost per avoided failure or reappearance		57
Academic benefit–cost ratio: event valued at USD 250		4.36
Academic benefit–cost ratio: event valued at USD 500		8.73
Break-even value per total-phone hour, 30 days		0.19
Break-even value per social-media hour, 30 days		0.26
Break-even value per total-phone hour, 60 days		0.10
Break-even value per social-media hour, 60 days		0.13
Panel C: Discernment benchmark		
	Effect used	Cost per 0.1 SD (USD)
Added Session 3	0.109 SD	1.05
Added Session 3 plus one hypothetical refresher	0.109 SD (scenario)	2.10
Amar et al., immediate	0.32 SD	1.51
Amar et al., four months	0.26 SD	1.86
Panel D: Outcome-specific standardized-effect ratios		
	Effect	Cost per 0.1 SD (USD)
Lower digital addiction, endline	0.078 SD	3.67
Sleep index, follow-up	0.114 SD	2.51
Digital literacy, follow-up	0.187 SD	1.53

Notes: Panel A reports incremental delivery costs: INR 2,000 in facilitator time, INR 200 from amortizing training over 100 sessions, INR 250 in materials, and INR 50 in coordination costs, divided by median attendance of 23 students. Using attendance rather than assignment in the denominator is conservative for intent-to-treat effects. The calibration assumes delivery in non-course institutional time and excludes participant time, development, research, evaluation, and institutional overhead; the design-average package weights the treatment arms equally. Panel B uses endline effects of 29.4 fewer total-phone minutes and 21.7 fewer social-media minutes. The academic calculation uses a 5.0-percentage-point reduction in failure or reappearance in the first regular semester exam after Session 2 ($p = 0.033$). Each calculation charges the full relevant package cost to one outcome and assigns zero value to all other gains, so its benefit–cost ratio is a lower bound for the measured bundle and its break-even value is an upper bound on what the focal outcome must be worth. Phone-use calculations assume the endline effect lasts 30 or 60 days; ratios omit sampling uncertainty. Panel C compares the added misinformation session with Amar et al. (2026). The refresher row is an untested scenario in which the effect after one later session equals the original endline effect; studies differ in instrument, population, delivery, price year, and cost components. Panel D reports outcome-specific standardized-effect ratios, which are not common welfare units and should not be added. Social-media use is nested in total phone use.

Table A43: Outcomes and pre-analysis status

Outcome family	Outcome or index reported	Status in PAP	Notes on implementation
Smartphone and social media use	Objective usage for total phone use, social media use, non-social media use, app categories, and selected social media apps; survey-reported social media use	Primary outcome: social media consumption	The plan specified survey and app-log measures and anticipated granular app-use outcomes.
Mental health and wellbeing	Pre-specified composite well-being index; digital addiction; social connectedness; happiness; life satisfaction; anxiety symptoms; depression symptoms	Primary outcome: well-being index	Digital addiction was pre-specified as a component of the composite and is reported separately in the paper.
Misinformation discernment	Overall discernment; false claim rejection; true claim recognition	Primary outcome: misinformation discernment	The paper uses separate positive IRT models for true recognition and false rejection, with overall weights matching the endline battery. It also reports the equally weighted true–false rating gap.
Commitment devices	Interest in setting limits; self-reported adoption of commitment devices or strategies	Secondary outcome: adoption of commitment devices	The plan specified both survey and Dashboard-based measures; the paper uses the survey measure.
Academic outcomes	Grade; failure or reappearance indicators	Secondary outcome: academic outcomes	Administrative grades are reported after linkage and timing checks.
Phone use during school hours	Phone use during scheduled class time and outside scheduled class time; social media and non-social media use during timetable windows	Secondary outcome: concentration during class hours	The paper implements this outcome by combining usage and class timetable data.
Phone use around bedtime	Phone use in the hour before reported bedtime	Secondary outcome: phone use around bedtime	The paper uses reported sleep time to construct phone use before sleep.
Sleep	Sleep index based on sleep timing and self-reported sleep quality	Secondary outcome: sleep quality	The paper combines closely related sleep outcomes into a single index and reports components separately.
Social comparison	Social comparison index and components	Secondary outcome: social comparison	The paper reports the social-comparison index.
Digital literacy and content knowledge	Digital-literacy index; curriculum content knowledge	Mechanism outcome	The paper reports digital literacy using vignettes and subject knowledge from the curriculum.
Locus of control and agency	Internal locus-of-control index	Mechanism outcome	The paper reports locus of control index.
Valuation of digital access	Willingness to accept deactivation and valuation-related outcomes	Exploratory outcome	The paper reports valuation of social media.
Aspirations and effort	Academic, professional, and personal plans; job-search effort	Exploratory outcome	The paper reports career aspirations and job-search effort where available

Continued on next page

Table A43 (continued)

Outcome group in the paper	Outcome or index reported	Status in pre-analysis plan	Notes on implementation
Microlearning survey experiment	Follow-up microlearning effects on perceptions, intentions, beliefs, and attitudes	Survey-experiment amendment	These outcomes are covered by the follow-up survey-experiment amendment and the paper reports results following the amendment.
Generative AI transfer	Follow-up survey measures of AI use, AI beliefs, AI confidence	Not in original classroom PAP	These outcomes are exploratory evidence on transfer to a new technology.
Attention checks	Follow-up attention-task performance	Not in original classroom PAP	These outcomes are exploratory evidence on attention.

Notes: The table maps outcomes and indices reported in the paper to the original classroom pre-analysis plan and the follow-up survey-experiment amendment. It excludes non-outcome exhibits. The status *Not in original classroom PAP* means that the outcome was not listed as a primary, secondary, exploratory, or mechanism outcome in the original classroom pre-analysis plan. Indices in the paper are standardized and inverse-covariance weighted unless otherwise noted in the corresponding outcome table.

Appendix C Design Appendix

C.1 Threats and Design Responses

In response to anticipated threats to internal validity, we implemented several design choices and field protocols to minimize their effects.

Participation and attrition. In this setting, students move in and out of the study. We work with participating batches and allow any student who is present on a day we visit to take part in a survey or session. This approach matches how such a program would be delivered in practice, and it lets us enroll students broadly rather than only those interested in the program. But it also means that who we observe changes from visit to visit, as attendance in higher-education institutions is irregular (Romer, 1993; Credé, Roch, and Kieszczynka, 2010).

We implemented several measures to reduce attrition. First, participants received monetary incentives for the baseline, endline, and follow-up surveys, and a certificate of participation for the endline and follow-up surveys. Second, we worked with class representatives, identified in coordination with college administrators or faculty, to circulate schedules and reminders.[62] Third, we created make-up opportunities for the curriculum and the surveys. Before the endline, students who had missed sessions but participated in at least one prior interaction were invited to a make-up session, where treated students received a condensed summary of the relevant curriculum content before completing the survey, and control students completed the survey only. If students could not be reached in person, we scheduled remote survey sessions after 4pm or on weekends, when response rates were higher, and password-protected the links to prevent early or unauthorized access.

Students could also enter the study after baseline. Survey visits reached whichever students were in class that day, and some students absent at baseline first appeared at a later wave. Turning them away would have disrupted the session, so we surveyed them and gave first-time respondents a version of the survey that included a short demographic module.[63]

Availability of objective usage data. Another issue is retention of Dashboard, our usage-tracking app. Most students were willing to install it and export data but not able to keep it running consistently in the background. The earlier version of the app, fielded for the first batches in Wave 1, required students to keep it running in the background. As we learned that most students uninstalled it (often to preserve space and battery on their phones), we redesigned Dashboard and introduced the export function for subsequent batches. As a result, baseline screen time data are available for only 59 of the 86 analysis batches. We updated the design of Dashboard to retrieve data already on the device, so students could easily install the app and contribute historical data

[62] Class representatives received certificates and tokens of appreciation when their batches achieved high attendance.

[63] Because entry after the first session could itself respond to treatment, the primary analyses include only students first observed at baseline or check-in 1, both before any classroom content was delivered, as discussed in Section 4.4.

through a quick export during a visit.[64]

Spillovers. We randomized at the batch level rather than the student level to reduce potential spillovers between groups. Control students do not sit in the sessions their treated peers attend, and because courses are organized by major and year, the two groups largely do not share classes. Students nonetheless interact across batch boundaries. To measure how much such contact occurs, the baseline survey asked students to name their closest friends, which allows us to compute each student's exposure to treated peers outside their own batch and to test for transmission directly.

Experimenter demand. Students may infer what the study is about and adjust their responses accordingly. Three features of the design limit this concern. First, control batches completed the same survey waves and check-ins as treated batches, in the same settings and on the same schedule, so contact with the research team is balanced across arms and any effect of simply being studied is differenced out. Second, our central outcomes do not rely on students' descriptions of their own behavior. We measure screen time from historical data on the student's phone, academic performance from university administrative records, and discernment and attention from task-based items. Third, for a subset of batches, the baseline survey included the five-item social-desirability scale of Hays, Hayashi, and Stewart (1989), which lets us test whether treatment effects are larger among students most prone to socially desirable responding.[65] Appendix Table A23 interacts treatment with this baseline measure for the main survey outcomes. The interactions are generally small and imprecise.

C.2 Measuring Discernment

Sample posts. In our surveys, the misinformation discernment module presented students with screenshots of social media posts. The posts include true claims, false claims that had circulated online, and novel false claims that the research team created for the survey and that were never in circulation. They cover several topic areas, including health, technology and science, global affairs, and current affairs. Each student rated a randomly drawn subset of posts. Appendix Figure A12 shows one post of each claim type from the technology and science topic area.

Discernment index. We rescale ratings to the interval from zero to one and reverse ratings of false claims. Separate graded-response IRT models for true recognition and false rejection account for differences in item difficulty, with all item discriminations constrained to be positive. The first two check-ins use a common discrimination within each truth category because they contain only two true items. Later surveys allow item-specific discriminations. Items not shown to a respondent are treated as missing, not incorrect. The overall index averages predicted accuracy with one-third

[64] As too few students kept the app installed, take-up of its limit feature was too low to serve as a measure, and we instead rely on self-reported commitment device use in the survey.

[65] The module was introduced partway through baseline data collection, so it covers the 43 batches surveyed thereafter.

weight on true recognition and two-thirds on false rejection, matching the two true and four false claims in the endline battery. We retain these weights at check-ins and follow-up and divide effects by the weighted endline control standard deviation of the six-item score.

Both treatment indicators enter each model, along with randomization stratum fixed effects and the demographic controls and missingness adjustment described in the exhibit notes. Treatment effects average predicted ratings under each assignment over the same students and item bank. Standard errors use the joint score covariance clustered at the batch level, including uncertainty in item parameters and covariance between the two models. No baseline misinformation score is included. Item banks differ across surveys, so the trajectory shows treatment effects at each survey rather than changes in students' scores on a common test.

Endline misinformation responses from March 24 and 25, 2025 are excluded because field records document a problem displaying claim images. Other outcomes and surveys on these dates are retained.

Appendix D Results Appendix

D.1 Robustness to Alternative Specifications

Appendix Table A39 shows that the conclusions from the main endline analysis are robust to alternative analytic choices. Across changes in sample inclusion, including restriction to students who completed the baseline survey in English, covariate adjustment, missing-baseline handling, treatment coding, and the level of aggregation, every estimate retains the sign of the main specification. Magnitudes are generally comparable, although some estimates shift and become less precise when the largest batches are excluded or the data are collapsed to batch means. Appendix Table A40 compares the preferred batch-weighted estimates, which give randomized batches approximately equal influence, with unweighted student-level estimates, which give larger batches proportionally more influence.

The objective-use estimates are stable across specification, support, and sample choices. Under alternative selection weights, observation weights, aggregation, and pre-treatment adjustments, the endline estimates retain their sign and magnitude (Appendix Table A25). Varying the measurement window from two to seven days produces similar proportional reductions in total phone use and social media use (Appendix Table A41). The minute estimates vary more in magnitude but remain negative and statistically significant throughout.

Appendix E Mechanisms Appendix

E.1 Information Provision and Delivery

The microlearning experiment separates the effect of a short private lesson from the incremental effect of information about peers. As described in Section 7, batches were randomly assigned to an English grammar lesson or a digital self-command lesson. Within batches assigned to digital self-command, students were randomized individually with equal probability to receive the core lesson alone (S1) or the core lesson with additional quotes based on the experiences of other students in the study (S2). The experiment followed measurement of the main classroom experiment's follow-up outcomes.

The additional peer information produces little incremental change. Panel B of Figure 10 compares S2 with S1, controlling for batch fixed effects and demographic characteristics. The estimates show little change in either content knowledge or perceptions about peers. Because the messages did not detectably shift perceptions, this comparison evaluates the particular messages used in S2 rather than the consequences of changing students' perceptions about peers.

The private lesson and the classroom program differ in dosage as well as delivery. The classroom program combined information with repeated exercises and classroom discussion. Their comparison therefore does not isolate the contribution of classroom interaction.

E.2 Peer Reinforcement

We use friendship nominations collected at baseline. Students could nominate up to three other students at their institution with whom they regularly interacted online or in person. We match nominations to survey respondents using the nominated student's name, program, and year, and use the matched respondent's batch to identify treatment assignment. The exposure measure uses friends in other randomized batches, since friends in the same batch share the student's assignment. Let N_i denote the set of student i's matched friends in other batches, $d_i = |N_i|$ the number of such friends, and T_j an indicator that friend j's batch was assigned to either treatment arm. Let C_s and M_s denote the numbers of batches and treated batches in original randomization stratum s. Conditional on student i's own assignment, friend j's treatment probability is

$$p_{ij} = \begin{cases} \dfrac{M_{s_j} - T_i}{C_{s_j} - 1}, & s_j = s_i, \\ \dfrac{M_{s_j}}{C_{s_j}}, & s_j \neq s_i. \end{cases}$$

The first case removes i's own batch from the assignment risk set. Residualized treated-friend exposure is

$$G_i = \frac{1}{d_i} \sum_{j \in N_i} (T_j - p_{ij}), \qquad \text{for } d_i > 0,$$

and $G_i = 0$ for students with no matched friend in another batch. Thus exposure measures

departures from the treated share expected under the randomization, rather than differences in treatment probabilities across strata.

For the comparison of internal locus of control and content knowledge in Figure A15, we estimate

$$Y_i = \alpha + \beta T_i + \rho G_i + \theta(T_i \times G_i) + \lambda_{h(i)} + \delta d_i + \varepsilon_i,$$

where $h(i)$ indexes the strata used in the main outcome regressions. The treatment probabilities above use the original assignment strata. Regressions use weights inversely proportional to batch size before the intervention and standard errors clustered at the batch level. The two outcomes use the same 2,073 students. We report slopes for a one-standard-deviation increase in G_i, using its standard deviation in this common sample. The control-group gradient is ρ, the treated-group gradient is $\rho+\theta$, and their difference is θ. Conditional on the baseline friendship network, variation in friends' exposure comes from the randomized assignment of their batches.

The two comparisons in the main text answer different questions. Comparing the locus-of-control and content-knowledge gradients among treated students asks whether exposure is associated more with perceived agency than with knowledge of the curriculum. The difference is 0.083 SDs ($p = 0.060$). We obtain its standard error from a joint regression of the two outcomes, allowing all coefficients and fixed effects to differ by outcome and accounting for their covariance through clustering. Comparing the locus-of-control gradient between treated and control students instead asks whether exposure complements direct receipt of the curriculum. The estimated difference is 0.061 SDs ($p = 0.162$).

Appendix Table A24 reports all eight outcomes on the same sample and adds baseline outcome controls where available. The false-discovery-rate adjustment treats the eight gradients among treated students as one family. The locus-of-control gradient has an adjusted q-value of 0.015. These estimates are consistent with peers making the curriculum's strategies easier to carry out, but do not separately identify changes in friends' behavior, social costs, or expectations.

Appendix F Program Cost-effectiveness

To measure cost-effectiveness, we construct a benchmark delivery cost covering facilitator time, training, materials, and coordination. At the observed median attendance of 23 students, the two-session *Deliberate Use* curriculum costs approximately USD 2.29 per student place, adding the misinformation session costs another USD 1.14, and the complete three-session curriculum costs USD 3.43. For outcomes estimated by pooling the two treatment arms, we use USD 2.86, the design average implied by equal assignment to the two curricula. Even doubling these amounts to allow for the opportunity cost of the class or student time used for delivery leaves the complete curriculum below USD 7 per student.[66]

[66] Our benchmark cost is INR 2500 per session, comprising INR 2000 for facilitators, INR 200 for training amortized over 100 sessions, INR 250 for materials, and INR 50 for coordination. We exclude curriculum development, which is sunk for future adopters as materials will be freely available, as well as research and evaluation costs. Dividing

Each row of Appendix Table A42 asks whether one outcome alone covers the entire relevant program cost. The calculation charges the full program cost to that outcome and assigns zero value to every other measured gain. Each reported ratio is therefore a lower bound on the return to the full bundle, and each break-even valuation is an upper bound on what the focal outcome must be worth for the program to pay for itself.

The academic effect clears this threshold under modest valuations. Treatment reduces the probability that a student fails or must reappear for an examination by 5.0 percentage points ($p = 0.033$). At the point estimate, assigning about 20.0 students to treatment prevents one such event, at a program cost of roughly USD 57, or USD 115 after doubling program costs. If preventing one failure or reappearance is worth at least USD 250 to the student and institution combined (e.g., by avoiding repeated examination and study effort, delayed academic progression, and consequences for labor-market outcomes), the academic effect alone generates at least 4.36 dollars in benefit per dollar of benchmark program cost, or 2.18 after doubling costs.

The effect on screen time alone could also justify the program cost under modest assumptions. Over 30 days, the endline reduction in total phone use covers the benchmark cost if a reduction of one hour is worth at least USD 0.19 for total phone use, or USD 0.26 for social media use. Over 60 days these thresholds fall to USD 0.10 and USD 0.13. They are easy to clear because the reduction falls on use students themselves wanted to cut, and the decline is concentrated in social media and longer sessions, while productivity and education use do not decline.

Under the benchmark cost, a 0.1 SD reduction in endline digital addiction costs approximately USD 3.67. This is a break-even shadow value: the improvement would cover the entire program cost if a 0.1 SD reduction in digital addiction were worth more than this amount.

The misinformation module is inexpensive relative to its effect and comparable to the closest external benchmark. Using the point estimate, adding Session 3 costs USD 1.14 per student and has an estimated endline discernment effect of 0.109 SD relative to the two-session curriculum, or about USD 1.05 per 0.1 SD, while the full curriculum costs about USD 1.87 per 0.1 SD relative to control. Amar et al. (2026) price a more intensive Indian classroom program focused on health misinformation at USD 4.84 per student in their full-cost model, which implies USD 1.51 per 0.1 SD immediately after the program and USD 1.86 using their 0.26 SD effect four months later.[67] Our effect on discernment is no longer detectable 10.5 months after Session 3, but one fully priced refresher would raise the complete curriculum cost from USD 3.43 to USD 4.58 per student, and would imply USD 2.50 per 0.1 SD if it restored the original effect. While the refresher scenario is a policy calculation rather than an experimental estimate, it suggests that accommodating decay need not make the program expensive.

The same features that make the program inexpensive per student also make it scalable. List

by median attendance rather than enrollment conservatively increases the cost per student. To account for scarce institutional or participant time, our high-cost sensitivity assigns an additional opportunity cost equal to the entire measured delivery cost of each student-session.

[67] We use their full-cost model because it prices facilitator time, training, and materials, as our benchmark does. The comparison is suggestive, since the two studies differ in population, measurement, delivery, and price year.

(2024) identifies several threats that cause interventions to lose voltage at scale, and the design speaks to each. The effects are measured across three independent sources, usage data, surveys, and administrative records, so they do not rest on a single instrument. Because recruitment ran at the institutional level, students were exposed through their batch rather than through self-selection, so the estimates are not inflated by prior interest in digital wellness. Because the experimental unit is also the delivery unit, the intent-to-treat effects already reflect treating a whole batch, including the peer reinforcement, under conditions mirroring those in potential rollouts: multiple waves, several facilitators, in-person and remote sessions, and incomplete take-up. Moreover, delivery uses standardized materials and facilitator scripts with training as a largely fixed cost, so the per-student costs calculated above could fall further at scale. The returns seem to come from the facilitated exercises, discussion, and shared classroom setting, not from the informational content. This suggests scaling is likely to work best by keeping the facilitated, classroom-based format rather than moving to individually distributed passive content or much larger groups, though the most effective delivery mode at scale is itself worth testing.

Appendix G Deviations from the pre-analysis plan

This section documents deviations from the analysis plan discussed in the pre-registration report and the corresponding analysis choices in the paper. Appendix Table A43 maps the paper's reported outcomes and indices to the original plan and the follow-up survey-experiment amendment.

Commitment device use. The original PAP proposed two measures of commitment device use: (i) a self-reported survey measure and (ii) an objective Dashboard-based measure constructed from app-level usage limits (including a limit tightness measure adapted from Allcott, Gentzkow, and Song (2022)). In the final analysis, we do not use app-based limits as the primary measure of commitment device use. Instead, our primary measure is self-reported use of commitment devices. The change reflects low take-up of the specific Dashboard in-built commitment device.

Sleep. We combine the two pre-specified sleep outcomes into a single index to reduce the number of tests. The subindices and components are reported in Appendix Table A11.

Microlearning component. The original PAP for the field experiment described a text-message microlearning component. We did not implement that component as planned because pilot testing showed consistently low take-up across multiple text-message delivery strategies. Instead, we implemented an in-person, survey-embedded microlearning experiment during the follow-up survey. This change was pre-specified in an amendment to the pre-registration report before being deployed in the field.

Discernment. We use separate positive IRT models for true recognition and false rejection rather than the pre-specified difference in rating totals because each student saw a randomly selected subset

of items drawn from a larger item bank. The discernment index weights true and false performance according to the endline battery. We also report an equally weighted true–false gap using the original ratings of circulating claims. Appendix C.2 describes the construction.

Sample and batch size. The pre-registration report anticipated a sample of 100 batches. In implementation, realized batch sizes were more variable than expected, which reduced the total number of batches that could be sampled within field constraints. The final sample includes 86 batches and 3,430 students. Because randomization occurred at the batch level and batch sizes vary substantially, our primary specification is a batch-weighted ITT that assigns equal total weight to each randomized batch (Athey and Imbens, 2017). This specification aligns the estimand with the unit of randomization and avoids mechanically giving more influence to larger batches. We report unweighted student-level estimates in Appendix D.1.

Instagram data. The preregistration report described collecting Instagram activity data from a consenting subset of participants for exploratory analyses (including text-based analyses of platform behavior). In practice, the consenting sample was too small to support informative analysis. We therefore do not include the planned Instagram-based exploratory analyses in the paper, consistent with the stated caveat in the PAP that these analyses depended on obtaining a sufficiently large consenting subsample.