Sleep: Educational Impact and Habit Formation*

Osea Giuntella
University of Pittsburgh
IZA and NBER

Silvia Saccardo
Carnegie Mellon University

Sally Sadoff
UC San Diego

March 21, 2025

Abstract

There is growing evidence on the importance of sleep for productivity, but little is known about the impact of interventions targeting sleep. In a field experiment among U.S. university students, we show that incentives for sleep increase both sleep and academic performance. Motivated by theories of cue-based habit formation, our primary intervention couples personalized bedtime reminders with morning feedback and immediate rewards for sleeping at least seven hours on weeknights. The intervention increases the share of nights with at least seven hours of sleep by 26 percent and average weeknight sleep by an estimated 19 minutes during a four-week treatment period, with persistent effects of about eight minutes per night during a one to five-week post-treatment period. Comparisons to secondary treatments show that immediate incentives have larger impacts on sleep than delayed incentives or reminders and feedback alone during the treatment period, but do not have statistically distinguishable impacts on longer-term sleep habits in the post-treatment period. We estimate that immediate incentives improve average semester course performance by 0.075 - 0.088 grade points, a 0.10 - 0.11 standard deviation increase. Our results demonstrate that incentives to sleep can be a cost-effective tool for improving educational outcomes.

JEL CODES: C93, I10, I23

Keywords: Sleep, Human Capital, Education, Habit Formation, Health Behaviors, Field-Experiments

---

*We are grateful to the seminar participants at the BEDI Conference, the Roybal Annual Retreat, University of Pittsburgh, the Center for Sleep and Circadian Rhythms Study, the Virtual Seminar on the Economics of Risky Behaviors (VERB), the CHIBE Behavioral Science and Health symposium, the University of Houston, the National University of Singapore, the Freie Universität Berlin, UCSD Spring School on Behavioral Economics, University of Chicago, Harvard University, RAND, University of Southern California, the UCLA Anderson School of Management, the New York Federal Reserve, and the NBER Economics of Health Meetings (Fall 2023). We are grateful to Mallory Avery, Daniel Banko-Ferran, Liam Goodwin, Kelly Hyde, Ilaria Prometti, Ben Schenck, Samuel Lindquist, Gabrielle Toborg, and William Wang for their valuable research assistance. We are thankful to the Pittsburgh Experimental Economics Lab for recruiting our participants and the University of Pittsburgh Registrar Office for their help in providing access to students' academic records. We also thank Tomáš Lintner for providing data related to educational program effects. We received generous funding from J-PAL North America (Sadoff), the National Institute of Aging (Saccardo, Grant #P30AG034546) and the Pitt Healthy Lifestyle Institute Pilot and Feasibility project (Giuntella).

1 Introduction

There is growing attention to the role of sleep for economic outcomes (Hillman et al., 2006; Mullainathan, 2014; Rao et al., 2021). At the same time, statistics suggest people are not sleeping enough. About a third of Americans sleep less than the recommended minimum of seven hours per night and a similar proportion state that they would like to sleep more (Jones, 2013; Ballard, 2019; Corkett, 2010; CDC, 2023). Sleep deprivation is even worse among adolescents and young adults, with about two-thirds regularly failing to meet sleep guidelines (Wheaton et al., 2018). These statistics have prompted policy concern that poor sleep may be worsening educational outcomes (Group et al., 2014). More broadly, Roenneberg (2013) referred to sleep deprivation as the most prevalent high-risk behavior in modern societies. Yet, little is known about whether interventions targeting sleep can improve productivity and performance.

Our study implements a randomized field experiment among U.S. university students to examine the impact of interventions targeting sleep on sleep habits and academic performance. Our work has three primary aims. First, we address long-standing gaps in understanding the causal relationship between sleep and human capital in a natural setting. Sleep's importance has garnered widespread attention in popular discourse, fueled by best-selling books (Walker, 2017; Huffington, 2016; Attia, 2023). However, much of the public conversation relies on limited evidence, often from small-scale studies, short-term laboratory experiments, or non-experimental investigations. Unlike previous work, we explore not only the direct causal effects of sleep interventions on sleep and performance but also their opportunity costs and mechanisms, leveraging data on time use, physical activity, well-being, and post-intervention outcomes.

Second, we examine the role for behavioral interventions in improving academic performance, particularly in higher education. The large body of prior work in this area has primarily targeted academic outcomes directly through either performance-based incentives, academic advising, or both. These policies have demonstrated limited impact and are often high cost (Lintner, 2024). We test whether a relatively low-cost intervention targeting sleep can have a downstream impact on educational performance and examine the mechanisms underlying our effects.

Finally, we investigate habit formation in sleep. To do so, we continue to track participants after the intervention ends and examine the persistence of sleep behavior post-intervention. In addition, we test variants of our primary intervention to identify the key features that drive habit formation.

We ran the experiment over seven semester-long waves from Spring 2019 to Spring 2022.

The 1,149 participants wore tracking devices (Fitbits) that measured sleep, heart rate, and physical activity; downloaded a custom smartphone app linked to their Fitbit data, which delivered our interventions; and, answered survey questions to capture information about their time use, performance on creativity and quantitative tasks, and well-being. The study included a one- to four-week baseline period, followed by a four-week intervention period and a one- to five-week post-intervention period (the post-intervention period lasts at least four weeks for 75% of participants).

Our intervention aimed to develop persistent sleep habits beyond the treatment period by leveraging theories of cue-based habit formation. This approach emphasizes the role of contextual cues, repetition, and immediate reinforcement of desired behaviors in generating "automatic" habits (Verplanken and Wood, 2006; Wood and Neal, 2007; Wood and Rünger, 2016). These theories suggest that repeatedly rewarding a behavior performed in response to a consistent cue can gradually strengthen the association between the cue and the reward that reinforces the action. Once the association is established, the cue may automatically trigger the desired action with little or no cognitive effort, even in the absence of a reward (Dickinson, 1985). Accordingly, we paired daily sleep cues with immediate rewards for meeting sleep goals, aiming to create an association between the cue and the desired sleep behavior during the intervention. In the post-intervention period, we retained the cue but removed the rewards to test whether sleep habits persisted in response to the cue alone.

Specifically, we set a goal for treated participants to sleep at least seven hours by 9 am on weeknights (Sunday - Thursday), following recommended guidelines (Panel et al., 2015). During the intervention period, participants received personalized bedtime reminders every weeknight, prompting them to follow a self-selected bedtime routine to get at least seven hours of sleep. On weekday mornings, they learned whether they had met their sleep goal and, if successful, received an immediate financial reward of $4.75. After the four-week intervention, we stopped the financial reward but maintained bedtime cues and morning feedback until the end of the semester (typically an additional four weeks). In secondary treatments, we tested variants of our primary intervention by providing rewards with a delay rather than immediately and by turning off either the rewards or the cue and feedback.

Our primary analysis compares the no intervention *Control* group to the *Immediate Incentives* group, in which participants received bedtime cues, morning feedback and immediate incentives for each weeknight they met the sleep goal during the intervention period.[1] At baseline, participants met the goal of sleeping at least seven hours on approximately 43% of

[1]The primary analysis pools two sub-treatments of the *Immediate Incentives* group: the main variation that continued to receive reminders and feedback in the post-treatment period and a variation that did not receive reminders and feedback in the post-treatment period. We do not find significant differences in the post-treatment effects of the two groups.

nights. During the treatment period, the intervention increases the rate of sleeping at least seven hours on weeknights by an estimated 11.5 percentage points ($p < 0.001$), 26 percent higher than baseline. The treatment effects persist into the post-intervention period but are smaller: an estimated 5.2 percentage points ($p < 0.001$), a 12 percent increase compared to baseline. We estimate that average weeknight sleep increases by 19 minutes during the treatment period and eight minutes during the post-treatment period ($p < 0.001$).[2]

Turning to the educational impact of sleep incentives, Immediate Incentives improve average semester course performance by an estimated 0.075 - 0.088 grade points ($p = 0.045$ and $p = 0.035$, respectively), a 0.10 - 0.11 standard deviation (SD) increase. We find evidence of similar sized treatment effects on grade point average in the semester following the intervention, but no impact two semesters after the intervention. Treatment effects are largest in classes that take place midday and, in exploratory analysis, are driven by impacts in STEM courses.

We next examine potential mechanisms for the impact of our intervention on academic performance. We find evidence that Immediate Incentives not only increase sleep hours via earlier bedtime but also sleep regularity, which persists after the intervention ends. Examining time use, the intervention leads to declines in self-reported screen time, which includes internet browsing, TV/videos and games, and excludes screen time for studying. The changes in screen time are similar in magnitude to the increases in sleep and are concentrated around bedtime. We find that total study time does not change during the intervention period, but there is suggestive evidence of a reallocation of study time from evening hours to morning hours, with little change in other time use. We do not find treatment effects on physical activity, or end-of-semester mental health, though we find evidence that treated participants report they are better able to cope with stress during the intervention period. We also find no evidence that the effects on performance are due to direct effects of receiving cash rewards. Together, our results suggest that incentives to sleep lead to more regular sleep habits, which displace screen time, and shift study time to earlier hours of the day, which may contribute to the improvement in academic performance.

Finally, we explore sleep habits and mechanisms of habit formation. As noted above, we find persistent but smaller increases in sleep in the post-intervention period, which lasts an average of four weeks. We also observed sustained improvements in sleep regularity, suggesting that the intervention led participants to establish more stable routines. However, we find little evidence that these routines are triggered by the intervention's external cue to go

[2]Our focus on weeknight sleep is in line with prior work that examines the impact of school and class start times, which occur on weekdays. We find no evidence of substitution between incentivized weeknight sleep and unincentivized sleep, including sleep that occurs during the day (i.e., naps), on weekends and during holidays.

to bed earlier, as there is no lasting impact on average bedtimes. Comparing our primary Immediate Incentives intervention to secondary treatments, we find that immediate incentives significantly outperform variants with delayed or no rewards during the intervention period. However, in the post-intervention period, the effects of the primary and secondary treatments are not statistically distinguishable. Together, our results suggest that interventions targeting sleep can shift bedtime habits, though it is less clear whether these habits become truly "automatic."

Our study is the first to show that an intervention targeting sleep can improve academic performance. Our findings contribute to the large literature on improving educational outcomes, particularly among college students (e.g., Angrist et al., 2014; Lavecchia et al., 2016, provide reviews). Compared to previously examined policies, such as financial aid, mentoring and support services, and performance-based incentives, our intervention is highly cost-effective. For example, in a recent meta-analysis, Lintner (2024) estimates that performance-based financial incentives improve college students' grade point average (GPA) by an average of 0.041 grade points with average costs of about $460 per student per semester. By comparison, we estimate that Immediate Incentives to sleep improve GPA by about twice the effect size (0.075 - 0.088 grade points) at less than one-quarter of the cost (about $112 including the cost of the Fitbit). Our results suggest that improving academic achievement through sleep interventions may be more cost-effective than incentivizing performance directly. This finding is akin to recent work showing that incentives for exercise can improve educational achievement (Cappelen et al., 2017).

We also contribute to the growing literature on the economics of sleep. Seminal studies using time-use surveys find a negative association between sleep and work hours, though they do not establish causality (Biddle and Hamermesh, 1990; Basner et al., 2007).[3] Studies using naturally occurring data reveal that later sunset times are associated with lower cognitive performance, poorer educational outcomes, reduced earnings, and worse physical and mental health. These effects are primarily attributed to reduced sleep (Giuntella et al., 2017; Gibson and Shrader, 2018; Giuntella and Mazzonna, 2019; Jin and Ziebarth, 2020; Lindquist and Sadoff, 2023, provide a review). In the context of education, U.S.-based estimates suggest that a one-hour later shift in sunrise or class start time increases sleep by an average of 6 - 36 minutes and has either no discernible impact on academic achievement or leads to an improvement in grades and test scores by 0.06 - 0.16 SD (Carrell et al., 2011; Heissel and Norris, 2018; Groen and Pabilonia, 2019). The causal impact of our intervention aligns with

[3] Related work finds a positive correlation between sleep and health outcomes (Cappuccio et al., 2010), as well as sleep and academic performance (Creswell et al., 2023). Other studies show that sleep deprivation affects decision-making, ethical behavior, social decisions, and voting behavior (Dickinson and McElroy, 2017; Dickinson and Masclet, 2023; McKenna et al., 2007; Holbein et al., 2019).

these findings from natural experiments: it increases weeknight sleep by an estimated 19 minutes during treatment and eight minutes post-treatment, and grades by 0.10 - 0.11 SD.

The non-experimental studies do not exogenously vary sleep or test policies aimed at improving sleep habits. Moreover, as illustrated in Figure A.1, there is a dearth of well-powered experimental studies that measure the impact of sleep on either labor market or academic performance.[4] Sleep laboratory studies find that short-term severe sleep deprivation worsens cognition and mood, but cannot estimate the effect of moderate sleep improvements in natural settings (Banks and Dinges, 2007; Killgore, 2010; Lim and Dinges, 2010). Avery et al. (2022), Barnes et al. (2017) and Breig et al. (2020) examine sleep in field experimental contexts but do not link sleep to productivity and performance. The only prior field experiment testing the impact of sleep interventions on performance does not find an impact of increased nighttime sleep on productivity among highly sleep-deprived workers in India (Bessone et al., 2021). By contrast, in our study, as in the broader U.S. context, average sleep is of higher quality and baseline sleep is closer to the recommended minimum of seven hours a night. Our study also takes place over a longer time horizon, which may better capture the effects of moderate, sustained increases in sleep on performance.

Lastly, we contribute to the literature on habit formation by studying interventions that combine cues, feedback, and rewards, which prior work has largely examined separately. For example, Wellsjo (2021) and Byrne et al. (2022) focus on cues and feedback, while studies leveraging incentives typically distribute rewards with a delay (Gneezy et al., 2011; Royer et al., 2015; Hussam et al., 2022; Beshears et al., 2021; Milkman et al., 2014). We show that making rewards immediate significantly increases their impact on repeated behaviors, extending prior findings that immediate rewards outperform delayed rewards for one-time actions (Levitt et al., 2016). Finally, our examination of habit formation suggests that participants establish their own routines rather than relying on externally provided bedtime cues. These results align with prior research showing that interventions allowing for flexibility tend to be more effective (Beshears et al., 2021).

The remainder of the paper is structured as follows. In Section 2, we describe the experimental design, data and analysis. Section 3 presents the the treatment effects on sleep hours and academic performance. Section 4 examines mechanisms for the effect of our intervention on educational outcomes. Section 5 discusses mechanisms for habit formation in sleep. Section 6 benchmarks the results relative to prior findings, and Section 7 concludes.

[4]The figure summarizes 123 studies referenced in books and leading articles on sleep (Attia, 2023; Huffington, 2016; Bessone et al., 2021; Creswell et al., 2023) including their sample size, citation count, study type (field experiment, lab experiment, non-experimental) and outcome measures (whether they include educational and/or labor market outcomes).

2 Experimental design and data

We conducted our experiment in seven semester-long waves from Spring 2019 to Spring 2022 among students at the University of Pittsburgh (Pitt). We measured sleep using wearable trackers and delivered our interventions targeting sleep via text messages and a custom smartphone app. Our outcome data come from the wearable trackers (sleep and physical activity), survey measures (time use, performance on creativity and quantitative tasks, and well-being); and administrative records (academic transcripts).

2.1 Wearable trackers and custom smartphone app

To gather objective measures of sleep in a natural setting, we had participants in our study wear Fitbits that estimate sleep patterns based on movement and heart rate data. The use of such wearable trackers allowed us to depart from dependence on sleep diary methods, which have been shown to significantly overestimate sleep (Lauderdale et al., 2008; Bessone et al., 2021). Fitbits, which are among the most popular wearable trackers, are well-suited for monitoring sleep in natural settings due to their portability and unobtrusiveness, and are the most utilized wearables for biomedical research purposes (Wright et al., 2017). In our study, we used Fitbit Charge HR, Charge HR 2, Charge HR 3, Alta HR, and Inspire 2, which all capture both movement and heart rate. Recent studies have validated the accuracy of heart rate-enabled Fitbits for population-based sleep studies when compared to actigraphy, a widely used method for outpatient sleep screening (Haghayegh et al., 2019). In the Online Appendix (Giuntella et al., 2024), we discuss in more detail the limitations and advantages of wearables like Fitbits to capture detailed sleep metrics.

One source of concern with studies that rely on wearable trackers is that the devices require continued engagement via daily syncing (i.e., regularly connecting the tracker to a smartphone to update the collected data). To ensure high sync levels for our study, we developed a custom-made smartphone app that connected to the Fitbit API, which allowed us to monitor sync rates daily and notify participants with low sync rates in order to keep them engaged. The custom-made app also allowed us to deliver our interventions to improve sleep habits via push notifications and the app itself. The app features included the ability to send reminders; provide immediate individualized feedback based on participants' sleep as measured by the Fitbit; and redeem rewards. We discuss the interventions in more detail in Section 2.3

2.2 Sample, recruitment and timeline of the experiment

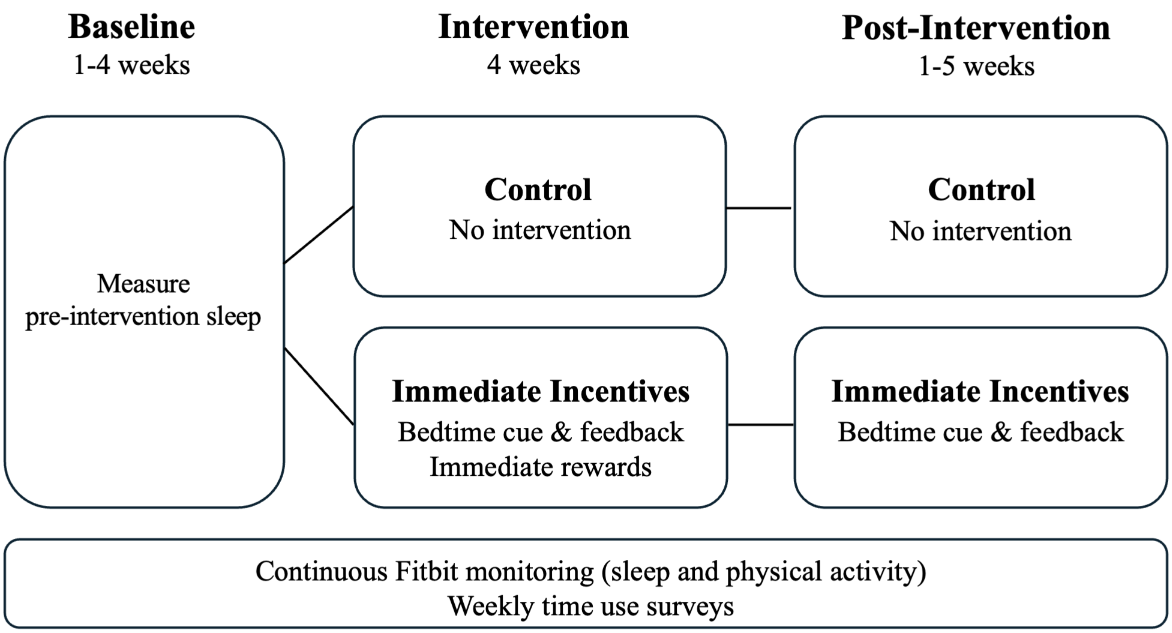
The experiment took place at the University of Pittsburgh, was approved by the University of Pittsburgh Institutional Review Board and was pre-registered in the AEA RCT registry (AEARCTR-0003235). We discuss deviations from the pre-registration plan in Appendix C.

We recruited participants through the Pittsburgh Experimental Economics Laboratory (PEEL) and invited them to participate in a semester-long study on wellness for a guaranteed minimum payment of $30 by the end of the semester, including a $6 upfront payment for the initial enrollment session, and the opportunity to receive additional earnings based on luck, as well as their choices during the experiment. To be eligible for our study, participants had to have a smartphone and be willing to wear and routinely synchronize a wearable device (Fitbit) during the semester. We began the experiment in Spring 2019 and enrolled participants every semester (Fall and Spring) until Spring 2022. We ran the experiment in seven consecutive waves, with modest-sized cohorts to accommodate the number of participants we could recruit through the lab every semester, as well as the number of Fitbits we had. Our final sample includes 1,149 participants.

Figure 1 shows the timeline of the study. In each wave, the study lasted for approximately 10 weeks. We initiated participant recruitment in the first few weeks of the semester and enrolled participants in the experiment on a rolling basis. Upon enrollment, we measured baseline sleep for one to four weeks. In our analysis, we restrict the baseline to the two weeks before the start of the intervention because fewer than half of subjects have more than two weeks of pre-intervention data. At the end of the baseline period, we randomly assigned participants to either a control group or treatments designed to improve sleep habits, which lasted for four weeks (*intervention period*). After the 4-week intervention period, we continued to follow participants for an additional 1-5 weeks until the end of the semester (*post-intervention period*), at which point we asked them to return the Fitbit and fill out an endline survey. The median number of weeks in the post-intervention period is four and 75% of participants have at least four weeks of post-intervention data. The study always ended during the last week of classes, before final exams. Across different waves, the start of the recruitment period depended on laboratory availability. This affected the timing of the treatment period and the length of the post-treatment period, which varied in each wave due to differences in recruiting start dates, the speed of participant enrollment, and the semester schedule. Importantly, the treatment itself was always four weeks long. The timeline of the experiment for each of the seven waves is depicted in Appendix Figure A.2.[5]

[5] In Fall 2019, due to recruitment issues at PEEL, we recruited two groups of participants and had them start the intervention in a staggered way, as shown in Figure A.2. In Spring 2020, the semester schedule was changed by the university closure prompted by the onset of the COVID-19 pandemic. Students enrolled in

Figure 1: Timeline of the experiment



To enroll in the study, participants completed an initial session at the laboratory (Spring 2019 - Spring 2020) or over Zoom (Fall 2020 - Spring 2022), completed an intake survey, received the Fitbit and installed our custom smartphone app.[6] During the intake session, participants consented to wear and sync the Fitbit throughout the semester, answer weekly surveys, and grant us access to their academic records. They were informed about their right to withdraw from the study at any time with no penalty. Participants left the initial session with a one-page reminder outlining what was expected of them during the study and agreed to return the Fitbit at the end of the semester. The intake survey administered to participants

the study in Spring 2020 learned about the university moving to remote learning during spring break (mid-March 2020), and continued to stay enrolled in the study until the end of the semester. In the Appendix, we conduct sensitivity analyses that exclude the Spring 2020 wave.

[6]From Fall 2020 onwards, we adjusted some of the intake procedures due to changes in the lab and university protocols during the COVID-19 pandemic.In March 2020, Pitt shifted to remote instruction for the remainder of the term due to the onset of the COVID-19 pandemic. In Fall 2020, Pitt used a "Flex@Pitt" model, combining in-person and remote instruction with the option for students to attend fully online depending on their preferences and safety concerns. In-person activities were limited. In Spring 2021, The Flex@Pitt model continued, with most courses remaining hybrid or fully online. Instead of filling out one unique survey during the intake session, the survey was split into an enrollment survey that participants filled out at enrollment while on Zoom and a follow-up survey that was emailed to them a few days later. In Spring 2019-Spring 2020 and Fall 2021-Spring 2022, participants picked up the Fitbit from PEEL and received a $6 payment. In Fall 2020 and Spring 2021, participants received the Fitbit via mail as the University of Pittsburgh combined in-person and remote instruction with the option for students to attend remotely depending on their preferences and safety concerns.

collected information on socio-demographic characteristics and baseline measures of well-being.

Over the course of the study, participants in all treatments received reminders to sync their Fitbit via text message and the app (see Figure B.5). They also received weekly surveys that captured time use, performance on creativity and quantitative tasks, and well-being. We describe the survey measures in more detail below (Section 2.4).

2.3 Treatments

In total, 1,219 individuals completed an enrollment survey. In order to be randomized, participants had to have at least one day of Fitbit data in the baseline period.[7] At the end of the baseline period, we randomized 1,149 participants to treatment groups, which are displayed in Table 1. Participants in the *Control* group ($N = 380$, waves 1-7), received no intervention and continued to wear their Fitbits and fill out surveys until the end of the semester. Participants in the treatment groups received interventions to improve sleep habits.

In all treatments, we set the goal of sleeping at least seven hours per night by 9 am on weeknights (Sunday through Thursday). We established 9 am as a key constraint, based on previous studies emphasizing the significance of sleep timing and the alignment of biological rhythms with the environmental light-dark cycle (Roenneberg and Merrow, 2016). Additionally, we aimed to reduce the likelihood that our intervention would encourage skipping classes scheduled at 9 am. Notably, about 80% of the participants wake up before 9 am at baseline.

Drawing on the habit formation framework outlined earlier, our *Immediate Incentives* intervention (468 participants, waves 1-7) leverages cues, rewards, and repetition to establish persistent sleep habits. To provide participants with a consistent cue, we sent them reminders—both through the app and via text—to meet their target goal of sleeping seven hours per night by 9 am every weeknight (Sunday - Thursday). These reminders had two major components. We established a personalized target bedtime for each participant, an hour earlier than their usual baseline bedtime, based on their individual sleep patterns, and sent reminders to go to bed half an hour before this new goal time. Figure B.1 displays the bedtime reminder.[8] Second, as the cue-based framework emphasizes the importance of a

[7] We mistakenly assigned eight participants to treatments who did not have any baseline Fitbit data. We conduct sensitivity analyses that exclude these individuals and results do not change (see Section 3).

[8] The bedtime was set approximately an hour before participants' average baseline bedtime rounded to the nearest 30 minutes with a latest bedtime goal of 1 am (e.g., for participants with an average baseline

Table 1: Treatments

Main Treatments	N	Waves	**Treatment** Reminders & Feedback	Rewards	**Post-Treatment** Reminders & Feedback
Control	380	1-7	–	–	–
Immediate Incentives	468	1-7	✓	Immediate	✓
Immediate Incentives, Post Cue/Feedback	356	1-7	✓	Immediate	✓
Immediate Incentives, No Post Cue/Feedback	112	5,7	✓	Immediate	–
Total:	848				
Secondary Treatments					
Delayed Incentives	103	1-3	✓	Delayed	✓
Delayed Incentives, No Cue/Feedback	97	1-3	–	Delayed	–
Cue/Feedback	101	1-3	✓	–	✓
Total	1,149				

Notes: The table reports the number of participants enrolled in each of the treatments; whether rewards were immediate or delayed, and whether they received reminders and feedback during and after the intervention. Immediate Incentives pools Immediate Incentives, Post Cue/Feedback and Immediate Incentives, No Post Cue/Feedback.

stable environment in triggering automatic behavior, we encouraged participants to engage in a specific bedtime behavior every weeknight before going to sleep. Participants selected their behavior from a menu of different options before the beginning of the intervention period. Examples included "Turn off your Phone", "Turn on bedtime music", "Turn off your computer", "Turn on meditation app".[9]

Next, to link sleeping behavior with a reward, we provided participants with immediate financial incentives upon meeting their sleep goal. Every weekday after 9 am, participants received feedback on whether they met their goal of sleeping at least seven hours via the app through push notifications and the app interface. Participants who met their goal received feedback about having achieved the goal and earned a $4.75 reward, which they redeemed by clicking a button on the app (see Figure B.2).[10] Participants in this treatment received

bedtime of 12-12:14 am, we set a goal bedtime of 11 pm; for participants with an average baseline bedtime of 12:15-12:30 am, we set a goal bedtime of 11:30 pm; for participants with an average baseline bedtime of 2 am or later, we set a goal bedtime of 1 am). We delivered the personalized bedtime reminder via text message. In the app, we delivered a standard message encouraging participants to go to bed early enough to sleep seven hours by 9 am.

[9]On the Friday before the beginning of the intervention period, participants received their intervention-related instructions. As part of these instructions, we asked participants to select a bedtime behavior to engage in before going to bed.

[10]Redemption rates were above 95% across waves.

a monetary reward of $4.75 through a Venmo transfer on the same day.[11] Participants who fell short of the sleep target were given feedback indicating that they had not achieved their goal and had missed out on receiving the reward. This feedback also included a negatively-valenced emoji to convey the injunctive message that sleeping less than seven hours was discouraged (see, e.g., Schultz et al., 2007), and encouragement to make another attempt to meet the sleep target. To encourage repetition of the incentivized sleep behavior, cues and rewards continued every weeknight and weekday of the four-week intervention period. In waves in which the treatment period spanned spring break, we paused the intervention during spring break. Results are unchanged when we restrict the analysis to terms where the intervention was not paused (see Online Appendix (Giuntella et al., 2024)).

At the end of the intervention period, we discontinued the financial rewards and notified participants via text message. In our main variation of the Immediate Incentives treatment, *Immediate Incentives – Post Cue/Feedback* ($N = 356$, waves 1-7), we continued sending participants bedtime reminders (i.e., the cue) and morning feedback on whether they had achieved their sleep goal throughout the post-intervention period, which lasted until the end of classes. The feedback remained identical to that of the intervention period, except that we removed mention of the financial reward, as displayed in Figure B.4. To examine the importance of maintaining the cue for habit persistence, we tested a variant of the Immediate Incentives treatment in which participants did not receive cues and feedback in the post-intervention period, *Immediate Incentives – No Post-Cue/Feedback* ($N = 112$, waves 5 and 7). Our primary analysis pools the two variants of Immediate Incentives. In waves in which the post-treatment period spanned Thanksgiving, we paused reminders and feedback during the week of Thanksgiving. Our primary analysis excludes holiday weeks (spring break and Thanksgiving).

Secondary treatments. Following our pre-analysis plan, in the first three waves of the study, we implemented secondary treatments to examine the importance of cues and feedback, and the timing of financial incentives for habit formation.

In the *Delayed Incentives* treatment ($N = 103$, waves 1-3), we provided participants with cues, feedback and rewards, as in the Immediate Incentives treatment. The only difference was that, although feedback about receiving the incentive was immediate, payment was not. Participants learned each day whether they met their sleep goal, but only received a single transfer with the total payment at the end of the study period, one to five weeks after

[11] For logistical reasons, the payment was received after 3 pm each day, which introduced a small delay between the performance of the behavior and the reward. However, the feedback about receiving a reward was provided as soon as participants synced the Fitbit after 9 am.

the intervention ended. Figure B.3 displays the feedback screens for the Delayed Incentives treatment. In the post-intervention period, participants continued to receive the bedtime cue and morning feedback, as in the Immediate Incentives treatment. This treatment aimed to test whether providing repeated immediate rewards increases their effectiveness during treatment compared to delayed rewards, as has been shown with one-time rewards (Levitt et al., 2016); and whether reinforcing behavior with immediate rewards enhances persistence of behavior after the reward is removed, more so than delayed rewards.

In the *Delayed Incentives – No Cue/Feedback* treatment ($N = 97$, waves 1-3), we removed the bedtime reminders (i.e., the cue) and the daily feedback about whether participants met their sleep goal but retained the financial incentive. At the start of the intervention period, we informed participants that they would receive $4.75 for every night they met the goal of sleeping at least seven hours by 9 am, with payment to be received via a Venmo transfer at the end of the semester. The participants did not receive reminders or feedback during the post-intervention period. This treatment is analogous to other work using financial incentives to create habits in the context of exercising (e.g., Charness and Gneezy, 2009; Royer et al., 2015) and aimed to test the importance of pairing cues and feedback with rewards.[12]

To test whether financial rewards are critical for establishing habits, we also conducted an additional treatment, *Cue/Feedback* ($N = 101$, waves 1-3), where we removed the financial reward. Participants in this treatment received the same bedtime reminders as participants in the Immediate Incentives and Delayed Incentives treatments. They also received daily feedback via the app on whether they had achieved their sleep goal. Instead of providing participants with a reward, the feedback screen included a positively- or negatively-valenced emoji depending on whether participants had achieved their sleep goal – i.e., the same feedback that the Immediate and Delayed incentives groups received in the post-intervention period (see Figure B.4). The participants in the Cue/Feedback treatment continued to receive the same reminders and feedback throughout the post-intervention period.

2.4 Data

2.4.1 Sleep

Our primary pre-registered sleep outcome is the share of weeknights (Sunday - Thursday) participants sleep at least seven hours. Our secondary measures of sleep include sleep hours per night, sleeping seven hours per night and sleeping between seven and nine hours without

[12]We did not test immediate incentives without providing cue/feedback because delivering immediate rewards inherently requires providing feedback on whether the goal was met. However, in the post-intervention period, we tested the effect of removing the cue/feedback for immediate incentives and, as we note, find little evidence that the post-cue/feedback had any significant impact (Appendix Table A.3 panel B, column 1).

restricting to weeknights, sleeping seven hours per night including naps, bedtime, wake-up time, sleep regularity, and sleep quality as measured by the Fitbit. In exploratory analyses, we also analyze sleeping between seven and nine hours on weeknights and sleeping at least six hours on weeknights in order to better compare these outcomes to our primary outcome measure. As a secondary measure, we also measured self-reported sleep and desired sleep in the intake survey.

We study sleep regularity in two ways. First, we use the sleep regularity index (SRI) (Fischer et al., 2021; Phillips et al., 2017; Windred et al., 2024), which measures how similar a person's sleep-wake cycles are from one day to the next using binary classifications of sleep and wake states at the minute-level. The SRI is a percentage score that ranges from 0 to 100, where higher values indicate more regular sleep patterns. It is calculated based on the presence or absence of sleep at each minute interval over a 24-hour period, comparing day-to-day variations. Second, we focus on a measure of variation or dispersion across the week, as measured by the within-person standard deviation in the outcome of interest (Fischer et al., 2021).

We measure sleep quality in terms of efficiency, Rapid Eye Movement (REM) sleep and deep sleep. Efficiency measures the percentage of time in bed that an individual is asleep. REM sleep is the stage of sleep in which individuals dream, which stimulates areas of the brain essential to learning. During REM sleep, heart rate and blood pressure rise. Studies suggest that REM sleep plays a key role in memory consolidation, emotional processing, and brain development (Marks et al., 1995; Boyce et al., 2016). Deep sleep is the most restorative form of sleep. During deep sleep the heart rate and breathing rate are at their lowest and our body repairs tissue. Deep sleep is important for regulating glucose metabolism and has also been linked to cognitive function and memory (Zhang and Gruber, 2019; Leproult and Van Cauter, 2010). We caution that while, as described above, there is growing evidence on the performance of recent Fitbit models in accurately measuring sleep duration (de Zambotti et al., 2018), the accuracy and reliability of these devices in capturing sleep stages needs further validation. In particular, sleep trackers have acceptable sensitivity but poor specificity when compared with sleep stages obtained using polysomnography (PSG). Given these limitations, sleep quality measures should be interpreted with caution. Furthermore, short naps (less than an hour) are not accurately recorded by Fitbit devices.

Sync rates throughout the study were relatively high. On average participants synced their devices for 85% of the days. As shown in Table A.2, there are higher sync rates during the intervention period among the Immediate Incentives group compared to the Control group (there are no significant differences in sync rates in the baseline and post-intervention periods). As noted above, 8 of the 1149 participants did not have any Fitbit data, including at baseline. Of the 1141 participants with baseline data, 23 did not report any data during treatment (2%) and 83 (7%) did not report any data in the post-intervention period.[13] In all of our analyses, for the nights with missing data, we replace missing data with an individual's baseline average following the approach of Bachireddy et al. (2019). We also conduct sensitivity analyses that do not replace missing data and report Lee (2009) Bounds for treatment and post-intervention effects. Results do not meaningfully change (Table A.3).

One concern with using Fitbits is that participants might lend their Fitbit to someone else for several nights to manipulate the reward system. To mitigate this potential concern, we make use of the resting heart rate data collected by the Fitbit. Research indicates that although resting heart rate can vary greatly between different people, it usually remains fairly stable within the same individual over time (Quer et al., 2020). We do not find a significant link between unusual fluctuations in resting heart rate (deviating more than one standard deviations from baseline) and treatment assignment (p-value$= 0.73$). We also find no evidence of data manipulation, such as heaping, rounding in recorded sleep times or manual logging of sleep (see Online Appendix for further discussion on wearables and measurement error (Giuntella et al., 2024)).

2.4.2 Educational outcomes

Our primary educational outcome is term Grade Point Average (GPA), measured using administrative data obtained on September 14, 2023. The Registrar's Office at the University of Pittsburgh supplied us with course data for participants enrolled in our experiment, covering each semester of their enrollment at the university from Fall 2018 through Spring 2023.[14]

This dataset provides comprehensive course information for our experimental participants across three key periods. In addition to data from the term during which our intervention took place (our primary outcome measure), it includes information from both before and after the intervention. Specifically, we have data for the term immediately preceding the intervention if participants were still enrolled at Pitt, as well as their high school GPA. We also have data for at least two terms following the intervention. Access to this post-intervention data allows us to conduct exploratory analyses on the intervention's longer-term impact, provided the student enrolled in graded classes during subsequent semesters.

We calculate term GPA based on all courses in which a student received a letter grade,

[13] Differences in syncing rates between the secondary treatments and the control group are reported in the Online Appendix (Giuntella et al., 2024).

[14] Because the administrative data was obtained after the intervention was concluded, we did not know the distribution of baseline GPA or GPA by course type in advance of the study. At the time of the pre-registration, we also did not know how many semesters of data and which secondary administrative data on educational outcomes would eventually be made available to us at study completion.

A+ through F, converted to a 4-point scale (see Table A.1 for the grading system). Term GPA is an average of the course grade points, weighted by the number of credits for each course. Our secondary outcome measures include course completion and credits earned; we also collected exploratory measures of withdrawals, course failures and course pass rates. Eighty-eight percent of the courses in our data receive a letter grade on the 4-point scale. Our main analysis excludes courses with grades outside this scale. We include these courses when examining the likelihood of having any grade, number of credits earned, as well as in the exploratory analysis examining course withdrawal, failure, and pass rates.

Of the 1,149 participants in our experiment, 1,128 have at least one course grade for the term of the intervention (98% of the sample). The 21 remaining participants had no available grades for the term of the intervention, but have academic records in other terms. Our analysis includes all available grade data for the relevant term. For 1,056 participants (92% of the sample) we have data on at least one course grade for one term after the intervention; for 969 participants (84% of the sample) we have at least one course grade for two terms after the intervention. From the Registrar data on participants' high school GPA, we could match 1,049 students (91% of the sample). For 74 of the 100 students with no high-school GPA, we have information on baseline GPA at the start of the term (cumulative GPA from all prior terms). This gives a total of 1,123 participants with baseline GPA (98% of our sample).

The match rates are similar if we limit the data to our primary analysis comparing the Control and Immediate Incentives groups. Of these 848 participants, we match 833 to course grades in the term of the intervention (98%), 784 to high school GPA (92%) and an additional 41 to a prior term GPA. We construct a baseline GPA variable from either prior semester GPA or, when not available, high school GPA. A total of 825 participants have a baseline GPA (97%). As shown in Table A.2, we have a higher proportion of baseline grades for the Immediate Incentives group than for the Control group. In sensitivity analysis, we limit the sample to participants with baseline GPA and results are similar (Table A.6). There is no difference between the groups in the likelihood of having course grades, which is our primary outcome measure (Table A.2).

The course data allow us to classify courses by class type and start time. Class types include lectures, seminars, credit laboratories, practicums, workshops, independent studies, directed studies, internships, and laboratories. Lectures comprise 80% of the classes. As shown in Figure A.6, non-lecture classes have significantly higher grades and lower variance than lecture classes. The average GPA (and standard deviation) in lectures is 3.44 (0.81) compared to 3.75 (0.51) in other classes. In lecture classes, 47% of students receive the highest possible grade and the lowest quartile is a B. By comparison, in non-lecture classes, 67% of students receive the highest possible grade and the lowest quartile is an A-. This raises concerns that the grading system in non-lecture courses leaves little scope for treatment effects. In our analyses, we therefore report estimated treatment effects for all course types together, as pre-registered, and an exploratory analysis with lectures alone (we report the effects for non-lecture courses in Table A.6).

In exploratory analysis, we also classify each course as STEM or non-STEM using the Department of Homeland Security 2023 list of STEM-designated CIP codes.[15] The average GPA (and standard deviation) in STEM classes is 3.26 (0.88), while it is 3.69 (.58) in non-STEM classes. In STEM classes, 39% of students receive the highest possible grade and the lowest quartile is a B. By comparison, in non-STEM classes, 61% of students receive the highest possible grade and the lowest quartile is an A-. Similar to non-lecture courses, the grading in non-STEM courses may limit the scope for treatment effects.

2.4.3 Additional outcomes

Time use. We implemented a time use survey once a week, rotating the weekday on which the survey was administered. Our time use measure follows the structure of American Time Use Survey (Abraham and Flood, 2009). From a drop-down menu, participants indicated how they allocated their time on the previous day. For each 30 minute interval over the course of 24 hours, participants could choose from a list of activities that included sleeping, grooming (self), watching TV/videos, surfing the internet, playing games, working, studying, preparing meals or snacks, eating or drinking, cleaning, laundry, grocery shopping, attending religious services, hanging out with friends, paying bills, exercising, commuting, or other activities. They could also indicate that they did not know or could not remember how they spent their time, or could refuse to respond. In our primary analysis, we examine "screen" time, which pools time spent watching TV/videos, surfing the internet and playing games and excludes screen time spent studying; we categorize time spent hanging out with friends as "social" time. We exclude from the analysis responses that report 24 hours of "other activities," which may reflect inattention in filling out the time use survey.

Creativity and quantitative tasks. We collected secondary measures of performance through creativity and quantitative tasks. We drew the math questions from the math section of the Graduate Record Examination (GRE) test. We measured creativity using an adapted version of the task employed by Charness and Grieco (2019), where we provided participants with a list of 10 words and asked them to use some or all of the words to write an interesting sentence. On alternate weeks, the weekly time use survey included either

[15] https://nces.ed.gov/ipeds/cipcode/Files/2023/Final-2023-CIP-STEM-List-Blog.pdf

one multiple choice math question or one creativity task. Both tasks were incentivized (see instructions in Appendix B). To assess the creativity task, we recruited raters from lab participants at the University of California San Diego and from Prolific ($N = 1,369$), and four undergraduate research assistants at the PEEL lab at the University of Pittsburgh. Raters received a random subset of the sentences produced by participants in the creativity task and rated them on a 1-5 scale. Each sentence was rated by a minimum of two raters; the median number of ratings per sentence is thirteen.

Physical health. From the Fitbits, we collected data on resting heart rate and physical activity (daily steps and active minutes). Resting heart rate measures heart beats per minute (BPM) at rest (i.e., when sitting, lying down or relaxing). Faster resting heart rates are associated with shorter life expectancy (Cooney et al., 2010; Dyer et al., 1980). Daily steps are the number of steps over the course of a 24 hour-period. Active minutes are measured as minutes in which a person is non-sedentary for a least 10 continuous minutes, where non-sedentary is defined as activity that raises heart rate enough to burn at least 3 metabolic equivalents (METs).[16]

Well-being. We collected measures of mental health in the intake survey (conducted upon enrollment) and in the endline survey at the end of the semester. We assessed depression using the Center for Epidemiologic Studies Depression scale (CES-D, Radloff, 1977), which is a 20-item validated instrument designed to assess the frequency of depressive symptoms on a scale from 0 ("Rarely or none of the time") to 3 ("Most or almost all the time"). An overall depression score is calculated by summing answers to all 20 items, with total scores ranging from 0 - 60. We also measured anxiety using the Generalized Anxiety Disorder scale (GAD-7, Williams, 2014), a 7-item scale designed to assess symptoms of Generalized Anxiety Disorder. The instrument assesses the frequency of anxiety-related symptoms using a scale ranging from 0 ("Not at all") to 3 ("Nearly every day"), with total scores ranging from 0 - 21. To measure well-being, we collected exploratory measures of mood, stress and ability to cope with stress (resilience). For mood, we asked participants to indicate, on a 10-point Likert scale, how happy they felt in that moment. For stress and resilience, participants indicated, using a 5-point Likert scale, 1) the extent to which they faced stress in their life at the time of answering the survey and 2) the extent to which they felt able to deal with the stress they were facing. Every week, we alternated between the mood and the stress/resilience questions. These measures were collected via text message and, each week,

[16] In practice, this measure sums the "lightly active", "fairly active" and "very active" minutes collected by the Fitbit.

participants were randomly assigned to receive the text message at different times of the day (11 am, 4 pm, 9 pm).

As shown in Table A.2, there is no difference in attrition rates between the Immediate Incentives group and the Control group for the additional outcomes discussed above.

2.5 Randomization and baseline characteristics

Randomization occurred at the end of the baseline period, the weekend before the start of the intervention period. We employed a block randomized design, stratifying our participants by gender and the share of weeknights participants slept more than seven hours (above vs below median).[17] In the initial waves of the study (Spring 2019 to Spring 2020) we randomized participants to one of five groups with equal probability (Control, Immediate Incentives, Delayed Incentives, Delayed Incentives No Cue/Feedback and Cue/Feedback). For the remaining waves, we randomly assigned participants to either the Control group or the Immediate Incentives treatment. In Waves 5 and 7 (Spring 2021 and 2022), we randomized participants in the Immediate Incentives treatment to either receive or not receive cue and feedback during the post-intervention period (Immediate Incentives - Post Cue/Feedback or Immediate Incentives - No Post Cue/Feedback).

Table 2 compares baseline characteristics in the Control group (column 1) to the Immediate Incentives group (column 2). We report demographic characteristics, baseline sleep behaviors and baseline academic characteristics (we discuss the baseline sleep in Section 3.1).

Students in the Control group are on average about 19 years old, with a large share of freshmen (52%). Sophomore, junior, and senior and above students make up 12%, 23%, and 12% of the Control group, respectively. Female and Asian students are over-represented compared to the full-time Pitt student population, and the U.S. college population in general. Approximately 55% (58%) of Pitt (U.S.) students are women, while women make up 72% of the Control group. Asian students make up 11% (7%) of the Pitt (U.S.) student population, while they represent 28% of the Control group. White students, which make up 56% of the Control group, are slightly under-represented compared to the Pitt student population (68%) and slightly over-represented compared to the U.S. college population (52%). The share of Black (8.8%) and Hispanic (4.0%) students is representative of the Pitt student population but lower than the U.S. college population, in which 13% of students are Black and 22% are Hispanic.[18] A quarter of the students in the Control group report their parents

[17] We did not balance the randomization on baseline GPA because, as discussed above, GPA data was not available at the time of randomization.

[18] Demographics for the 2021-22 U.S. college population are available at: https://www.statista.com/statistics/236360/undergraduate-enrollment-in-us-by-gender, accessed on November 18 2023. Demographics for the Pitt student population in 2021-22 are available at: https://www.ir.pitt.edu/sites/

Table 2: Treatment-Control differences in baseline characteristics, Immediate Incentives

Variable	Control	Immediate Incentives	Difference
Demographics			
Female	0.721	0.726	-0.002
	(0.449)	(0.446)	(0.031)
Age	19.463	19.344	-0.110
	(2.982)	(1.964)	(0.170)
White	0.548	0.568	0.023
	(0.498)	(0.496)	(0.034)
Asian	0.285	0.261	-0.017
	(0.452)	(0.439)	(0.031)
Black	0.088	0.068	-0.021
	(0.283)	(0.253)	(0.019)
Hispanic	0.040	0.053	0.010
	(0.196)	(0.225)	(0.015)
Other	0.040	0.049	0.005
	(0.196)	(0.216)	(0.014)
Highest parent educ:			
less than college	0.255	0.284	0.028
	(0.437)	(0.452)	(0.031)
college	0.287	0.288	0.001
	(0.453)	(0.454)	(0.032)
more than college	0.447	0.427	-0.018
	(0.498)	(0.495)	(0.034)
Baseline sleep outcomes			
Sleep hours	6.625	6.659	0.012
	(0.958)	(0.902)	(0.064)
Sleep ≥ 7	0.438	0.426	-0.020
hours	(0.276)	(0.274)	(0.019)
Sleep ≥ 6	0.706	0.713	0.001
hours	(0.258)	(0.258)	(0.018)
Bedtime	25.265	25.211	-0.073
	(1.313)	(1.297)	(0.091)
Wake up time	7.956	7.935	-0.062
	(1.302)	(1.238)	(0.086)
Baseline academic characteristics			
Freshman	0.521	0.530	-0.006
	(0.500)	(0.500)	(0.031)
Sophomore	0.118	0.120	0.009
	(0.324)	(0.325)	(0.022)
Junior	0.226	0.212	-0.012
	(0.419)	(0.409)	(0.028)
Senior and above	0.124	0.139	0.020
	(0.330)	(0.346)	(0.023)
STEM major	0.582	0.571	-0.004
	(0.494)	(0.496)	(0.035)
Number of courses	5.167	5.158	-0.038
	(1.282)	(1.420)	(0.095)
Number of early sessions	1.523	1.667	0.166
	(1.562)	(1.545)	(0.109)
High-School GPA	4.140	4.131	-0.011
	(0.440)	(0.434)	(0.032)
Baseline term GPA	3.429	3.457	0.016
	(0.530)	(0.465)	(0.038)
Joint p-value			0.881
Observations	380	468	848

Notes: The sample is restricted to individuals in the Control and Immediate Incentives treatment groups. Early class sessions are classes starting at 10 am or earlier. All estimates in column 3 include wave fixed effects. Robust standard errors are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1.

did not receive a college degree (either some or no college); 29% report that at least one of their parents has a college degree; and 45% report that at least one of their parents has a post-graduate degree.

About 58% of participants in the Control group report to be in STEM majors. They are enrolled in an average of 5.2 courses with an average of 1.5 class sessions per week beginning before 10 am (early classes). The average high school GPA in our sample of 4.14 is representative of the overall University of Pittsburgh student population: the interquartile range of students offered admission at Pitt in 2022 had a weighted average GPA ranging from 3.91 to 4.42.[19]

In column 3, we estimate the treatment-control difference for each baseline characteristic from a regression that includes an indicator for the Immediate Incentives group and wave fixed effects. We do not find any statistically significant differences between average baseline characteristics in the Control group and the Immediate Incentives group. We also estimate treatment-control differences for each treatment group separately in Giuntella et al. (2024) and find statistically significant differences at the expected rate (e.g., about five percent of tests are significant at the $p < 0.05$ level).

2.6 Analysis

For outcome measures that are observed repeatedly throughout the study (e.g., nightly sleep), our primary regression analysis estimates treatment effects during the intervention period and the post-intervention period relative to the Control group. Formally, we estimate the following OLS model, unless otherwise noted:

$$Y_{it} = \beta_1 D_i * T_t + \beta_2 D_i * P_t + X_i + \rho_t + w_t + \mu_t + d_t + \epsilon_{it} \quad (1)$$

where Y_{it} is the outcome measure of interest; D_i is an indicator equal to one if an individual was assigned to the treatment group of interest; T_t is an indicator equal to one for any observation during the four-week intervention period; P_t is an indicator equal to one for any observation in the post-intervention period; X_i includes an individual's baseline value of the outcome variable, baseline sleep (percent of weeknights with at least seven hours of sleep)[20], baseline GPA, indicators for the number of classes starting before 10 am in a week (ranging from 0 - 5), and demographic controls for gender, age in years (dummies), race/ethnicity (Asian, Black, Hispanic, White, other), and indicators for parents' highest education (less

default/files/assets/CDS_2021-2022_Pittsburgh%20Campus_2.pdf, accessed on November 18 2023.

[19]Data available at: https://admissions.pitt.edu/first-year-student/class-profile, accessed on November 18 2023.

[20]We exclude baseline sleep in regressions for sleep outcomes due to collinearity with the the baseline value of the outcome variable.

than college degree, which includes high school degree only or some college; college degree; or, more than a college degree). For all individual characteristics, we included a missing indicator for whether the variable is missing. The variables ρ_t, w_t, μ_t, d_t are fixed effects for the wave of the experiment, week of the experiment, month of the year, and day of the week, respectively. Standard errors are clustered at the individual level. For outcome measures that are observed only once during the study (e.g., course grades), we estimate the following OLS model, unless otherwise noted:

$$Y_i = \beta_1 D_i + X_i + \epsilon_i \tag{2}$$

where the variables are as described above. In regressions on course grades, the level of observation is the course weighted by the number of credits. Standard errors are clustered at the individual level.

Our main analysis compares the Control group to the Immediate Incentives treatment. As pre-registered, we also present the analysis for the primary outcomes comparing the Control group to the pooled incentives treatments (see column 7 of Table A.3 and column 8 of Table A.6) column.

We report both unadjusted p-values and statistical significance adjusted for multiple hypothesis testing (MHT) within families of secondary measures (see Tables C.1 and C.2 for the families of outcomes). For that purpose, we use the method described in Anderson (2008), calculating Anderson False Discovery Rate (FDR) q-values and noting which estimates are robust to adjustment.[21] In the Results section, we report unadjusted p-values and note which estimates are robust to adjustment.

In our main specifications, we include all participants who have outcome data. In the Appendix, we conduct sensitivity analyses that limit the sample to those who have both Fitbit and course grades data (Table A.3, column 3 and Table A.6, column 8).

3 Results

We first examine treatment effects on sleep hours. We then turn to the impact of our intervention on educational outcomes.

[21] Note that adjusted q-values can be both larger or smaller than unadjusted p-values. This is because, as noted by Anderson (2008), sharpened FDR q-values can be less than unadjusted p-values when many hypotheses are rejected.

3.1 Sleep

Baseline. Data from the baseline (pre-intervention) period reveals three findings about our college student participants: (1) students are sleep deprived relative to recommended guidelines; (2) they overestimate their sleep; and (3) they would like to sleep more than they do.

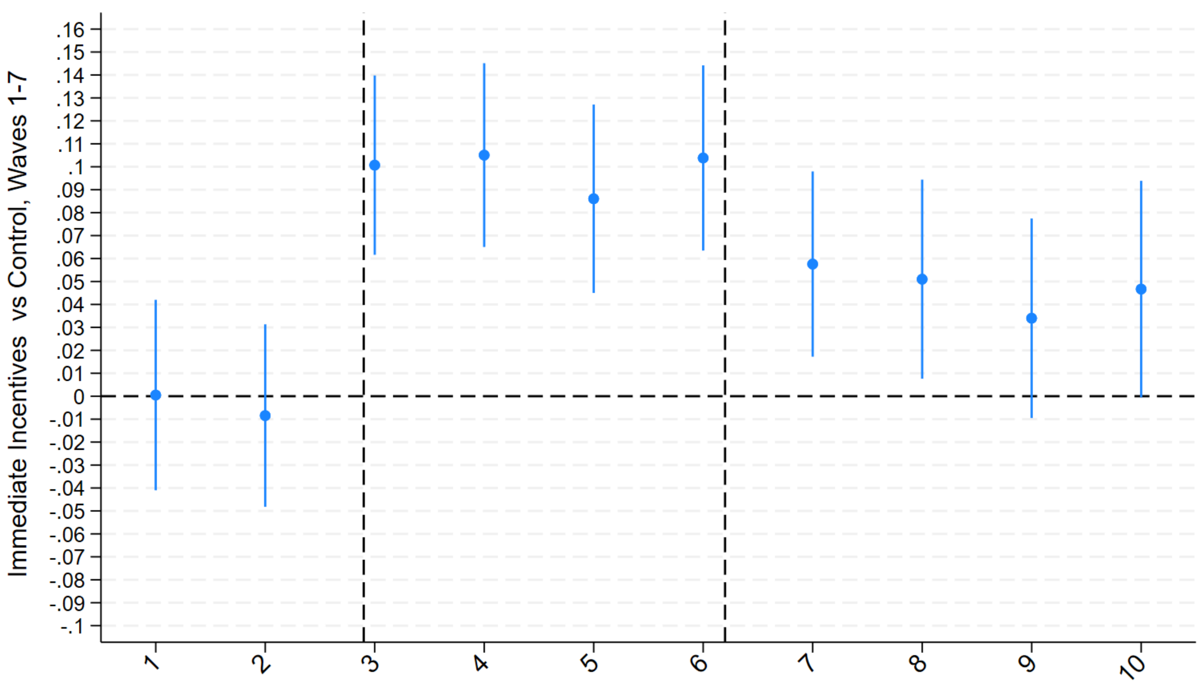
First, as shown in Table 2, on weeknights, participants sleep an average of 6.6 hours, meeting the recommendation of sleeping at least seven hours on approximately 43% of weeknights and sleeping less than six hours on approximately 29% of weeknights. About half of the participants have an average bedtime after 1 am, and about a quarter go to bed after 2 am on average. These data indicate that sleep deprivation is prevalent in our sample, consistent with a recent report by the National Institutes of Health indicating that more than 70% of college students sleep less than eight hours per night (Hershner and Chervin, 2014). In our sample, participants sleep less than eight hours on approximately 84% of weeknights. As shown in Appendix Figure A.4 panel A, only 4% of the participants achieve the goal of sleeping at least seven hours on all five weeknights while only 10% meet the goal on four nights. About a quarter of participants meet the goal on only one night, and one-fifth meet the goal on no nights.

Second, while only about one-third of participants meet the seven hour sleep goal on at least three out of five weeknights (as measured by the Fitbit), over two-thirds report sleeping at least seven hours on a typical weeknight in the enrollment survey. This suggests that students systematically overestimate their sleep duration.

Third, participants would like to sleep more than they actually do—both in comparison to their Fitbit-measured sleep and their (over-estimated) self-reported sleep. On average, their stated optimal sleep time is an hour longer than their reported typical sleep duration. Comparing optimal to actual sleep, Appendix Figure A.4 panel B, shows that the distribution of optimal sleep is skewed to the right relative to actual sleep (as measured by the Fitbit), with 97% of the students reporting that their optimal weeknight sleep would be at least seven hours and the modal stated optimal sleep being eight hours. Appendix Figure A.4 panel C compares actual and optimal sleep hours at the individual level. Ninety-four percent of individuals fall above the 45-degree line, indicating that they sleep less than they believe they should. These results suggest potential scope for interventions that help individuals increase their sleep, as they state they would like to.

Intervention and post-intervention period. We first estimate treatment effects on the primary measure of sleep, which we incentivized: sleeping at least seven hours on weeknights (Sunday - Thursday). This outcome variable only includes nighttime sleep, and excludes weekends, holidays and naps (defined as episodes of sleep that start between 7 am and 8 pm). Figure 2 plots the estimated difference in the rate of sleeping at least seven hours between the treatment (Immediate Incentives) and Control groups, by week. The estimates are from regressions by week in which individual-nights are the level of observation and we include an indicator for the treatment group with no additional covariates (the Control group is the omitted group). Standard errors are adjusted for clustering at the individual level (the bars in the figure indicate 95% confidence intervals).[22] As shown in the figure, there are no differences at baseline (weeks 1-2). Treatment effects emerge in the first week of the intervention (week 3) and persist throughout the four-week treatment period (weeks 3-6). Treatment effects decline as soon as the intervention ends (week 7) but remain positive and fairly steady throughout the post-intervention period (weeks 7-10).

Figure 2: Immediate incentives and sleep $\geq$ 7hrs (weeknights), excluding naps



Notes: The sample is restricted to weeknights (Sunday-Thursday nights). On the horizontal axis we report week of the study: baseline (weeks 1-2), treatment (weeks 3-6), post-treatment (weeks 7-10). The coefficient reports the difference in the likelihood of sleeping at least 7hrs between individuals in the Immediate Incentives treatment and those in Control by week. Standard errors are clustered at the individual level. Bars indicate 95% confidence intervals.

In Table 3, we present regression estimates of the treatment and post-treatment impacts

[22] We present the analogous figures for sleep hours in Figure A.5 and distributions of sleep hours in Figure A.6.

of Immediate Incentives on sleeping at least seven hours on weeknights (column 1) and on weeknight sleep hours (column 2), following the specification described in Equation 1. At baseline, participants meet the goal of sleeping at least seven hours on approximately 43.8% of weeknights.[23] During the intervention period, Immediate Incentives increase the rate of sleeping at least seven hours by an estimated 11.5 percentage points, a 26% increase ($p < 0.001$). The treatment effects persist into the post-intervention period but are about half the size: an estimated 5.2 percentage points, 12% higher than baseline ($p < 0.001$).[24] We estimate that total sleep hours increase by an estimated 19 minutes on average during the intervention period and by an estimated eight minutes during the post-intervention period. All estimates are significant ($p < 0.001$) and are robust to adjusting for multiple hypothesis testing (see Tables A.9). Our estimated effects on sleep duration are similar to those found for sleep medications, such as magnesium (Mah and Pitre, 2021), melatonin (Choi et al., 2022) and zolpidem (Xiang et al., 2021).[25]

The effects at the mean reflect shifts throughout the distribution of sleep, as measured by sleep hours and share of nights sleeping at least seven hours (Figure A.6). In Table A.4, we estimate treatment effects by baseline quartile of sleep (share of nights sleep at least seven hours in panel A and sleep hours in panel B). We find similar effects across quartiles during the intervention period; and evidence of larger post-intervention effects among participants with lower levels of sleep at baseline. These results suggest that our intervention has the

[23]The baseline average reported in Table 3 is slightly different from that reported in Table 2 as we are pooling Immediate Incentives and Control. Furthermore, in Table 2 we calculate the baseline average at the individual level and in Table 3 we calculate it at the night level, and not all participants have the same number of nights in the baseline period due to rolling enrollment.

[24]We conduct the following sensitivity analyses in Table A.3. In panel A columns 2 - 6, respectively we: limit the covariates to wave fixed effects, gender, baseline sleep and baseline GPA; limit the sample to participants who have term GPA, exclude missing nights rather than replacing missing data with individual baseline means; exclude wave 3 (onset of COVID-19); and, reweight the sample with respect to gender to make it representative of the gender composition of U.S. college students. The results do not change. We estimate treatment effects of 11.0-12.7 percentage points and post-treatment effects of 5.2-6.1 percentage points. The estimated effects pooling incentives (column 7) are similar, but slightly lower: 10.5 percentage points during treatment and 4.9 percentage points in post-treatment. Lee (2009) bounds range between 9.6 and 15.5 percentage points for the treatment and between 3.4 and 6.1 percentage points for the post-treatment (panel B columns 2 - 3).

[25]We do not find evidence that potential spillovers between the treatment and control groups are biasing our estimates. In waves 2 - 7, we asked subjects whether they knew someone else participating in the experiment. Out of 798 individuals, 271 reported knowing someone in the experiment (34% of our sample in waves 2-7); 84 (10.5%) reported knowing someone who either received reminders or rewards; and 69 (8.6%) reported knowing someone who received rewards. In Table A.3, we exclude each of the "spillover" groups from the analysis (panel B, columns 4 - 6) and also add controls for each of these categories (column 7). The point estimates decrease slightly when we exclude those who know anyone in the experiment and are almost identical when we exclude those who know someone in the treatment groups. Finally, we find no evidence that either treatment effects or baseline rates of sleeping at least seven hours increase across waves, further suggesting an absence of spillover effects over time (see Online Appendix (Giuntella et al., 2024)).

Table 3: Immediate Incentives and sleep

	(1)	(2)
	Sleep $\geq$ 7 hrs	Sleep hours
Treatment	0.115***	0.318***
	(0.013)	(0.035)
Post-Treatment	0.052***	0.139***
	(0.015)	(0.036)
Observations	39,035	39,035
Mean of Dep. Var.	0.438	6.653
Std. dev.	0.496	1.465
Number of individuals	840	840

Notes: The sample is restricted to individuals in the Immediate Incentives treatment and individuals in the Control group. All estimates include day of the week, week of the experiment, wave, and month fixed effects, baseline value of the outcome variable, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable is missing. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

most persistent impact on those who are most sleep deprived. In exploratory analysis, we examine heterogeneity with respect to time preferences measured at baseline and find little evidence of differential effects on sleep (see Online Appendix (Giuntella et al., 2024)).

As shown in Table A.5 (panel A), we do not find any evidence of substitution between incentivized weeknight sleep and unincentivized sleep during the day, on weekends, or during holidays (spring break for the treatment period and Thanksgiving for the post-treatment period). If anything, we find small positive spillovers, with some evidence of an increase in the likelihood of sleeping at least seven hours during weekends in the post-intervention period.[26]

In panel B of Table A.5, we examine additional sleep outcomes. Similar to our main results, we find that our intervention significantly increases the share of nights participants sleep at least six hours and the share of nights they sleep seven to nine hours, with persistent but smaller impacts in the post-treatment period.[27]

[26] Including naps, holidays and weekends, we estimate the intervention increased the share of nights with at least seven hours of sleep by 8.1 percentage points during treatment and 4.1 percentage points during post-treatment; and, increased total sleep hours by an estimated 14 minutes in the treatment period and 6 minutes in the post-treatment period ($p < 0.01$ for all estimates). As noted above, Fitbits may not fully capture short naps.

[27] Sleeping less than six hours is a common metric of sleep deprivation (Hafner et al., 2017). The recom-

3.2 Educational outcomes

Next, we investigate the impact of our primary treatment on educational outcomes. Table 4 presents regression estimates of the impact of the intervention on semester GPA and secondary educational outcomes.[28] The regressions follow the specification of Equation 2. In columns 1 - 2 we estimate treatment effects on our primary outcome, course grade, in the term the intervention took place. Column 1 includes all course types (lectures, seminars, labs, independent studies and other classes) whereas column 2 restricts the analysis to lectures (which account for approximately 80% of course types). Columns 3 - 4 report the same analysis for the term following the intervention. In columns 5 and 6, we examine the persistence of the effects two terms after the intervention.

Table 4: GPA, Immediate Incentives

	(1)	(2)	(3)	(4)	(5)	(6)
	Term of intervention		Term +1		Term +2	
	All classes	Lectures	All classes	Lectures	All classes	Lectures
Immediate Incentives	0.075**	0.088**	0.068*	0.091**	0.004	0.004
	(0.037)	(0.042)	(0.038)	(0.042)	(0.042)	(0.046)
Observations	4,300	3,413	4,087	3,298	3,842	3,080
Mean of dep. var.	3.502	3.436	3.553	3.494	3.547	3.505
Std. dev.	0.763	0.805	0.756	0.795	0.774	0.806
Number of individuals	833	827	784	782	727	718

Notes: All estimates include demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), baseline sleep, indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Observations are weighted by the number of credits taken in the semester. Standard errors are clustered at the individual level. Mean of dep. var is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

As shown in columns 1 and 2, we estimate that Immediate Incentives improved average course performance by 0.075 grade points in all classes ($p = 0.045$) and 0.088 grade points in lecture classes ($p = 0.035$).[29] The estimated GPA impacts of 0.075 - 0.088 grade points

mendation of sleeping seven to nine hours draws on studies that link excessive sleep duration to detrimental effects on health (Hirshkowitz et al., 2015; Jike et al., 2018).

[28] As discussed in Section 2.4, we are missing GPA for 1.9 percent of our participants. We examine differential attrition on the GPA measure in Table A.2 and find no evidence for differential attrition on term GPA.

[29] As shown in panel C of Table A.6, we find no treatment effects in non-lecture classes (i.e., seminars, labs, internships, directed studies). As discussed in Section 2.4, over two-thirds of students in these classes receive an A, and the lowest quartile is A-, leaving little room to improve grades (see Figure A.6).

during the intervention semester are equivalent to a 0.10 - 0.11 standard deviation (SD) increase in grades. These effect sizes are about 25-29% of the GPA gap (0.30 grade points) between students with above versus below median high school GPA. We estimate a treatment effect of similar magnitude on course performance in the semester following the intervention, providing suggestive evidence of persistent effects (column 3 ($p = 0.074$) and column 4 ($p = 0.029$)).[30] However, we do not find treatment effects in the semester two terms after the intervention (columns 5 and 6).

We conduct the following sensitivity analyses in Table A.6: limit the covariates to wave fixed effects, gender, baseline sleep and baseline GPA; limit the sample to participants who have post-intervention term grades; limit the sample to participants who have baseline GPA (high school or baseline term GPA); limit the sample to participants who have sleep data at baseline; exclude wave 3 (onset of COVID-19); and reweight the sample with respect to gender to make it representative of the gender composition of U.S. college students. The estimated effects are slightly smaller when we limit the covariates, exclude participants missing grades data or reweight the sample: 0.060 - 0.075 grades points for all classes (p-values ranging between 0.024 and 0.090) and 0.074 - 0.088 for lectures (p-values ranging between 0.019 and 0.076). Interestingly, when we exclude the Spring 2020 semester (wave 3), when students were sent home mid-semester due to the onset of the COVID-19 pandemic, our estimated treatment effects increase slightly to an estimated 0.090 - 0.105 grade points. During this wave of the study, our participants experienced the abrupt closure of the university in the middle of the semester and disruptions in sleep and other lifestyle habits (Giuntella et al., 2021). When pooling all incentives treatments, the estimates are similar with slightly smaller average impacts, as shown in Table A.3 column 7 for sleep and Table A.6 column 8 for grades.

We also explore the effects of our intervention on other measures of course performance (Table A.8). This analysis includes the courses in our main analysis as well as courses that do not contribute to GPA because they do not have a grade on a four-point scale, such as pass, honors, and incomplete (see Table A.1 for the grading system). We find that students in the treatment group are less likely to receive any grade, which is primarily due to small increases in the likelihood of withdrawing from a class (column 2). At the same time, students in the treatment group are marginally less likely to fail a course (column 3). These results suggest that treated students are more likely to withdraw from classes they would otherwise fail. As a result, there are no significant differences in the likelihood of passing a class (column 4) nor in the number of credits completed in a term. We note that the effects

[30] The effect one term after the intervention is not robust to multiple hypothesis testing (q-value=0.18, Table A.9).

on the likelihood of having any grade, withdrawing, or failing are not statistically significant after adjusting for multiple hypothesis testing (Table A.9).

Table 5: Immediate Incentives and GPA: Heterogeneity by schedule and class type

	(1) Course grade	(2) Class start: before 10am	(3) Class start: 10am-2pm	(4) Class start: after 2pm	(5) Class type: non-STEM	(6) Class type: STEM
Panel A: All classes						
Immediate Incentives	0.075**	0.056	0.094**	0.064	0.013	0.132**
	(0.037)	(0.064)	(0.044)	(0.055)	(0.034)	(0.057)
Observations	4,300	959	1,634	1,568	2,351	1,948
Mean of Dep. Var.	3.502	3.471	3.493	3.497	3.696	3.267
Std. dev.	0.763	0.810	0.744	0.773	0.574	0.888
Number of individuals	833	607	773	751	794	694
Panel B: Lectures						
Immediate Incentives	0.088**	0.022	0.115**	0.095	0.030	0.132**
	(0.042)	(0.072)	(0.048)	(0.065)	(0.040)	(0.059)
Observations	3,413	735	1,385	1,229	1,717	1,695
Mean of Dep. Var.	3.436	3.403	3.447	3.426	3.668	3.202
Std. dev.	0.805	0.859	0.774	0.815	0.598	0.912
Number of individuals	827	523	731	697	710	692

Notes: All estimates include demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), baseline sleep, indicators for the number of classes starting before 10 am, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Observations are weighted by the number of credits taken in the semester. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Turning to heterogeneity, Table 5 shows that the results are not driven by performance in early-morning classes but rather the directionally largest effects are in late morning and early afternoon classes that occur between 10 am and 2 pm, followed by afternoon/evening classes (after 2 pm). This is in line with the findings from Carrell et al. (2011) that early class start time affects performance in all classes, not just classes taking place early in the morning.

We additionally find evidence that our overall effects are driven by performance in STEM courses: on average our intervention leads to a 0.132 grade point increase in STEM courses, which corresponds to a 0.15 SD increase in STEM grades. By contrast, we estimate small increases of 0.013 grade points in non-STEM courses. This finding suggests that sleep may have a more significant impact on quantitative courses. However, given the large proportion of STEM majors in our sample (58%), another possibility is that improved sleep could enhance performance in courses important for a given major. The results may also partially reflect that there is more room to move grades in STEM courses, which have an average GPA of 3.27, compared to non-STEM courses with an average GPA of 3.69.

In exploratory analysis (Table A.7), we examine heterogeneous treatment effects on sleep and GPA by individual characteristics. We find larger effects among women, both on sleep and GPA. Effects on GPA are large among STEM majors, but effects on sleep are similar for STEM and non-STEM majors. Further, first-term freshmen students exhibit substantially larger effects than other students. These findings are consistent with the idea that habits may be more malleable among freshman who have not fully developed their routines (Creswell et al., 2023).

4 Mechanisms for effects on educational outcomes

Our intervention targeted increased sleep hours as a primary mechanism for improving academic performance. In this section, we examine additional mechanisms through which our treatment effects could operate. First, we examine treatment impacts on sleep measures other than total hours, such as timing, regularity and quality, which could also affect educational outcomes. Second, we examine how increased time spent sleeping affects time spent on other activities in order to understand the opportunity costs of sleep and the extent to which shifts in non-sleep activities could drive academic performance impacts. Third, we examine treatment effects on creativity and quantitative tasks, physical health and mental well-being in order to understand the effect of sleep on these important outcomes, as well as their potential role as mechanisms for the impact of the intervention on academic performance. Finally, we explore the extent to which the immediate rewards themselves could be driving our results through direct effects of cash.

4.1 Sleep timing, regularity and quality

In the first two columns of Table 6 panel A, we estimate treatment effects on bedtime and wake-up time. During the intervention period, treated participants go to bed an estimated 19 minutes earlier than participants in the control group ($p < 0.001$) with directionally earlier wake up times ($p = 0.15$). This pattern does not persist in the post-treatment period when the incentives ended but the bedtime reminders continued. Instead, average bedtime largely reverts to baseline levels and treated participants wake up slightly later ($p = 0.074$). As shown in Figure A.5, both bedtime and wake-up time get progressively later over the course of the intervention period and stabilize during the post-intervention period.

We examine sleep regularity in two ways. First, in column 3, we examine effects on

Table 6: Sleep Timing and Regularity

	(1)	(2)	(3)
Panel A: Daily data	Bedtime	Wake up time	Sleep Regularity Index (SRI)
Treatment	-0.313***	-0.048	1.417***
	(0.036)	(0.034)	(0.352)
Post-Treatment	-0.027	0.068*	0.600
	(0.037)	(0.038)	(0.388)
Observations	39,035	39,035	39,035
Mean of Dep. Var.	25.23	7.954	87.76
Std. dev.	1.560	1.557	9.705
Number of individuals	840	840	840
Panel B: Weekly data - Regularity Within-individual weekly s.d.	Bedtime	Wake up time	Sleep hours
Treatment	-0.048**	-0.072**	-0.105***
	(0.023)	(0.031)	(0.033)
Post-Treatment	-0.057**	-0.094***	-0.053
	(0.025)	(0.036)	(0.035)
Observations	7,807	7,807	7,807
Mean of Dep. Var.	0.892	0.936	1.160
Std. dev.	0.481	0.592	0.664
Number of individuals	840	840	840

Notes: The sample is restricted to individuals in the Immediate Incentive treatment and individuals in the Control group. All estimates include day of the week, week of the experiment, wave, and month fixed effects, baseline value of the outcome variable, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for the number of classes starting at 10am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable is missing. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

the Sleep Regularity Index (SRI), which captures day-by-day similarity in sleep and wake patterns (see Section 2.4 for details). In panel B, we examine the regularity of bedtime, wake-up time and sleep hours, which we measure by the within-individual standard deviation at the week level (e.g., how much a person's nightly bedtime varies across the week where the level of observation is the individual-week).[31] We find that the intervention significantly increases SRI by about 0.15 SD and within-individual variability in bedtime, wake-up time and sleep hours by about about 0.10 to 0.16 SD during the treatment period. The magnitude of the effects during the intervention and post-intervention periods are of similar size for bedtime

[31] The regressions follow the specification of Equation 1, except we exclude day of the week fixed effects given the analysis is at the weekly level.

and wake-up time (with smaller post-treatment effects on SRI and total sleep hours). These findings suggest that, while treated participants do not on average sustain earlier bedtimes after the intervention ends, they do establish more stable bedtime and wake-up time routines.

On our measures of sleep quality, we find small positive increases in minutes of REM sleep, no significant impact on minutes of deep sleep, and small, marginally significant impacts on sleep efficiency (Appendix Table A.5, panel B). We note that, in our sample, baseline efficiency is high: participants are asleep an estimated 93.5% of the time they are in bed. By comparison, Bessone et al. (2021) estimate an efficiency of 70% among their experimental participants in India. As discussed in Section 2.4 and in Giuntella et al. (2024), the accuracy and reliability of Fitbits in capturing sleep stages and assessing sleep quality remain limited.

4.2 Time use

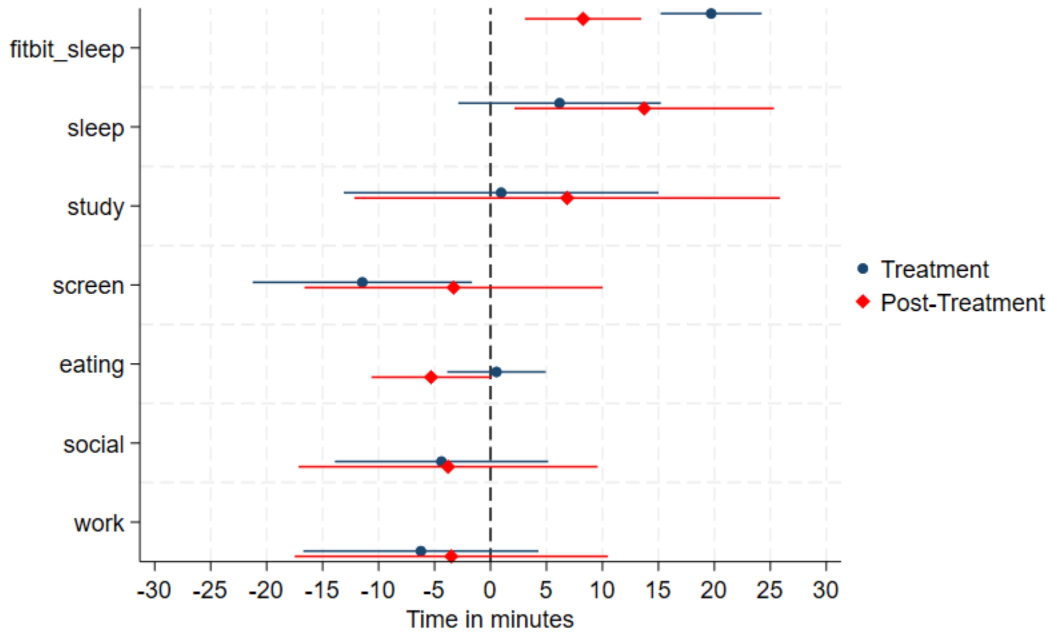
We next focus on our survey measures of time use that we asked weekly throughout the study (see Section 2.4 for details). Figure 3 shows estimated treatment and post-treatment effects on time spent (in minutes) for the top six time use categories, using the regression specification in Equation 1. In the first row of the figure, we show the estimated effects from the Fitbit data that we report in column 2 of Table 3. As discussed earlier, our intervention increased sleep by 19 minutes during treatment and eight minutes after the removal of the incentives. Immediate Incentives directionally increase self-reported sleep in both the intervention and post-intervention periods by about 6 - 14 minutes on average per day (Table A.10). We also find that subjects were 7 (6) percentage points more likely to report at least seven hours of sleep during the intervention (post-intervention) period.

During the intervention period, incentives to sleep significantly decrease average screen time, which includes internet browsing, TV/videos and games, excluding screen time for studying, by an estimated 11.5 minutes per day ($p < 0.01$). We estimate smaller and not statistically significant treatment effects on screen time during the post-intervention period. We do not find evidence of meaningful changes in time spent studying, socializing, eating or working during the intervention.[32]

In Figure 4, we report treatment effects on sleep, screen time, social time and study time over the course of the day during the intervention period. The effects on sleep and screen time are concentrated at night (8 pm - 4 am). Interestingly, while total study time

[32]The effect on screen time is marginal once adjusting for multiple hypothesis testing ($q = 0.12$). We estimate treatment effects separately for internet, TV/videos and games in Table A.10 and the overall impact is largely driven by decreases in TV/video time. The table also reports effects on other time use categories. At baseline, we estimate the following average minutes per day for each category: sleep (494 minutes), study time (321 minutes), screen time (172 minutes), eating and preparing food (95 minutes), social time (101 minutes), and work time (92 minutes).

Figure 3: Incentives to sleep and time use (minutes)



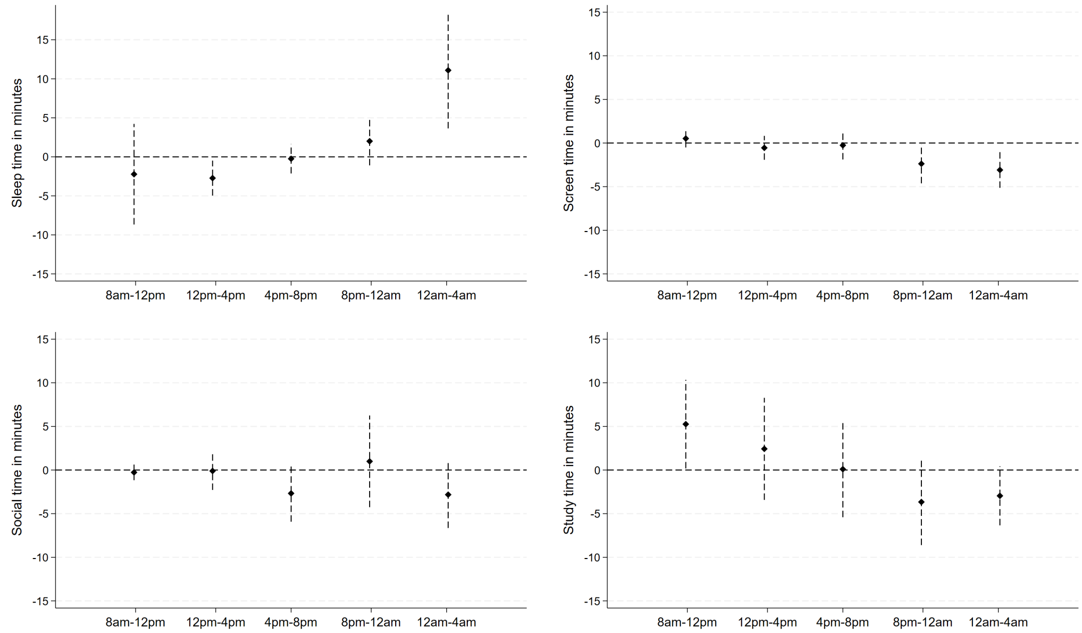
Notes: The figure reports differences between the Immediate Incentives treatment and Control group in time-use during the intervention (in navy) and in the post-intervention period (in red). All the coefficients are obtained from regressions including wave, month, and day of the week fixed effects, baseline value of the outcome variable, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Standard errors are clustered at the individual level. Bars indicate 95% confidence intervals.

does not increase, we observe a reallocation of study time from the evening/night (8 pm - 4 am) to the morning (8 am - 12 pm), although not precisely estimated. These results suggest that incentives to sleep led participants to develop sleep habits characterized by earlier screen disengagement at night and more focus on study time during the day. We find no significant changes in time spent working or participating in social activities. We also estimate treatment effects during the post-intervention period and find similar, but weaker, patterns (Figure A.7).

4.3 Task performance, physical health and well-being

To assess performance, we collected measures of quantitative reasoning and creative output on alternating weeks throughout the study (see Section 2.4 for details). We do not find any impact of the intervention on these measures (see Table A.11, columns 1 and 2), which could be due to the intervention not affecting math or creativity or to our coarse measures

Figure 4: Immediate Incentives to sleep and time use over the day: Intervention period



Notes: The figure reports differences between participants in the Immediate Incentives treatment and Control group in the minutes allocated to different time-use activities during the intervention throughout the day. All the coefficients are obtained from regressions including wave and day of the week fixed effects, baseline value of the outcome variable, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Standard errors are clustered at the individual level. Bars indicate 95% confidence intervals.

not being able to capture the impact of our intervention on performance. In addition, sleep could have influenced other measures of performance and cognition through channels like attention or memory consolidation, which were not captured by our measures (Diekelmann and Born, 2010).

Our final outcomes of interest are well-being and physical health. Previous work suggests that there is a positive relationship between sleep and both mental well-being and physical health (Giuntella and Mazzonna, 2019; Giuntella et al., 2017; Jin and Ziebarth, 2020). As discussed in Section 2, we use the Fitbit to measure participants' heart rate, daily steps and physical activity. We present estimates of treatment and post-treatment effects in Table A.11. We find no evidence of treatment effects on any of the physical health measures (columns 3, 4, and 5).

To investigate the impact of the intervention on well-being, we sent participants weekly text messages to collect data on mood, stress, and resilience to stress. Additionally, we utilize the Generalized Anxiety Disorder (GAD-7) scale to assess anxiety levels and the Center for

Table 7: Immediate Incentives and well-being

	(1) Happiness	(2) Stress	(3) Resilience	(4) CES-D	(5) GAD-7
Treatment	-0.093	0.063	0.153***		
	(0.108)	(0.059)	(0.055)		
Post-Treatment	-0.089	-0.034	0.072	0.394	0.078
	(0.111)	(0.068)	(0.063)	(0.886)	(0.404)
Observations	4,221	3,678	3,605	1,466	1,466
Mean of Dep. Var.	6.388	3.114	2.988	15.79	6.834
Std. dev.	1.649	1.116	0.995	10.23	4.871
Number of individuals	807	811	805	838	838

Notes: Estimates in columns 1 - 3 include day of the week, wave, and month fixed effects, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. For mood, participants indicated, on a 10-point Likert scale, how happy they felt in that moment. For stress and resilience, participants indicated, using a 5-point Likert scale, 1) the extent to which participants faced stress in their life at the time of answering the survey and 2) the extent to which they felt able to deal with the stress they were facing. For columns 4 and 5, outcomes are measured at endline, and estimates include all of the controls listed above, except for day of week, week of the experiment, and month fixed effects. CES-D is the Center for Epidemiologic Studies Depression Scale. GAD-7 is the General Anxiety Disorder-7. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

Epidemiologic Studies Depression (CES-D) scale to gauge depression levels. These scales were administered at baseline and endline only, so we are only able to estimate treatment effects on post-intervention end-of-semester anxiety and depression. Table 7 shows that the intervention does not have a significant impact on mood or stress levels (columns 1 and 2). However, it led to a statistically significant increase in resilience—participants' self-reported ability to cope with stress—by approximately 0.15 standard deviations (column 3), which is significant at the 10% level after adjusting for multiple hypothesis testing. The intervention did not show any significant effects on post-treatment measures of depression and anxiety (columns 4 and 5), with point estimates small in magnitude and not statistically significant.

4.4 Effects of immediate cash rewards

Finally, we examine whether receiving immediate cash rewards could have had direct effects on academic performance. Students in the Immediate Incentives group received an average of $52 in immediate rewards over the four-week intervention period ($4.75 on 55% of the 20 treatment days).[33] If rewards of this size alleviate financial constraints in the middle of the semester that hinder academic achievement, they are most likely to do so among students

[33] As discussed above, students in all conditions received $6 at the start of the study a few weeks into the semester. Students in Control condition received an additional $24 at the end of the study just before final exams.

who are expected to face cash shortages. To examine whether this occurs, in Appendix Table A.12, we examine treatment effects on GPA by demographic subgroups that are more likely to experience an impact of direct cash: students who report receiving financial aid vs. not (columns 1-2), first generation college students vs. not (columns 3-4), and students who report working at baseline vs. not (columns 5-6). We find no evidence that disadvantaged students have larger treatment effects – point estimates are always smaller for the lower socioeconomic status subgroup compared to the more advantaged subgroup.

Second, we asked participants about their consumption of caffeinated beverages, which could improve alertness and ability to focus (we asked about caffeine consumption in every other weekly time use survey). As shown in Appendix Table A.11 columns 6-7, we find no evidence that participants in the immediate incentives group increased their consumption of caffeinated beverages compared to those in control.

Third, in waves 2-7, we asked participants who received incentives what they spent the rewards on. In Appendix Table A.12, we estimate treatment effects excluding treatment group participants who report spending the incentives on food, drinks or school related expenses (column 7); and, treatment effects including only those treatment group participants who report spending on those categories (column 8). Note that we include the full control group in all analyses. We find no evidence that treatment effects are driven by participants who use the rewards to purchase food, beverages or school supplies. The estimated effects are directionally larger when these participants are excluded and directionally smaller when the analysis includes only these participants in the treatment group.

Finally, as discussed in more detail in Section 6, there is little evidence from prior work that the size of the rewards in our study would have a meaningful impact on academic performance. As shown in Figure 5, programs that provide grants of similar size find no impact on GPA (e.g., Evans et al., 2020[A]).

Taken together, we do not find evidence for the direct effects of the cash rewards as a mechanism for our treatment effects. Instead, our results point to earlier bedtime and increases in sleep regularity, decreases in screen time, reallocation of study time from evening to morning hours and improved capability to deal with stress as potential mechanisms. Beyond total sleep hours, sleep timing and regularity may be important for cognition and performance. Prior work suggests that irregular sleep among college students is associated with delayed circadian rhythms and lower academic performance (Phillips et al., 2017; Windred et al., 2024; Smarr, 2015). Similarly, shifting study time to the morning, when students may be more alert and better able to concentrate, could enhance learning outcomes (Pope, 2016). Finally, reductions in screen time, particularly at night, could contribute to the impact of our intervention on academic performance. Prior experimental and quasi-experimental work does not find significant effects of social media on grades (Braghieri et al., 2022; Collis and Eggers, 2022). However, Braghieri et al. (2022) do find that social media worsens college students' mental health and increases the likelihood they report adverse impacts of mental health on their academic performance. Our findings that Immediate Incentives to sleep both reduce screen time and enhance resilience provides suggestive evidence that sleep, social media use, and mental health may interact in important ways that affect educational outcomes, highlighting the need for future research to directly examine these mechanisms.

5 Mechanisms for habit formation in sleep

We designed the Immediate Incentives treatment with the aim of creating persistent bedtime habits by pairing bedtime cues with immediate rewards for meeting sleep goals. As discussed above, the intervention leads to earlier bedtimes, which was the intention of the bedtime reminder. However, these effects are not sustained after the incentives end, even though bedtime reminders remain in place. Instead, we find evidence of persistent treatment effects on sleep duration and sleep regularity for about four weeks after the intervention period ends.

To better understand the mechanisms for the effects of Immediate Incentives on sleep behaviors, we now turn to an examination of our secondary treatments. As discussed in Section 2 and summarized in Table 1, we designed the secondary treatments to vary elements of our primary Immediate Incentives treatment in order to investigate which of the features of Immediate Incentives drive habit formation and persistence. These include: (1) Delayed Incentives, which is identical to Immediate Incentives except that the rewards are distributed at the end of the study about a month after treatment; (2) Delayed Incentives No Cue/Feedback, which is identical to Delayed Incentives except that participants do not receive cues or feedback; and, (3) Cue/Feedback which only provides cues (bedtime reminders) and feedback with no rewards. Table 8 estimates the effects of our primary and secondary treatments on sleep hours and sleeping at least seven hours on weeknights. We restrict the analysis to waves 1 - 3 of the experiment when the secondary treatments were conducted.

We first examine the importance of immediate rewards compared to delayed or no rewards. As shown in columns 1 - 2, the effects of Immediate Incentives during treatment are about 48% to 73% percent higher than the effect of Delayed Incentives (with or without cues and feedback); and about three to four times higher than the effects of Cue/Feedback alone. The differences between the estimated impact of Immediate Incentives and each of the secondary treatments are all significant at the 10% level.[34] During the post-intervention period,

[34]All differences remain significant at the 10% level after adjusting for multiple hypothesis testing, except

Table 8: Secondary treatments

	(1) Sleep ≥ 7	(2) Sleep hours
Treatment		
Immediate incentives	0.126***	0.304***
	(0.020)	(0.052)
Delayed incentives	0.085***	0.177***
	(0.021)	(0.056)
Delayed incentives, no feedback	0.080***	0.176***
	(0.021)	(0.057)
Cue/Feedback only	0.029	0.109**
	(0.019)	(0.051)
Post-Treatment		
Post: Immediate incentives	0.032	0.182***
	(0.024)	(0.065)
Post: Delayed incentives	0.004	0.043
	(0.025)	(0.070)
Post: Delayed incentives, no feedback	0.048*	0.090
	(0.027)	(0.073)
Post: Cue/Feedback only	0.008	0.068
	(0.026)	(0.070)
Observations	27,130	27,130
Mean of Dep. Var.	0.429	6.677
Std. dev.	0.495	1.529
Number of individuals	589	589
Difference between treatments (*p*-value)		
Immediate vs delayed incentives	0.093	0.052
Immediate vs delayed incentives no feedback	0.053	0.047
Immediate vs only feedback	0.001	0.001
Post: Immediate vs delayed incentives	0.273	0.059
Post: Immediate vs delayed incentives no feedback	0.568	0.211
Post: Immediate vs feedback only	0.354	0.120

Notes: The sample is restricted to waves 1 - 3. All estimates include day of the week, week of the experiment, wave, and month fixed effects, baseline value of the outcome variable, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable is missing. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

the estimated effects of Immediate Incentives are generally larger in magnitude than the secondary treatments, particularly for sleep hours. However, the effects are not statistically indistinguishable.

We next examine the importance of bedtime cues and feedback. Comparing the two Delayed Incentives treatments, we find that the effects of Delayed Incentives are similar with or without cues and feedback. If anything, the variant with no cue/feedback has directionally larger effects in the post-intervention period. In exploratory analysis reported in Appendix Table A.3, panel B column 1, we estimate the post-treatment effects of the Immediate Incentives intervention separately for the subgroup of participants who continued to receive reminders and feedback in the post-treatment period (Immediate Incentives Post Cue/Feedback) and for the subgroup of participants who stopped receiving them at the end of the intervention period (Immediate Incentives No Post Cue/Feedback). [35] Our estimates reveal no significant differences between these two subgroups, suggesting that receiving bedtime cues after incentives stopped did not help sustain the sleep patterns developed during the intervention period.

Collectively, the results in this section do not provide strong evidence that individuals built automatic habits in response to our external cue. Early bedtimes do not persist in the post-intervention period when cues (but not rewards) remained in place. And, treatment variants with cues do not outperform variants without them, in either the intervention or post-intervention periods. However, the intervention did lead to persistent improvements in sleep duration and sleep regularity, as evidenced by reduced variability in bedtime and wake-up time. This suggests that participants may have identified personal bedtimes and wake-up times that they could sustain independently, even without continued external reinforcement.

Although habit formation is often conceptualized as an automatic response to an external cue, our findings suggest that sustained behavior change may have instead resulted from participants developing self-sustaining sleep routines, possibly reinforced by self-identified cues. Alternatively, sustained behavior change may have occurred through other mechanisms, such as acquiring a greater preference for sleep, recognizing its benefits more clearly, or restructuring personal schedules in ways that reduced the costs required to maintain regular sleep (Volpp and Loewenstein, 2020). Future research could explore these alternative pathways to sustained behavior change and examine when external cues can effectively promote habit formation.

for the difference between the treatment effect on sleeping at least seven hours of Immediate Incentives vs. Delayed Incentives (Table A.13).

[35] We restrict the analysis to waves 5 and 7 in which we ran both variants. The interaction term "Immediate Incentives*No Cue in Post-Treatment" estimates the difference in sleep in the post-intervention period between Immediate Incentives No Post Cue/Feedback and Immediate Incentives with Post Cue/Feedback.

6 Benchmarking our results

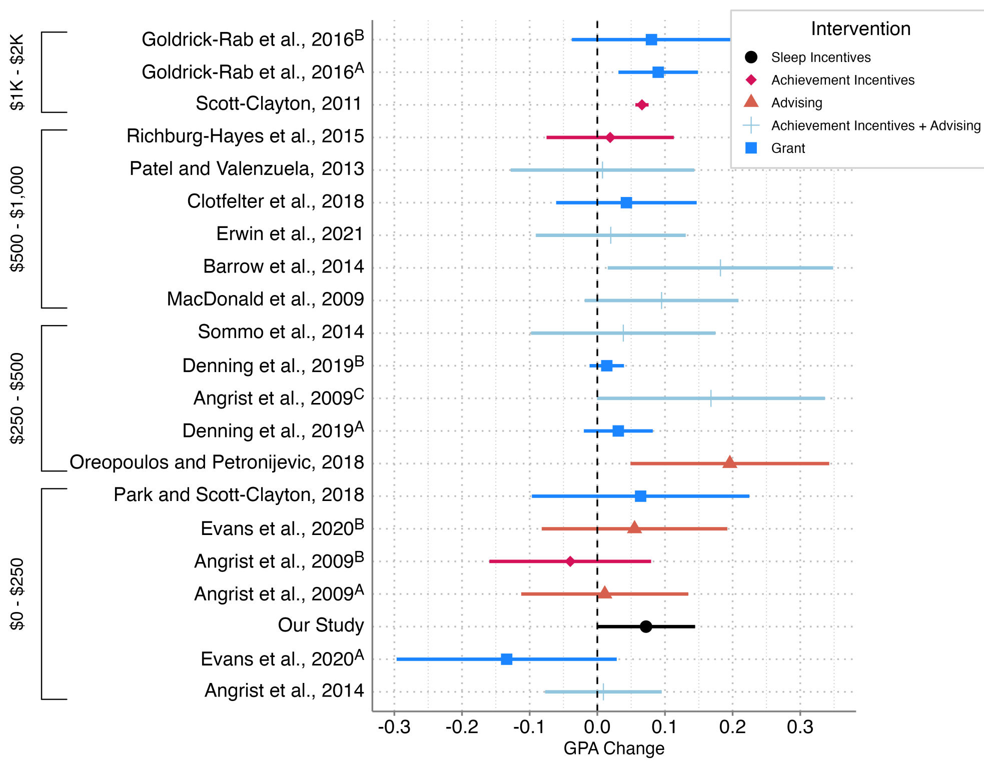
We benchmark our results in two ways. First, we compare our effects to casual estimates of the relationship between sleep and academic performance from naturally occurring data. As discussed above, prior work examines the effect of shifts in sunset and school start times on sleep, grades and test scores. For example, Carrell et al. (2011) estimates that shifting the start time of college students' first class by an hour from 7:00 am to 8:05 am improves overall academic performance by 0.12 - 0.14 SD. The study does not directly measure students' sleep. Other studies using self-reported sleep estimate that a one-hour later school start time increases sleep by 36 minutes among American children, with a 0.16 SD improvement in reading and no change in math (Groen and Pabilonia, 2019). Related studies find that the the sun rising one hour later increases average sleep among American children by an estimated six minutes, with a 0.082 SD increase in math scores and a 0.057 SD improvement in reading scores (Heissel and Norris, 2018). Taken together, these studies suggest that a one hour shift increases sleep by 6 - 35 minutes and has either a null effect on academic performance or improves grades and test scores by 0.06 - 0.16 SD.[36] Our impacts of a 19 minute average increase in weeknight sleep during treatment, an eight minute average increase in post-treatment, and a 0.10 - 0.11 SD improvement in grades, falls within the range of the estimates in prior work on the causal relationship between shifts in sleep and changes in academic performance.

Second, we compare the cost effectiveness of our intervention to prior work examining policies aimed at improving college students' outcomes, including those that condition rewards on academic performance. Angrist et al. (2014) summarize the work on performance-based incentives, including their own, and conclude that, "the picture that emerges. . . is one of mostly modest effects. . . [And there are] similarly discouraging results from studies of state-based merit aid programs. A few studies report positive effects, most notably Scott-Clayton (2011)'s evaluation of West Virginia PROMISE," which conditions free tuition on meeting a minimum GPA. Scott-Clayton (2011) finds similar-sized GPA effects to our study, 0.066 grade point improvements, but at over ten times the cost, an estimated \$1,250 per student per semester. By comparison, we estimate that incentives to sleep increase semester GPA by 0.075 - 0.088 grade points and cost approximately \$112 per participant for the semester and would cost about \$160 per participant per year. This includes \$60 for the cost

[36] Outside the U.S., Lusher et al. (2019) estimates that shifting class start times by an hour increases average sleep by about four minutes among Vietnamese University students with no effect on performance (Lusher et al., 2019). Jagnani (2021) estimates that the sun setting one hour earlier increases sleep by an average of 30 minutes among Indian children and that the sun setting 10 minutes earlier improves test scores by 0.10 SD and leads to 0.14 more years of schooling.

of the Fitbit and an estimated average of $52 per participant per semester for the incentives (participants in the Immediate Incentives group received the incentives of $4.75 per night on 55% of the 20 nights we offered it).

Figure 5: Cost-Effectiveness



Notes: The figure compares our Immediate Incentives effect on GPA to estimates from previous interventions aimed at improving college academic performance. Studies are grouped on the vertical axis based on their cost per subject per semester. Bars represent 95% confidence intervals. Superscripts above paper names denote different treatment arms or treatment groups. For Goldrick-Rab et al. (2016), superscript A is an estimate for the first cohort studied and B is their pooled estimate for the second and third cohort. For Denning et al. (2019), A and B are estimates for first-year and returning students, respectively. For Angrist et al. (2009), A is an estimate for an advising and peer-support treatment arm, B is for a financial incentives arm, and C for an arm combing A and B. For Evans et al. (2020), A estimates a grant treatment arm, and B estimates combined grant aid with academic advising. For Oreopoulos and Petronijevic (2018), course grades on a 0-100 scale have been divided by 25 for comparability to 4.0 scale GPA effects. Effect sizes for Richburg-Hayes et al. (2015), Patel and Valenzuela (2013), Erwin et al. (2021), MacDonald et al. (2009), and Sommo et al. (2014) are from Lintner (2024).

In Figure 5, we report the estimated effects on GPA from our study, alongside prior work examining the impact of achievement incentives, advising, and grants, ordered from most to least costly.[37] As depicted in the figure, our intervention is characterized by relatively

[37] Achievement incentives include performance-based incentives and merit aid. Advising includes advising and support services. We report authors' OLS estimates of effects on non-cumulative GPA, either at the semester or year-level. Costs are per program participant per semester. We note that for some of these programs the primary outcome may have been enrollment, persistence or graduation and GPA may have been a secondary outcome. We do not adjust program costs for inflation (note this overestimates the cost of our intervention relative to others). See Table A.14 for more information on each study.

low costs, while the estimated effects are equal to or greater than those observed in most previous studies. Only a handful of interventions surpass ours in terms of impact, but they come with a two to five-fold higher cost per participant. In particular, our intervention compares favorably to achievement incentives that condition rewards on GPA. As noted above, in a recent meta-analysis, Lintner (2024) estimates that performance-based financial incentives improve college students' GPA by an average of 0.041 grade points with average costs of $457 per student per semester (or $462 if weighted by study sample size). Our intervention finds effect sizes of almost twice the magnitude (0.075 - 0.088 grade points) at less than one-quarter the cost (about $112).

7 Conclusion

In this paper, we show that an intervention targeting sleep habits improves academic performance. Our results suggest that targeting sleep may be a more cost-effective way to improve student performance than incentivizing performance directly. This could be because incentives based on sleep can be immediate whereas incentives based on grades are typically offered with a delay (e.g., at the end of the term). It could also be the case that students do not fully understand the production function for grades – and in particular, may not recognize the role of sleep.

Future research could explore how beliefs and information about sleep shape individual behavior and educational outcomes. Our students self-report wanting to sleep more than they do, but we do not know students' beliefs about the causal relationship between sleep and academic performance. Our study also highlights that the mechanisms for the impact of sleep on performance could have important interactions with behaviors that have largely been examined separately, including social media use and mental health. A better understanding of these interactions can help inform which behaviors to target to maximize downstream positive impacts on both performance and well-being (Lindquist and Sadoff, 2023).

Policies building on our results could focus on both the individual and structural level. Interventions that leverage a better understanding of beliefs could help bridge the gap between students' sleep intentions and their actual behavior. The timing of such interventions may also matter – for example, targeting first-term freshman and the middle of the semester may be particularly effective in shifting student behavior (Creswell et al., 2023; Liu et al., 2022). Institutions could also test structural-level interventions, such as adjusting course start times or creating sleep-friendly campus environments. Identifying the most effective and scalable interventions is critical to providing actionable insights for educators and policymakers seeking to improve student outcomes through better sleep.

References

Abraham, K. G. and Flood, S. M. (2009). American time use survey data extract builder (ATUS-X). *International Journal for Time Use Research*, 6:167–168.

Anderson, M. 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.

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

Angrist, J., Oreopoulos, P., and Williams, T. (2014). When opportunity knocks, who answers? New evidence on college achievement awards. *Journal of Human Resources*, 49(3):572–610.

Attia, P. (2023). *Outlive: The science and art of longevity*. Harmony.

Avery, M., Giuntella, O., and Jiao, P. (2022). Why don't we sleep enough? A field experiment among college students. *Review of Economics and Statistics, Forthcoming*.

Bachireddy, C., Joung, A., John, L. K., Gino, F., Tuckfield, B., Foschini, L., and Milkman, K. L. (2019). Effect of different financial incentive structures on promoting physical activity among adults: A randomized clinical trial. *JAMA Network Open*, 2(8):e199863–e199863.

Ballard, J. (2019). 40% of Americans don't generally wake up feeling well-rested. *YouGov*. Available at: https://today.yougov.com/topics/health/articles-reports/2019/03/13/sleep-habits-americans-survey-poll Accessed: March 2019.

Banks, S. and Dinges, D. F. (2007). Behavioral and physiological consequences of sleep restriction. *Journal of Clinical Sleep Medicine*, 3(5):519–528.

Barnes, C. M., Miller, J. A., and Bostock, S. (2017). Helping employees sleep well: Effects of cognitive behavioral therapy for insomnia on work outcomes. *Journal of Applied Psychology*, 102(1):104.

Barrow, L., Richburg-Hayes, L., Rouse, C. E., and Brock, T. (2014). Paying for performance: The education impacts of a community college scholarship program for low-income adults. *Journal of Labor Economics*, 32(3):563–599.

Basner, M., Fomberstein, K. M., Razavi, F. M., Banks, S., William, J. H., Rosa, R. R., and Dinges, D. F. (2007). American time use survey: sleep time and its relationship to waking activities. *Sleep*, 30(9):1085–1095.

Beshears, J., Lee, H. N., Milkman, K. L., Mislavsky, R., and Wisdom, J. (2021). Creating exercise habits using incentives: The trade-off between flexibility and routinization. *Management Science*, 67(7):4139–4171.

Bessone, P., Rao, G., Schilbach, F., Schofield, H., and Toma, M. (2021). The economic consequences of increasing sleep among the urban poor. *The Quarterly Journal of Economics*, 136(3):1887–1941.

Biddle, J. E. and Hamermesh, D. S. (1990). Sleep and the allocation of time. *Journal of Political Economy*, 98(5, Part 1):922–943.

Boyce, R., Glasgow, S. D., Williams, S., and Adamantidis, A. (2016). Causal evidence for the role of REM sleep theta rhythm in contextual memory consolidation. *Science*, 352(6287):812–816.

Braghieri, L., Levy, R., and Makarin, A. (2022). Social media and mental health. *American Economic Review*, 112(11):3660–3693.

Breig, Z., Gibson, M., and Shrader, J. (2020). Why do we procrastinate? present bias and optimism. *Present Bias and Optimism (August 27, 2020)*.

Byrne, D. P., Goette, L., Martin, L. A., Miles, A., Jones, A., Schob, S., Staake, T., and Tiefenbeck, V. (2022). The habit forming effects of feedback: Evidence from a large-scale field experiment. Available at SSRN: https://ssrn.com/abstract=3974371.

Cappelen, A. W., Charness, G., Ekström, M., Gneezy, U., and Tungodden, B. (2017). Exercise improves academic performance. *NHH Dept. of Economics Discussion Paper*, (08).

Cappuccio, F. P., D'Elia, L., Strazzullo, P., and Miller, M. A. (2010). Sleep duration and all-cause mortality: A systematic review and meta-analysis of prospective studies. *Sleep*, 33(5):585–592.

Carrell, S. E., Maghakian, T., and West, J. E. (2011). A's from zzzz's? The causal effect of school start time on the academic achievement of adolescents. *American Economic Journal: Economic Policy*, 3(3):62–81.

CDC (2023). About CDC: Sleep and sleep disorders. Accessed Nov 5, 2023. https://www.cdc.gov/sleep/data-and-statistics/adults.html.

Charness, G. and Gneezy, U. (2009). Incentives to exercise. *Econometrica*, 77(3):909–931.

Charness, G. and Grieco, D. (2019). Creativity and incentives. *Journal of the European Economic Association*, 17(2):454–496.

Choi, K., Lee, Y. J., Park, S., Je, N. K., and Suh, H. S. (2022). Efficacy of melatonin for chronic insomnia: Systematic reviews and meta-analyses. *Sleep medicine reviews*, 66:101692.

Clotfelter, C. T., Hemelt, S. W., and Ladd, H. F. (2018). Multifaceted aid for low-income students and college outcomes: Evidence from North Carolina. *Economic Inquiry*, 56(1):278–303.

Collis, A. and Eggers, F. (2022). Effects of restricting social media usage on wellbeing and performance: A randomized control trial among students. *PloS one*, 17(8):e0272416.

Cooney, M. T., Vartiainen, E., Laakitainen, T., Juolevi, A., Dudina, A., and Graham, I. M. (2010). Elevated resting heart rate is an independent risk factor for cardiovascular disease in healthy men and women. *American Heart Journal*, 159(4):612–619.

Corkett, S. (2010). 2020 sleep in america® poll shows alarming level of sleepiness and low levels of action. *National Sleep Foundation*.

Creswell, J. D., Tumminia, M. J., Price, S., Sefidgar, Y., Cohen, S., Ren, Y., Brown, J., Dey, A. K., Dutcher, J. M., Villalba, D., et al. (2023). Nightly sleep duration predicts grade point average in the first year of college. *Proceedings of the National Academy of Sciences*, 120(8):e2209123120.

De Paola, M., Scoppa, V., and Nistico, R. (2012). Monetary incentives and student achievement in a depressed labor market: Results from a randomized experiment. *Journal of Human Capital*, 6(1):56–85.

de Zambotti, M., Goldstone, A., Claudatos, S., Colrain, I. M., and Baker, F. C. (2018). A validation study of fitbit charge 2™ compared with polysomnography in adults. *Chronobiology International*, 35(4):465–476.

Denning, J. T., Marx, B. M., and Turner, L. J. (2019). Propelled: The effects of grants on graduation, earnings, and welfare. *American Economic Journal: Applied Economics*, 11(3):193–224.

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 308(1135):67–78.

Dickinson, D. L. and Masclet, D. (2023). Unethical decision making and sleep restriction: Experimental evidence. *Games and Economic Behavior*, 141:484–502.

Dickinson, D. L. and McElroy, T. (2017). Sleep restriction and circadian effects on social decisions. *European Economic Review*, 97:57–71.

Diekelmann, S. and Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*, 11(2):114–126.

Dyer, A. R., Persky, V., Stamler, J., Paul, O., Shekelle, R. B., Berkson, D. M., Lepper, M., Schoenberger, J. A., and Lindberg, H. A. (1980). Heart rate as a prognostic factor for coronary heart disease and mortality: Findings in three Chicago epidemiologic studies. *American Journal of Epidemiology*, 112(6):736–749.

Erwin, C., Binder, M., Miller, C., and Krause, K. (2021). Performance-based aid, enhanced advising, and the income gap in college graduation: Evidence from a randomized controlled trial. *Educational Evaluation and Policy Analysis*, 43(1):134–153.

Evans, W. N., Kearney, M. S., Perry, B., and Sullivan, J. X. (2020). Increasing community college completion rates among low-income students: Evidence from a randomized controlled trial evaluation of a case-management intervention. *Journal of Policy Analysis and Management*, 39(4):930–965.

Fischer, D., Klerman, E. B., and Phillips, A. J. (2021). Measuring sleep regularity: theoretical properties and practical usage of existing metrics. *Sleep*, 44(10):zsab103.

Gibson, M. and Shrader, J. (2018). Time use and labor productivity: The returns to sleep. *Review of Economics and Statistics*, 100(5):783–798.

Giuntella, O., Han, W., and Mazzonna, F. (2017). Circadian rhythms, sleep, and cognitive skills: Evidence from an unsleeping giant. *Demography*, 54(5):1715–1742.

Giuntella, O., Hyde, K., Saccardo, S., and Sadoff, S. (2021). Lifestyle and mental health disruptions during COVID-19. *Proceedings of the National Academy of Sciences*, 118(9):e2016632118.

Giuntella, O. and Mazzonna, F. (2019). Sunset time and the economic effects of social jetlag: Evidence from US time zone borders. *Journal of Health Economics*, 65:210–226.

Giuntella, O., Saccardo, S., and Sadoff, S. (2024). Online appendix. https://tinyurl.com/m9khyrht.

Gneezy, U., Meier, S., and Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, 25(4):191–210.

Goldrick-Rab, S., Kelchen, R., Harris, D. N., and Benson, J. (2016). Reducing income inequality in educational attainment: Experimental evidence on the impact of financial aid on college completion. *American Journal of Sociology*, 121(6):1762–1817.

Groen, J. A. and Pabilonia, S. W. (2019). Snooze or lose: High school start times and academic achievement. *Economics of Education Review*, 72:204–218.

Group, A. S. W., ADOLESCENCE, C. O., HEALTH, C. O. S., Au, R., Carskadon, M., Millman, R., Wolfson, A., Braverman, P. K., Adelman, W. P., Breuner, C. C., et al. (2014). School start times for adolescents. *Pediatrics*, 134(3):642–649.

Hafner, M., Stepanek, M., Taylor, J., Troxel, W. M., and Van Stolk, C. (2017). Why sleep matters—the economic costs of insufficient sleep: A cross-country comparative analysis. *Rand Health Quarterly*, 6(4).

Haghayegh, S., Khoshnevis, S., Smolensky, M. H., Diller, K. R., and Castriotta, R. J. (2019). Accuracy of wristband fitbit models in assessing sleep: Systematic review and meta-analysis. *Journal of Medical Internet Research*, 21(11):e16273.

Heissel, J. A. and Norris, S. (2018). Rise and shine the effect of school start times on academic performance from childhood through puberty. *Journal of Human Resources*, 53(4):957–992.

Hershner, S. D. and Chervin, R. D. (2014). Causes and consequences of sleepiness among college students. *Nature and Science of Sleep*, pages 73–84.

Hillman, D. R., Murphy, A. S., Antic, R., and Pezzullo, L. (2006). The economic cost of sleep disorders. *Sleep*, 29(3):299–305.

Hirshkowitz, M., Whiton, K., Albert, S. M., Alessi, C., Bruni, O., DonCarlos, L., Hazen, N., Herman, J., Hillard, P. J. A., Katz, E. S., et al. (2015). National sleep foundation's updated sleep duration recommendations. *Sleep Health*, 1(4):233–243.

Holbein, J. B., Schafer, J. P., and Dickinson, D. L. (2019). Insufficient sleep reduces voting and other prosocial behaviours. *Nature Human Behaviour*, 3(5):492–500.

Huffington, A. (2016). *The sleep revolution: Transforming your life, one night at a time*. Harmony.

Hussam, R. N., Rabbani, A., Reggiani, G., and Rigol, N. (2022). Rational habit formation: Experimental evidence from handwashing in India. *American Economic Journal: Applied Economics*, 14(1):1–41.

Jagnani, M. (2021). Children's sleep and human capital production. *The Review of Economics and Statistics, Forthcoming*.

Jike, M., Itani, O., Watanabe, N., Buysse, D. J., and Kaneita, Y. (2018). Long sleep duration and health outcomes: A systematic review, meta-analysis and meta-regression. *Sleep Medicine Reviews*, 39:25–36.

Jin, L. and Ziebarth, N. R. (2020). Sleep, health, and human capital: Evidence from daylight saving time. *Journal of Economic Behavior & Organization*, 170:174–192.

Jones, J. (2013). In U.S., 40% get less than recommended amount of sleep. *Gallup*. Available at: https://news.gallup.com/poll/166553/less-recommended-amount-sleep.aspx.

Killgore, W. D. (2010). Effects of sleep deprivation on cognition. *Progress in Brain Research*, 185:105–129.

Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K., and Rathouz, P. J. (2008). Self-reported and measured sleep duration: How similar are they? *Epidemiology*, 19(6):838–845.

Lavecchia, A. M., Liu, H., and Oreopoulos, P. (2016). Behavioral economics of education: Progress and possibilities. In *Handbook of the Economics of Education*, volume 5, pages 1–74. Elsevier.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102.

Leproult, R. and Van Cauter, E. (2010). Role of sleep and sleep loss in hormonal release and metabolism. *Pediatric Neuroendocrinology*, 17:11–21.

Levitt, S. D., List, J. A., Neckermann, S., and Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8(4):183–219.

Lim, J. and Dinges, D. F. (2010). A meta-analysis of the impact of short-term sleep deprivation on cognitive variables. *Psychological bulletin*, 136(3):375.

Lindquist, S. and Sadoff, S. (2023). Understanding the interactions of sleep, social media and mental health for productivity and performance: The role of field experiments. *Annual Proceedings of the Upton Forum*.

Lintner, T. (2024). Effects of performance-based financial incentives on higher education students: A meta-analysis using causal evidence. *Educational Research Review*, page 100621.

Liu, T. X., Malmendier, U., Wang, S. W., and Zhang, S. (2022). Not too early, not too late: Encouraging engagement in education [working paper]. Available at: https://eml.berkeley.edu/~ulrike/Papers/NotesTaking_Public.pdf.

Lusher, L., Yasenov, V., and Luong, P. (2019). Does schedule irregularity affect productivity? Evidence from random assignment into college classes. *Labour Economics*, 60:115–128.

MacDonald, H., Malatest, R., Assels, R., Baroud, R., and Gong, L. (2009). *Final Impacts Report: Foundations for Success Project*. Canada Millennium Scholarship Foundation.

Mah, J. and Pitre, T. (2021). Oral magnesium supplementation for insomnia in older adults: a systematic review & meta-analysis. *BMC complementary medicine and therapies*, 21:1–11.

Marks, G. A., Shaffery, J. P., Oksenberg, A., Speciale, S. G., and Roffwarg, H. P. (1995). A functional role for rem sleep in brain maturation. *Behavioural Brain Research*, 69(1-2):1–11.

McKenna, B. S., Dickinson, D. L., Orff, H. J., and Drummond, S. P. (2007). The effects of one night of sleep deprivation on known-risk and ambiguous-risk decisions. *Journal of Sleep Research*, 16(3):245–252.

Milkman, K. L., Minson, J. A., and Volpp, K. G. (2014). Holding the hunger games hostage at the gym: An evaluation of temptation bundling. *Management Science*, 60(2):283–299.

Mullainathan, S. (2014). Get some sleep, and wake up the G.D.P. *The New York Times*. Available at: https://www.nytimes.com/2014/02/02/business/get-some-sleep-and-wake-up-the-gdp.html.

Oreopoulos, P. and Petronijevic, U. (2018). Student coaching: How far can technology go? *Journal of Human Resources*, 53(2):299–329.

Panel, C. C., Watson, N. F., Badr, M. S., Belenky, G., Bliwise, D. L., Buxton, O. M., Buysse, D., Dinges, D. F., Gangwisch, J., Grandner, M. A., et al. (2015). Recommended amount of sleep for a healthy adult: A joint consensus statement of the american academy of sleep medicine and sleep research society. *Journal of Clinical Sleep Medicine*, 11(6):591–592.

Park, R. S. E. and Scott-Clayton, J. (2018). The impact of Pell Grant eligibility on community college students' financial aid packages, labor supply, and academic outcomes. *Educational Evaluation and Policy Analysis*, 40(4):557–585.

Patel, R. and Valenzuela, I. (2013). Moving forward: Early findings from the performance-based scholarship demonstration in arizona. the performance-based scholarship demonstration. *MDRC*.

Phillips, A. J., Clerx, W. M., O'Brien, C. S., Sano, A., Barger, L. K., Picard, R. W., Lockley, S. W., Klerman, E. B., and Czeisler, C. A. (2017). Irregular sleep/wake patterns are associated with poorer academic performance and delayed circadian and sleep/wake timing. *Scientific reports*, 7(1):3216.

Pope, N. G. (2016). How the time of day affects productivity: Evidence from school schedules. *Review of Economics and Statistics*, 98(1):1–11.

Quer, G., Gouda, P., Galarnyk, M., Topol, E. J., and Steinhubl, S. R. (2020). Inter-and intraindividual variability in daily resting heart rate and its associations with age, sex, sleep, bmi, and time of year: Retrospective, longitudinal cohort study of 92,457 adults. *Plos one*, 15(2):e0227709.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401.

Rao, G., Redline, S., Schilbach, F., Schofield, H., and Toma, M. (2021). Informing sleep policy through field experiments. *Science*, 374(6567):530–533.

Richburg-Hayes, L., Patel, R., Brock, T., de la Campa, E., Rudd, T., and Valenzuela, I. (2015). Providing more cash for college: Interim findings from the performance-based scholarship demonstration in california. *New York: MDRC*.

Roenneberg, T. (2013). The human sleep project. *Nature*, 498(7455):427–428.

Roenneberg, T. and Merrow, M. (2016). The circadian clock and human health. *Current biology*, 26(10):R432–R443.

Royer, H., Stehr, M., and Sydnor, J. (2015). Incentives, commitments, and habit formation in exercise: Evidence from a field experiment with workers at a fortune-500 company. *American Economic Journal: Applied Economics*, 7(3):51–84.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

Scott-Clayton, J. (2011). On money and motivation a quasi-experimental analysis of financial incentives for college achievement. *Journal of Human Resources*, 46(3):614–646.

Smarr, B. L. (2015). Digital sleep logs reveal potential impacts of modern temporal structure on class performance in different chronotypes. *Journal of Biological Rhythms*, 30(1):61–67.

Sommo, C., Boynton, M., Collado, H., Diamond, J., Gardenhire, A., Ratledge, A., Rudd, T., and Weiss, M. J. (2014). Mapping success: Performance-based scholarships, student services, and developmental math at hillsborough community college. *New York: MDRC, October*.

Verplanken, B. and Wood, W. (2006). Interventions to break and create consumer habits. *Journal of Public Policy & Marketing*, 25(1):90–103.

Volpp, K. G. and Loewenstein, G. (2020). What is a habit? Diverse mechanisms that can produce sustained behavior change. *Organizational Behavior and Human Decision Processes*, 161:36–38.

Walker, M. (2017). *Why we sleep: Unlocking the power of sleep and dreams*. Simon and Schuster.

Wellsjo, A. S. (2021). Simple actions, complex habits: Lessons from hospital hand hygiene [working paper]. Available at: https://economics.ucr.edu/wp-content/uploads/2023/01/1-18-23-wellsjo.pdf.

Wheaton, A. G., Jones, S. E., Cooper, A. C., and Croft, J. B. (2018). Short sleep duration among middle school and high school students—united states, 2015. *Morbidity and Mortality Weekly Report*, 67(3):85.

Williams, N. (2014). The GAD-7 questionnaire. *Occupational Medicine*, 64(3):224–224.

Windred, D. P., Burns, A. C., Lane, J. M., Saxena, R., Rutter, M. K., Cain, S. W., and Phillips, A. J. (2024). Sleep regularity is a stronger predictor of mortality risk than sleep duration: a prospective cohort study. *Sleep*, 47(1):zsad253.

Wood, W. and Neal, D. T. (2007). A new look at habits and the habit-goal interface. *Psychological Review*, 114(4):843.

Wood, W. and Rünger, D. (2016). Psychology of habit. *Annual Review of Psychology*, 67:289–314.

Wright, S. P., Hall Brown, T. S., Collier, S. R., and Sandberg, K. (2017). How consumer physical activity monitors could transform human physiology research. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 312(3):R358–R367.

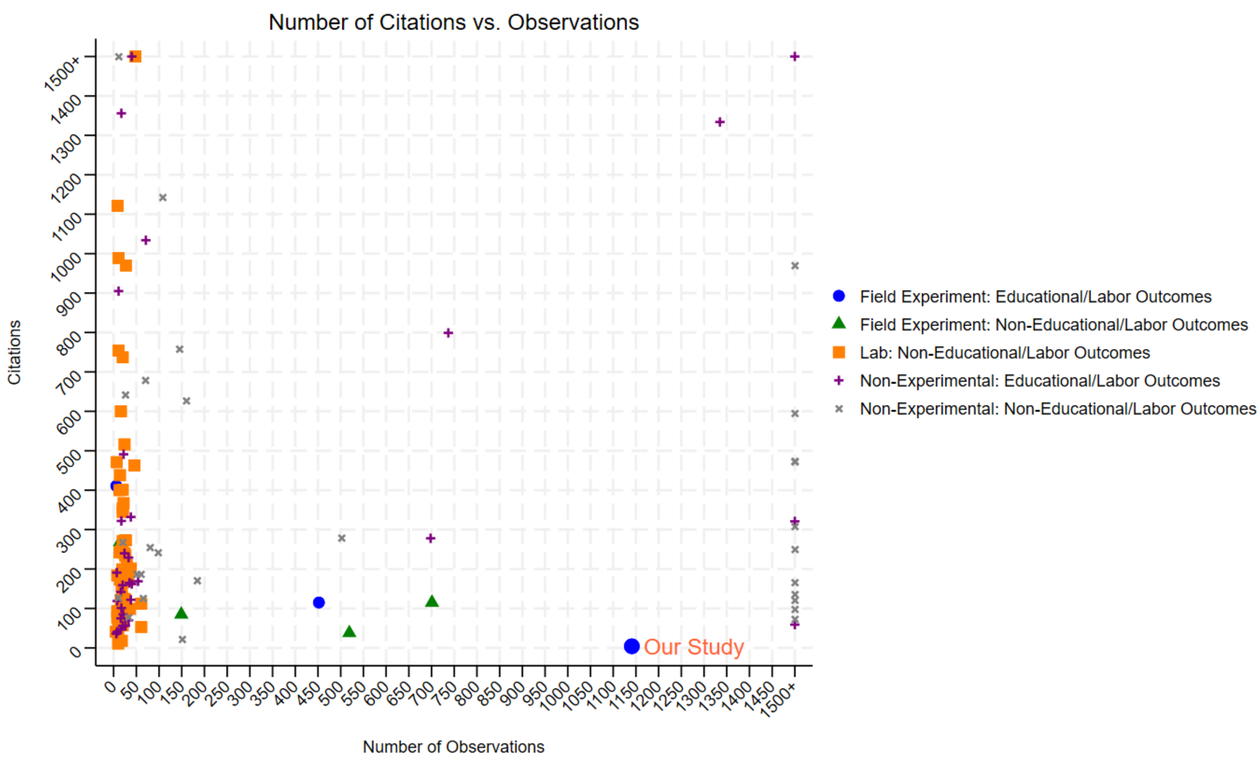
Xiang, T., Cai, Y., Hong, Z., and Pan, J. (2021). Efficacy and safety of zolpidem in the treatment of insomnia disorder for one month: a meta-analysis of a randomized controlled trial. *Sleep Medicine*, 87:250–256.

Zhang, Y. and Gruber, R. (2019). Can slow-wave sleep enhancement improve memory? A review of current approaches and cognitive outcomes. *The Yale Journal of Biology and Medicine*, 92(1):63–80.

Appendix

A. Figures and tables

Figure A.1: Previous work on sleep



Notes: The figure illustrates the relationship between the number of citations and the number of observations across field experiments, lab experiments and non-experimental studies. We further classify studies by outcome measure (include or do not include educational and/or labor outcomes). Citations are measured using Google Scholar, as of October 15 2024. Studies with citations or observations exceeding 1500 are grouped under "1500+" to improve readability. We built this figure by reviewing 123 studies referenced in one or more of the following: Peter Attia's book "Outlive" (Attia, 2023) (chapter 16 which is devoted to sleep), Arianna Huffington's book "The Sleep Revolution" (Huffington, 2016) (chapters 11 and 12 focusing on sleep and performance), Bessone et al. (2021), Creswell et al. (2023), and the Lim and Dinges (2010) meta-analysis.

Figure A.2: Timeline of the experiment

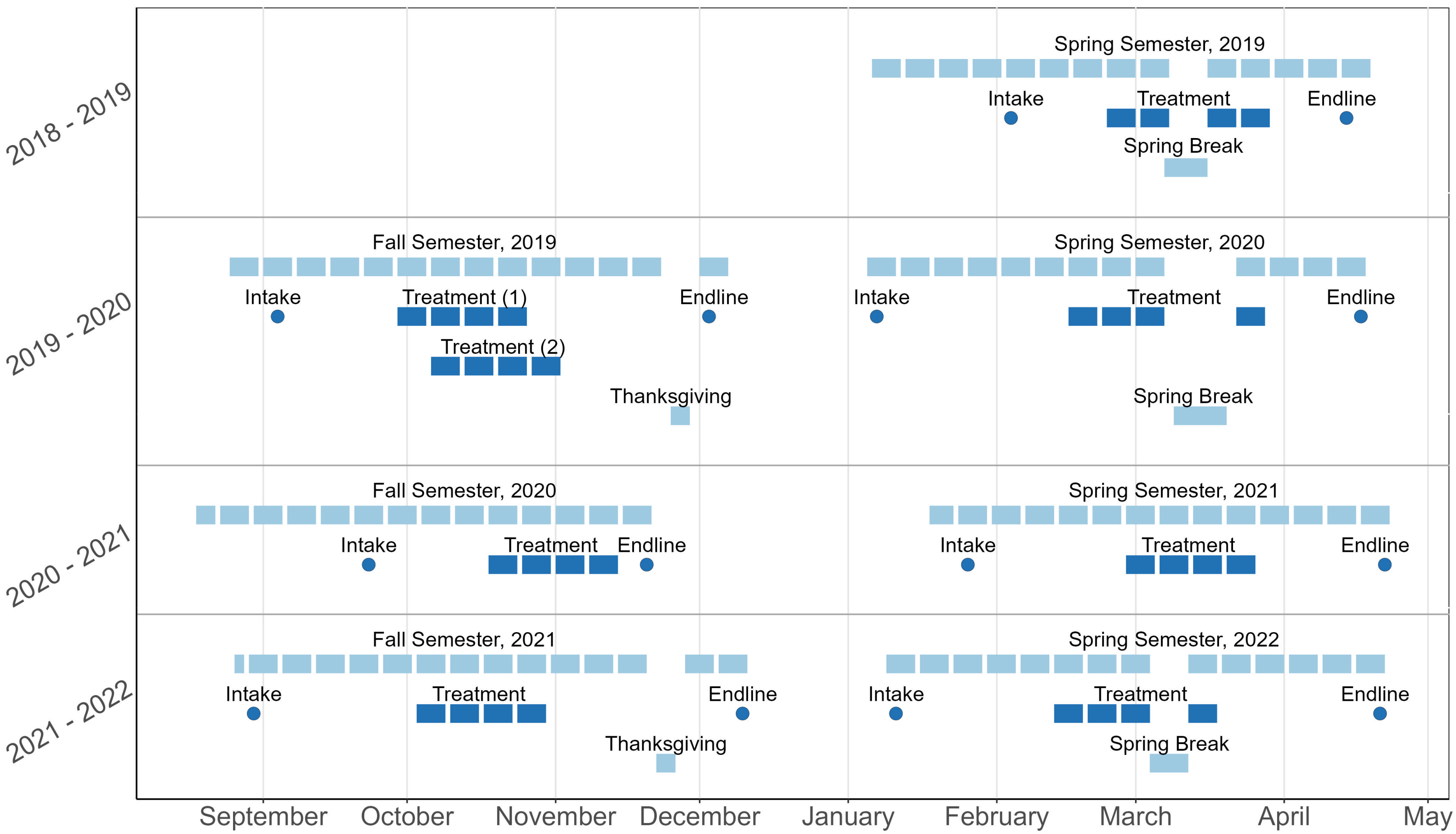
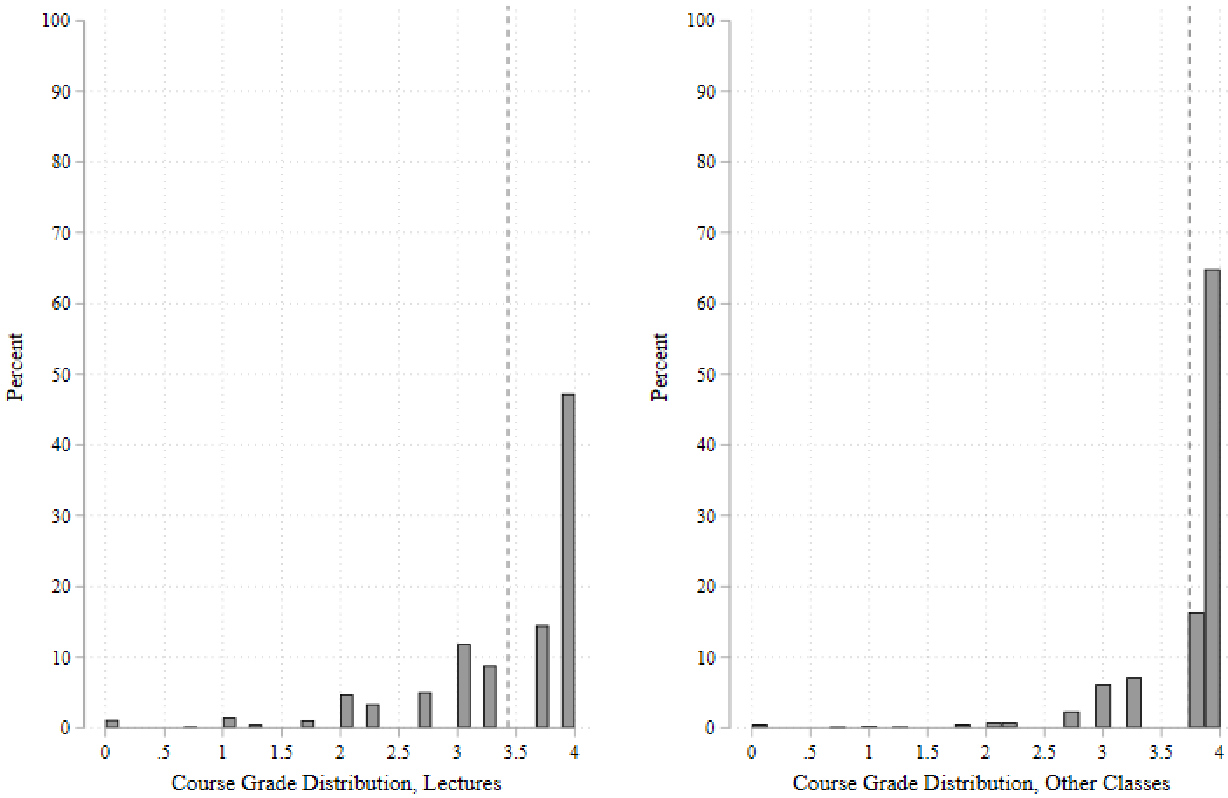


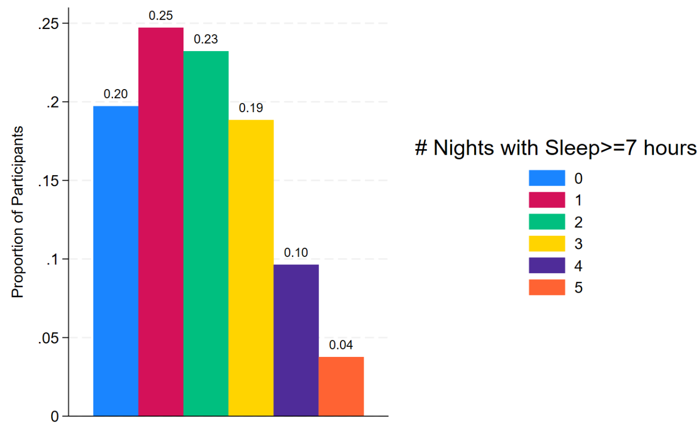
Figure A.3: Grades distribution



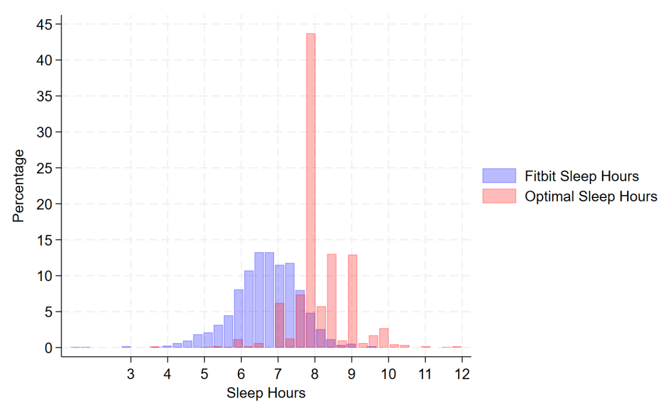
Notes: The figure reports the distribution of grades in lectures and other classes. The dashed vertical line identifies the average grade in these class types.

Figure A.4: Sleep measures

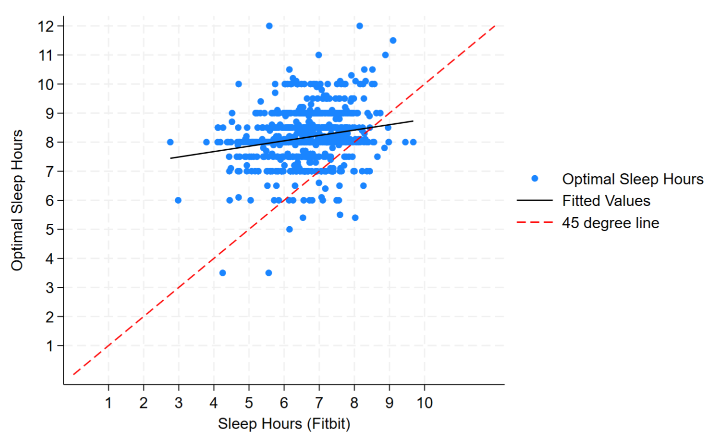
Panel A



Panel B

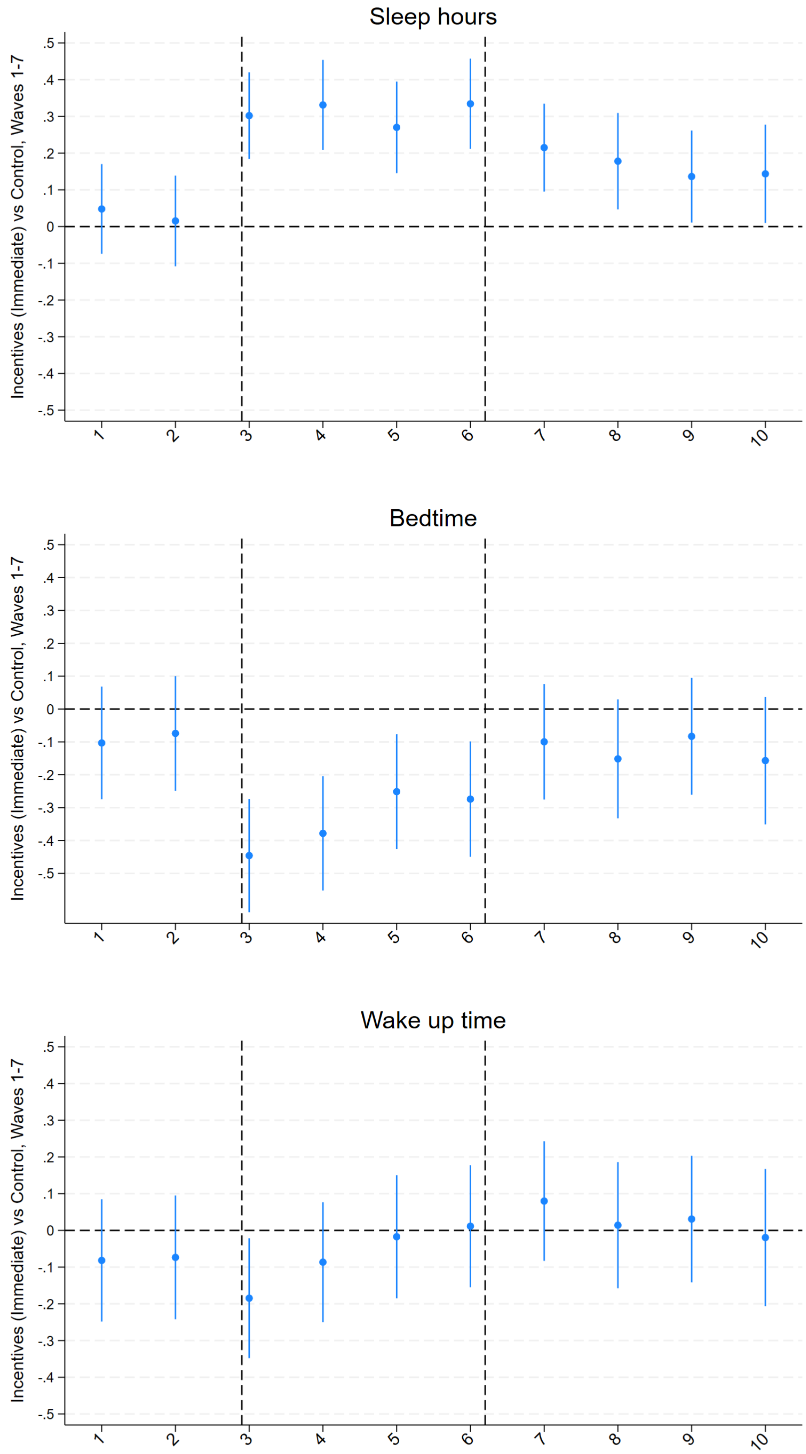


Panel C



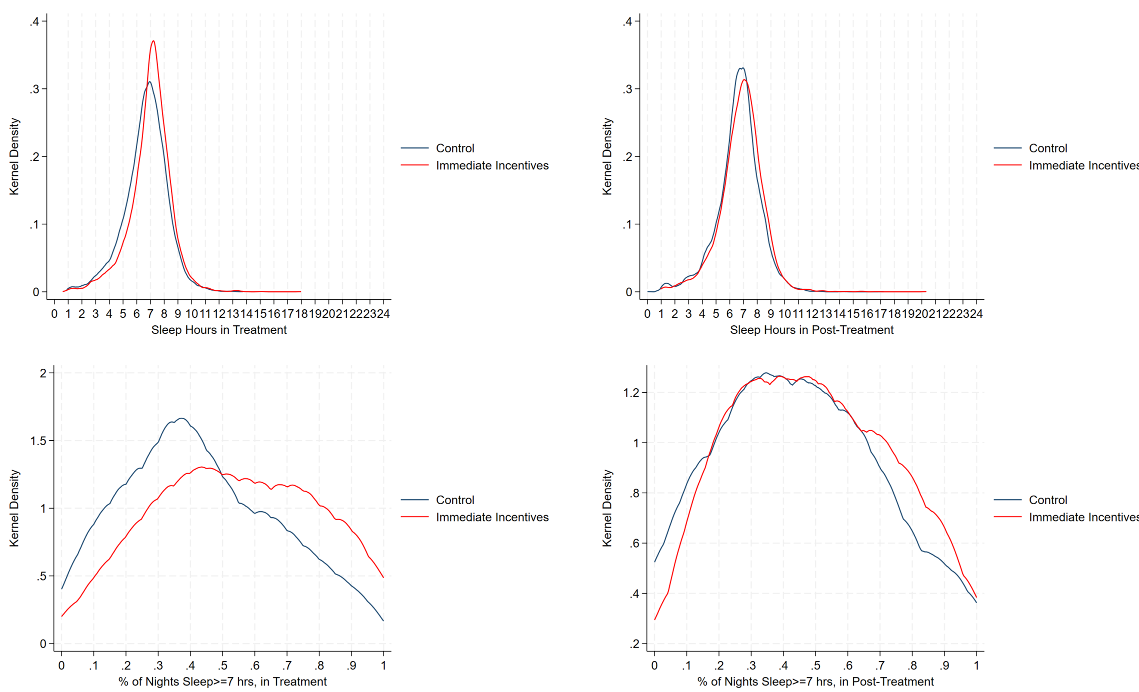
Notes: Panel A shows the distribution of nights individuals slept more than 7 hours at baseline. Panel B plots the distribution of actual (measured via Fitbit) and optimal (self-reported) sleep hours for participants in the sample at baseline. Panel C is a scatterplot that compares individual-level actual sleep hours (measured via Fitbit; x-axis) to optimal sleep hours (y-axis). The red dashed 45-degree line represents a scenario where actual and optimal sleep are equal. Points above the line indicate individuals who sleep less than they believe they should.

Figure A.5: Immediate Incentives, sleep hours, bedtime and wake-up time



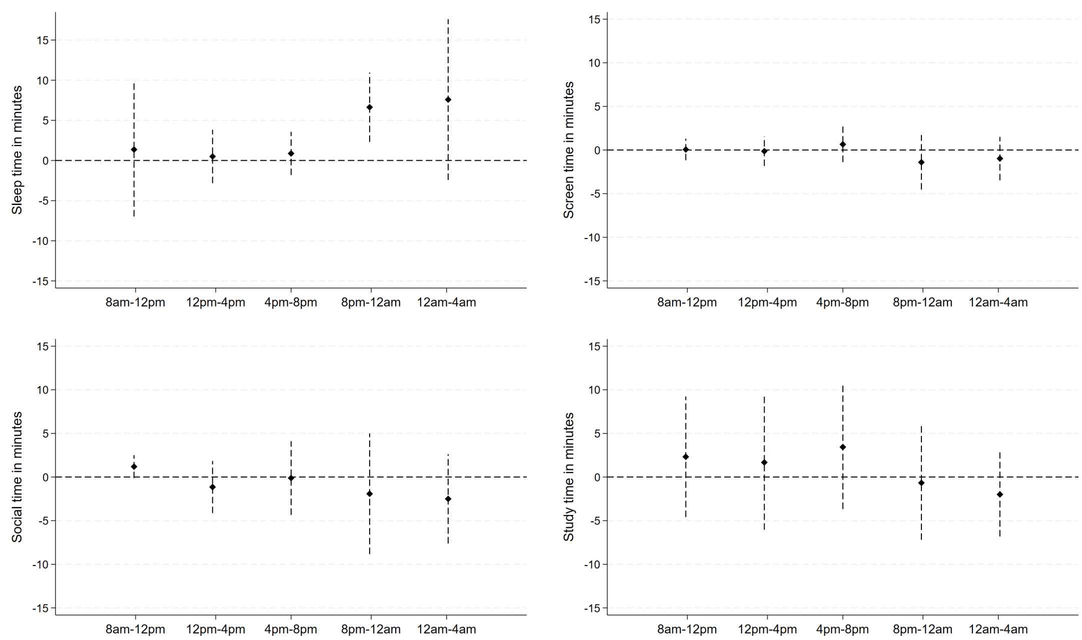
Notes: The sample is restricted to weekdays (Sunday-Thursday nights). On the horizontal axis we report time in weeks since the study started (week 3 is the first week of treatment, week 6 is the last week of treatment). The coefficient reports the differences in average sleep hours, bedtime, and wake-up time between individuals in the Immediate Incentives treatment and those in Control by week. Standard errors are clustered at the individual level. Bars indicate 95% confidence intervals.

Figure A.6: Immediate incentives and distribution of sleep during and after the intervention



Notes: The figure reports kernel densities of the Immediate Incentives (red) and Control (navy) groups for a) sleep hours (top panel) and b) the proportion of nights with sleep over seven hours (bottom panel), during treatment (leftward graphs) and post-treatment (rightward graphs).

Figure A.7: Immediate Incentives to sleep and time use over the day: Post-Intervention period



Notes: The figure reports differences between participants in the Immediate Incentives treatment and Control groups in the minutes allocated to different time-use activities post-treatment throughout the day. All the coefficients are obtained from regressions including wave, month and day of the week fixed effects, baseline value of the outcome variable, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Standard errors are clustered at the individual level. Bars indicate 95% confidence intervals.

Table A.1: Grading system

(1) Grade	(2) GPA	(3) Quality points	(4) Has a grade	(5) Withdrawn	(6) Passed	(7) Credit completed
A+	YES	4	YES	NO	YES	YES
A	YES	4	YES	NO	YES	YES
A-	YES	3.75	YES	NO	YES	YES
B+	YES	3.25	YES	NO	YES	YES
B	YES	3	YES	NO	YES	YES
B-	YES	2.75	YES	NO	YES	YES
C+	YES	2.25	YES	NO	YES	YES
C	YES	2	YES	NO	YES	YES
C-	YES	1.75	YES	NO	YES	YES
D+	YES	1.25	YES	NO	YES	YES
D	YES	1	YES	NO	YES	YES
D-	YES	0.75	YES	NO	YES	YES
F	YES	0	YES	NO	NO	NO
G	–	–	NO	NO	NO	NO
H	–	–	YES	NO	YES	YES
HS	–	–	YES	NO	YES	YES
I	–	–	NO	NO	NO	NO
N	–	–	NO	NO	NO	NO
NC	–	–	NO	NO	NO	NO
NG	–	–	NO	NO	NO	NO
R	–	–	NO	NO	NO	NO
S	–	–	YES	NO	YES	YES
U	–	–	YES	NO	NO	NO
W	–	–	NO	YES	NO	NO

Notes: Non-grade outcomes (G-W) represent the following: G, unfinished or ongoing course work due extenuating personal circumstances; H, honors, exceptional completion of coursework; HS, highly satisfactory completion of coursework, used only by School of Medicine; I, unfinished or ongoing course work due to nature of course; N, non-credited or graded course, such as a course audit; NC, non-credit course; NG, non-credit course due unfinished course work; R, student resigned from University; S, satisfactory completion of requirements; U, unsatisfactory completion of class requirements; W, student withdrew from course. Source:https://www.registrar.pitt.edu/sites/default/files/pdf/Grading%20System.pdf

Table A.2: Incentives and attrition

	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A			
	# synced days before intervention	# synced days during intervention	# synced days post intervention	Has HS GPA	Has baseline GPA	Has course grades
Immediate incentives	0.179	0.703**	-0.093	0.006	0.032	0.008
	(0.258)	(0.321)	(0.405)	(0.010)	(0.024)	(0.010)
Observations	840	840	840	840	840	840
Mean of Dep. Var.	10.39	17.07	12.29	0.932	0.786	0.982
Std. dev.	5.761	5.870	7.446	0.252	0.411	0.133
			Panel B: Baseline			
	Has time use	Has math task	Has creativity task	Has mood survey	Has resilience survey	Has mental health
Immediate incentives	-0.015	0.019	-0.011	0.018	0.009	0.009
	(0.016)	(0.024)	(0.024)	(0.013)	(0.014)	(0.014)
Observations	840	840	840	840	840	840
Mean of Dep. Var.	0.902	0.714	0.831	0.718	0.707	0.958
Std. dev.	0.297	0.452	0.375	0.450	0.455	0.200
			Panel C: During or After the Intervention			
	Has time use	Has math task	Has creativity task	Has mood survey	Has resilience survey	Has mental health
Immediate incentives	0.005	-0.014	0.025	0.014	0.001	-0.036
	(0.015)	(0.021)	(0.022)	(0.016)	(0.017)	(0.028)
Observations	840	840	840	840	840	840
Mean of Dep. Var.	0.956	0.900	0.894	0.937	0.927	0.718
Std. dev.	0.205	0.300	0.308	0.243	0.260	0.450

Notes: The table reports the difference between the Immediate Incentives treatment and Control groups in attrition rate across the different outcome measures. All estimates include wave fixed effects. Robust standard errors are in parenthesis. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

Table A.3: Immediate incentives and sleep: Sensitivity analysis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A:	Primary Specification	Basic Controls	Has term GPA	Excludes missing nights	Excludes wave 3	Weighted by gender	Incentives pooled
Treatment	0.115***	0.110***	0.114***	0.127***	0.117***	0.110***	0.105***
	(0.013)	(0.011)	(0.013)	(0.012)	(0.013)	(0.013)	(0.011)
Post-Treatment	0.052***	0.058***	0.052***	0.053***	0.061***	0.053***	0.049***
	(0.015)	(0.011)	(0.015)	(0.014)	(0.015)	(0.015)	(0.014)
Observations	39,035	39,035	38,330	33,052	35,570	39,035	48,195
R-squared	0.214	0.209	0.211	0.157	0.216	0.213	0.211
Mean of Dep. Var.	0.438	0.438	0.437	0.434	0.442	0.438	0.435
Std. dev.	0.496	0.496	0.496	0.496	0.497	0.496	0.496
Number of individuals	840	840	825	840	763	840	1040
Panel B:	No cue in Post	Lower Lee Bounds	Upper Lee Bounds	Excluding those who knew someone in the experiment	Excluding those who knew someone receiving rewards or feedback	Excluding those who knew someone receiving rewards	Baseline sample controlling for (4-6)
Treatment	0.121***	0.096***	0.034***	0.088***	0.107***	0.113***	0.110***
	(0.021)	(0.009)	(0.011)	(0.017)	(0.015)	(0.014)	(0.013)
Post-Treatment	0.065***	0.155***	0.061***	0.035*	0.050***	0.055***	0.049***
	(0.025)	(0.009)	(0.012)	(0.019)	(0.017)	(0.016)	(0.016)
Immediate Incentives*	0.008						
No Cue in Post-Treatment	(0.027)						
Observations	17,850	16,800	13,835	24,915	31,520	33,980	37,145
R-squared	0.233			0.238	0.224	0.220	0.217
Mean of Dep. Var.	0.456	0.434	0.434	0.440	0.441	0.433	0.434
Std. dev.	0.498	0.263	0.263	0.496	0.497	0.495	0.496
Number of individuals	357	840	840	529	677	729	798

Notes: The sample in all columns except panel A column 7 is restricted to individuals in the Immediate Incentives treatment and individuals in the Control group. Column 7 includes individuals from the pooled incentives treatments and individuals in the Control group. Individuals in the Cue/Feedback treatment were not included in this analysis. All estimates except those in panel A column 2 include day of the week, week of the experiment, wave, and month fixed effects, baseline value of the outcome variable, indicators for the number of classes starting at 10 am or earlier, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Estimates in column 2 include only wave fixed effects, baseline value of the outcome variable, controls for gender, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). Panel A column 4 does not replace missing nights with baseline data as in our main analysis. Panel B column 1 restricts the analysis to waves 5 and 7. Columns 9 and 10 restrict the analysis to the intervention and post-intervention period, respectively. Panel B columns 4 - 7 restrict the analysis to waves 2 - 7. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

Table A.4: Immediate Incentives and sleep: By quartiles of sleep at baseline

	(1)	(2)	(3)	(4)
	Q1	Q2	Q3	Q4
	Panel A: Sleep≥7			
Treatment	0.149***	0.108***	0.149***	0.095***
	(0.023)	(0.024)	(0.022)	(0.026)
Post-Treatment	0.094***	0.033	0.039	0.025
	(0.021)	(0.028)	(0.030)	(0.030)
Observations	10,020	9,755	8,143	8,287
Mean of Dep. Var.	0.110	0.318	0.518	0.799
Std. dev.	0.313	0.466	0.500	0.401
Number of individuals	213	211	206	210
	Panel B: Sleep hours			
Treatment	0.368***	0.344***	0.406***	0.272***
	(0.072)	(0.077)	(0.070)	(0.083)
Post-Treatment	0.216***	0.188**	0.206**	0.075
	(0.072)	(0.087)	(0.092)	(0.089)
Observations	10,020	8,266	8,143	8,287
Mean of Dep. Var.	5.779	6.382	6.912	7.642
Std. dev.	1.380	1.420	1.509	1.079
Number of individuals	213	211	206	210

Notes: The sample is restricted to individuals in the Immediate Incentives treatment and individuals in the Control group. Individuals in the Cue/Feedback treatment were not included in this analysis. All estimates include day of the week, week of the experiment, wave, and month fixed effects, baseline value of the outcome variable, indicators for the number of classes starting at 10 a.m. or earlier, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable in the control group at baseline. Std. dev. is the standard deviation of the dependent variable in the control group at baseline. *** p<0.01, ** p<0.05, * p<0.1.

Table A.5: Immediate incentives and sleep: Additional outcomes

Panel A:	(1) Any nap	(2) Sleep ≥ 7 weekends	(3) Sleep ≥ 7 weekends & holidays	(4) Sleep ≥ 7 all nights & naps	(5) Sleep hours all nights & naps
Treatment	0.004	0.018	0.020	0.081***	0.248***
	(0.004)	(0.015)	(0.015)	(0.011)	(0.032)
Post-Treatment	-0.005	0.045***	0.047***	0.041***	0.098***
	(0.004)	(0.016)	(0.017)	(0.012)	(0.031)
Observations	56,166	14,170	17,131	56,166	56,166
Mean of Dep. Var.	0.0506	0.508	0.508	0.560	7.199
Std. dev.	0.212	0.500	0.500	0.496	1.614
Number of individuals	840	840	840	840	840

Panel B:	Sleep ≥ 6	Sleep 7-9	Efficiency	REM sleep	Deep sleep
Treatment	0.080***	0.109***	0.249*	2.533***	0.448
	(0.009)	(0.011)	(0.138)	(0.769)	(0.650)
Post-Treatment	0.036***	0.047***	0.069	1.763**	0.129
	(0.010)	(0.012)	(0.175)	(0.869)	(0.698)
Observations	39,035	39,035	39,035	35,115	35,115
Mean of Dep. Var.	0.711	0.384	93.50	84.22	74.59
Std. dev.	0.431	0.462	5.623	31.89	24.90
Number of individuals	840	840	840	798	798

Notes: The sample is restricted to individuals in the Immediate Incentives treatment and individuals in the Control group. Individuals in the Cue/Feedback treatment were not included in this analysis. All estimates include day of the week, week of the experiment, wave, and month fixed effects, baseline value of the outcome variable, indicators for the number of classes starting at 10 am or earlier, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

Table A.6: Immediate Incentives and course grade: Sensitivity analysis

	(1) Primary specification	(2) Basic controls	(3) Has grade in term+1 or term+2	(4) No missing HS/baseline GPA	(5) Excludes obs with no sleep data	(6) Excludes wave 3 (Covid)	(7) Weighted by gender	(8) Incentives, pooled
				Panel A: All classes				
Incentives	0.075**	0.064*	0.060*	0.067*	0.075**	0.090**	0.070*	0.061*
	(0.037)	(0.038)	(0.035)	(0.037)	(0.037)	(0.040)	(0.039)	(0.035)
Observations	4,300	4,300	4,102	4,216	4,256	3,934	4,300	5,254
Mean of dep. var.	3.502	3.502	3.500	3.502	3.499	3.491	3.502	3.498
Std. dev.	0.763	0.763	0.759	0.765	0.766	0.776	0.763	0.763
Number of individuals	833	833	791	815	825	757	833	1027
				Panel B: Lectures				
Incentives	0.088**	0.076*	0.074*	0.078*	0.088**	0.105**	0.082*	0.075*
	(0.042)	(0.043)	(0.040)	(0.042)	(0.042)	(0.045)	(0.044)	(0.040)
Observations	3,413	3,413	3,255	3,340	3,382	3,130	3,413	4,197
Mean of dep. var.	3.436	3.436	3.435	3.435	3.434	3.423	3.436	3.435
Std. dev.	0.805	0.805	0.799	0.807	0.807	0.819	0.805	0.801
Number of individuals	827	827	787	809	819	752	827	1021
				Panel C: Other classes (seminars, labs, etc.)				
Incentives	0.001	-0.008	-0.004	0.001	0.001	0.012	0.013	-0.012
	(0.040)	(0.040)	(0.045)	(0.040)	(0.040)	(0.041)	(0.045)	(0.036)
Observations	887	887	726	876	874	804	887	1,057
Mean of dep. var.	3.753	3.753	3.759	3.755	3.751	3.753	3.753	3.748
Std. dev.	0.505	0.505	0.503	0.503	0.507	0.502	0.505	0.517
Number of individuals	562	562	449	554	554	510	562	674

Notes: All estimates except those in column 2 include demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), baseline sleep, indicators for the number of classes starting at 10 a.m. or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). Estimates in column 2 include only wave fixed effects, baseline sleep, controls for gender, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Observations are weighted by the number of credits taken in the semester. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A.7: Immediate Incentives, sleep and GPA: Heterogeneity

	(1) Male	(2) Female	(3) First-term	(4) Other students	(5) No-STEM major	(6) STEM major
			Panel A: Sleep $\geq$ 7 hours			
Treatment	0.088***	0.129***	0.150***	0.108***	0.107***	0.125***
	(0.024)	(0.015)	(0.029)	(0.014)	(0.018)	(0.017)
Post-Treatment	0.056**	0.054***	0.122***	0.042**	0.045*	0.061***
	(0.026)	(0.018)	(0.034)	(0.016)	(0.024)	(0.019)
Observations	10,615	28,225	6,455	32,580	16,545	22,295
R-squared	0.189	0.212	0.189	0.222	0.214	0.213
Mean of Dep. Var.	0.359	0.469	0.421	0.442	0.474	0.413
Std. dev.	0.480	0.499	0.494	0.497	0.499	0.492
Number of individuals	229	607	160	680	356	480
			Panel B: Course grades, all classes			
Immediate Incentives	0.029	0.083*	0.172**	0.058	-0.018	0.130***
	(0.076)	(0.044)	(0.072)	(0.043)	(0.052)	(0.049)
Observations	1,148	3,131	772	3,528	1,775	2,504
Mean of dep. var.	3.384	3.545	3.528	3.496	3.552	3.467
Std. dev.	0.851	0.724	0.705	0.775	0.719	0.791
Number of individuals	229	600	160	673	352	477
			Panel C: Course grades, lectures			
Immediate Incentives	0.028	0.101**	0.197**	0.070	0.002	0.141**
	(0.085)	(0.051)	(0.086)	(0.048)	(0.059)	(0.056)
Observations	939	2,455	615	2,798	1,401	1,993
Mean of dep. var.	3.330	3.478	3.465	3.430	3.497	3.395
Std. dev.	0.873	0.773	0.742	0.818	0.758	0.833
Number of individuals	227	596	160	667	348	475

Notes: The sample is restricted to individuals in the Immediate Incentives treatment and individuals in the Control group. All estimates include demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), baseline sleep, indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Observations are weighted by the number of credits taken in the semester. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

Table A.8: Incentives and other metrics of academic performance

	(1) Has a grade	(2) Withdrawn	(3) Failed	(4) Passed	(5) Credits
	Panel A: All classes				
Immediate Incentives	-0.011**	0.009*	-0.008	-0.003	0.025
	(0.005)	(0.005)	(0.005)	(0.008)	(0.031)
Observations	4,772	4,772	4,772	4,772	4,772
Mean of dep. var.	0.982	0.0142	0.00964	0.972	2.755
Std. dev.	0.133	0.119	0.0977	0.164	1.002
Number of individuals	840	840	840	840	840
	Panel B: Lectures				
Immediate Incentives	-0.014**	0.010*	-0.011*	-0.003	-0.009
	(0.006)	(0.005)	(0.006)	(0.009)	(0.033)
Observations	3,728	3,728	3,728	3,728	3,728
Mean of dep. var.	0.981	0.0161	0.0118	0.969	2.919
Std. dev.	0.138	0.126	0.108	0.174	0.874
Number of individuals	829	829	829	829	829

Notes: The sample is restricted to individuals in the Immediate Incentives treatment and individuals in the Control group. All estimates include demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), baseline sleep, indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Observations are weighted by the number of credits taken in the semester. Standard errors are clustered at the individual level. Mean of dep. var. is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A.9: Correcting for multiple hypothesis testing

Incentives and Sleep		
	p-value	q-value
Treatment:		
Sleep hours	0.00	0.00
Sleep $\geq$ 6 hours	0.00	0.00
7 $\leq$ sleep hours$\leq$ 9	0.00	0.00
Efficiency	0.07	0.04
REM sleep	0.00	0.00
Deep sleep	0.49	0.17
Bedtime	0.00	0.00
Wake up time	0.15	0.07
SRI	0.00	0.00
Sleep hours regularity	0.00	0.02
Bedtime regularity	0.02	0.02
Wake up time regularity	0.04	0.03
Post-Treatment:		
Sleep hours	0.00	0.00
Sleep $\geq$ 6 hours	0.00	0.00
7$\leq$ sleep hours$\leq$ 9	0.00	0.00
Efficiency	0.69	0.20
REM sleep	0.04	0.03
Deep sleep	0.85	0.20
Bedtime	0.46	0.16
Wake up time	0.07	0.04
SRI	0.12	0.07
Sleep hours regularity	0.13	0.07
Bedtime regularity	0.02	0.02
Wake-up time regularity	0.01	0.01
Incentives and Academic Achievement		
Term+1 GPA	0.07	0.18
Term+2 GPA	0.92	0.53
Has a grade	0.03	0.18
Withdrawn	0.07	0.18
Failed	0.10	0.18
Passed	0.72	0.44
Credits completed	0.41	0.26

Notes: The sample is restricted to individuals in the Immediate Incentives treatment and individuals in the Control group. All estimates include day of the week, week of the experiment, wave, and month fixed effects, baseline value of the outcome variable, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable is missing. Standard errors are clustered at the individual level.

Table A.10: Incentives and time use (in minutes), excluding careless respondents

	(1) Sleep	(2) Sleep $\geq$ 7 hours	(3) Study	(4) Social	(5) Work	(6) Eating & Preparing Food	(7) Exercise
Treatment	6.168	0.070***	0.951	-4.380	-6.219	0.534	1.268
	(4.612)	(0.017)	(7.163)	(4.855)	(5.351)	(2.242)	(1.783)
Post-Treatment	13.730**	0.061***	6.846	-3.790	-3.506	-5.311*	-3.656
	(5.905)	(0.023)	(9.689)	(6.807)	(7.129)	(2.713)	(2.529)
Observations	5,754	5,754	5,754	5,754	5,754	5,754	5,754
Mean of Dep. Var.	494.1	0.734	321.9	101.3	92.48	95.14	22.68
Std. dev.	108.8	0.442	192.2	123.3	144.2	51.17	42.87
Number of individuals	836	836	836	836	836	836	836
	(8) House errands	(9) Personal care	(10) Screen	(11) TV & other videos	(12) Internet	(13) Games	(14) Other
Treatment	-0.685	0.874	-11.453**	-8.262**	-2.718	-0.003	-0.552
	(1.364)	(1.635)	(4.989)	(3.238)	(3.166)	(2.171)	(7.302)
Post-Treatment	-1.576	-1.204	-3.300	-4.442	0.634	0.933	-11.151
	(1.789)	(2.321)	(6.784)	(4.435)	(4.141)	(3.297)	(9.555)
Observations	5,754	5,754	5,754	5,754	5,754	5,754	5,754
Mean of Dep. Var.	18.14	54.09	171.8	70.75	78.96	22.09	454.7
Std. dev.	36.85	40.27	136.7	92.15	90.09	63.62	171.5

Notes: The sample is restricted to individuals in any of the Incentived treatments (Immediate Incentives, Delayed Incentives, and Delayed Incentives, No Cue/Feedback) and individuals in the Control group. Individuals in the Cue/Feedback treatment were not included in this analysis. We also exclude participants deemed careless (e.g. those who gave the reported "other activities" for all time periods within the last 24 hours) All the estimates include controls for wave, month and day of the week fixed effects, indicators for the number of classes starting before 10 am, gender, race (dummies for Asian, Black, Hispanic, other) and ethnicity, parental education (dummies for less than college, college degree, and post-college degree), number of classes starting before 10 a.m., quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing), and the average time spent on the activity at baseline. Standard errors are clustered at the individual level. Mean of dep. var is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

Table A.11: Immediate Incentives, additional measures

	(1) Correct math answer	(2) Creativity score	(3) RHR	(4) # steps	(5) Active minutes	(6) #caffeinated drinks	(7) Any caffeinated drink
Treatment	0.002	0.002	-0.244	28.775	-1.297	-0.004	-0.011
	(0.023)	(0.057)	(0.186)	(154.303)	(3.885)	(0.048)	(0.027)
Post-Treatment	-0.031	0.000	-0.162	-84.946	-8.044	-0.014	0.018
	(0.032)	(0.059)	(0.204)	(232.343)	(6.019)	(0.053)	(0.031)
Observations	3,181	3,243	38,640	39,035	39,035	3,382	3,382
R-squared	0.188	0.052	0.837	0.356	0.293	0.458	0.398
Mean of Dep. Var.	0.363	3.307	65.72	8415	228.4	0.790	0.513
Std. dev.	0.481	0.717	8.241	5179	120.3	0.949	0.500
Number of individuals	809	803	832	840	840	816	816

Notes: The dependent variable in column 1 is an indicator equal to 1 if the respondent answered correctly the math question on the survey. The dependent variable in column 2 is a creativity score (see Section 2.4). RHR corresponds to participants Resting Heart Rate. Steps corresponds to participants' daily steps as measured via the Fitbit. Active minutes capture any activity at or above about 3 metabolic equivalents (METs). The sample is restricted to individuals in the Immediate Incentives treatments and individuals in the Control group. All estimates include day of the week, week of the experiment, wave, and month fixed effects, baseline value of the outcome variable, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Standard errors are clustered at the individual level. Mean of dep. var is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

Table A.12: Incentives and GPA, by SES, (All classes)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Waves	All Waves	All Waves	All Waves	All Waves	All Waves	All Waves	Waves 2-7	Waves 2-7
Subsample	Financial Aid	Non-Financial Aid	First-generation college	Non first-generation college	Work	Non-Work	Excludes from treatment those reporting food, drink, school expenses	Includes in treatment only those reporting food, drink, school expenses
Immediate incentives	-0.006	0.172***	0.020	0.103**	0.057	0.072*	0.080*	-0.022
	(0.047)	(0.062)	(0.063)	(0.044)	(0.073)	(0.043)	(0.043)	(0.078)
Observations	2,298	2,002	1,208	3,092	889	3,411	3,130	2,044
Mean of Dep. Var.	3.480	3.527	3.422	3.533	3.527	3.495	3.519	3.494
Std. dev.	0.772	0.753	0.807	0.743	0.727	0.773	0.727	0.755
Number of individuals	443	390	228	605	172	661	601	398

Notes: All estimates include demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), baseline sleep, indicators for the number of classes starting at 10am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable was missing. Observations are weighted by the number of credits taken in the semester. Standard errors are clustered at the individual level. Mean of dep. var is the mean of the dependent variable at baseline. Std. dev. is the standard deviation of the dependent variable at baseline. *** p<0.01, ** p<0.05, * p<0.1.

Table A.13: Correcting for multiple hypothesis testing, secondary treatments

	Sleep$\geq$ 7		Sleep hours	
	p-value	q-value	p-value	q-value
Treatment:				
Immediate incentives vs control	0.00	0.00	0.00	0.00
Delayed incentives vs control	0.00	0.00	0.00	0.01
Delayed incentives, no feedback vs control	0.00	0.00	0.00	0.01
Cue/Feedback only vs control	0.13	0.13	0.03	0.07
Immediate vs delayed incentives	0.09	0.12	0.05	0.08
Immediate vs delayed incentives no feedback	0.05	0.08	0.05	0.08
Immediate vs only feedback	0.00	0.00	0.00	0.01
Post-Treatment:				
Immediate incentives vs control	0.17	0.15	0.01	0.01
Delayed incentives vs control	0.89	0.40	0.54	0.28
Delayed incentives, no feedback vs control	0.07	0.10	0.22	0.16
Cue/Feedback only vs control	0.77	0.35	0.33	0.21
Immediate vs delayed incentives	0.27	0.19	0.06	0.09
Immediate vs delayed incentives no feedback	0.57	0.28	0.21	0.16
Immediate vs only feedback	0.35	0.22	0.12	0.13

Notes: The sample is restricted to waves 1 - 3. All estimates include day of the week, week of the experiment, wave, and month fixed effects, baseline value of the outcome variable, and demographic controls for gender, age (dummies), race and ethnicity (Asian, Black, Hispanic, White, other), indicators for the number of classes starting at 10 am or earlier, indicators for whether parents' highest academic title was less than college, college degree, more than a college degree, and quartile of baseline GPA (high school GPA if non-missing, prior term GPA if high school GPA is missing). For all demographic characteristics, we included a missing indicator for whether the variable is missing. Standard errors are clustered at the individual level.

Table A.14: Review of post-secondary interventions

Paper	Treatment	Setting	Findings	GPA & Costs
Angrist et al. (2009)*	A) Financial incentives for academic achievement B) Peer advising and study groups C) Treatments A and B combined	Field experiment with first-year students at Canadian 4-year university	A) GPA: -0.04 (0.061) B) GPA: 0.011 (0.063) C) GPA: 0.168 (0.086) Academic probation: -0.069 (0.036)	GPA: Table 6, Panel A, Column 1 Costs: Bottom of page 160
Angrist et al. (2014)*	Financial incentives for academic achievement	Field experiment with students at public university in Ontario	GPA: 0.009 (0.044)	GPA: Table 4b, "Fall" Panel, Column 9 Costs: Table 3, "Fall" Panel, Column 9
Barrow et al. (2014)*	Extra grant aid and counseling services as part of the Opening Doors Louisiana Program	Field experiment with low-income community college students in Louisiana	GPA: 0.182 (0.085) Credits: 1.234 (0.30)	GPA: Table 8, Column 2 Costs: Table 2, "First semester" panel, Column 1
Clotfelter et al. (2018)	Extra state grant aid due to crossing income threshold for Carolina Covenant Grant eligibility	Regression discontinuity with low-income students attending the University of North Carolina, Chapel Hill	GPA: 0.043 (0.053) 4-year degree[1]: 0.068 (0.040)	GPA: Table 6, Panel A, Column 1 Costs: Table 3, Panel B, Column 1
Denning et al. (2019)	Extra Pell and state grant aid due to crossing threshold for $0 Expected Family Contribution	Regression discontinuity with 4-year university and community college students in Texas	GPA, FTIC[1]: 0.031 (0.026) 4-year degree, FTIC: 0.022 (0.012) GPA, returning students: 0.014 (0.013)	GPA: Table 3, Panel B, Column 3 Costs: Table 2, Column 2

[1] FTIC stands for "First Time in College"

Notes: The asterisk (*) symbol signifies that the study is included in Lintner (2024). Column 4 reports average treatment effects with standard errors in parentheses. We report OLS estimates of effects on non-cumulative GPA, either at the semester or year-level. We report multiple GPA effects when authors reported on multiple treatment arms (e.g. Angrist et al. (2009); Evans et al. (2020)) or cohorts (e.g. Goldrick-Rab et al. (2016); Denning et al. (2019)). We also report other statistically significant effects, such as credits completed or degree completion, when applicable. Effect sizes on "4-Year Degree" report the rate at which people receive a 4-year degree in 4 years, while "Credits" reports effect sizes on credits taken in one school year. When multiple GPA effects were reported, we selected semester-level estimates if available and used the authors' preferred specification if indicated. When per-person treatment costs were not reported, we divided overall program costs per semester by the intent-to-treat sample size (if total program costs were reported by year, we divided in half to calculate per semester costs).

Table A.14: Review of post-secondary interventions (continued)

Paper	Treatment	Setting	Findings	GPA & Costs
De Paola et al. (2012)*	Financial incentives for academic achievement	Field experiment with business administration students at the University of Calabria	Cumulative points: 6.023 (3.06) Credits: 2.34 (1.22)	GPA: NA[2] Costs: Page 64
Erwin et al. (2021)*	Financial incentives and advising for academic achievement	Field experiment with students at the University of New Mexico	GPA: 0.02 (0.06) Credits: 0.03 (0.02)	GPA: From Lintner (2024) Costs: Page 148[3]
Evans et al. (2020)	A) Access to emergency grant funding B) Treatment A as well as advising services	Field experiment with low-income community college students in Texas	A) GPA: -0.134 (0.083) B) GPA: 0.055 (0.07) Enrollment, female students: 0.04 (0.041)	GPA: Treatment A provided by authors, Treatment B from Table 8, Column 2 Costs: Pages 958-959
Goldrick-Rab et al. (2016)	Extra grant aid as part of the Wisconsin Scholars Grant	Field experiment with low-income first-year students at public universities in Wisconsin	GPA, cohort 1: 0.08 (0.06) Credits, cohort 1: 0.9 (1.7) GPA, cohorts 2 & 3: 0.09 (0.03) Credits, cohorts 2 & 3: 2.1 (0.7)	GPA: Table 5, “First Semester" Panel, Columns 2 & 5 Costs: Bottom of page 1772
MacDonald et al. (2009)*	Financial incentives and advising for academic achievement as part of the Foundations for Success program	Field experiment with students at three Canadian colleges	GPA: 0.10 (0.06)	GPA: From Lintner (2024) Costs: Page 87, Table 9-5[4]

[2] We includeDe Paola et al. (2012) in the cost calculations for the Lintner (2024) meta-analysis, but do not include the study in Figure 5 because it does not report a GPA measure.

[3] We use the authors' per-person cost estimate for scholarships alone (excluding advising costs), as the costs of advising are not given explicitly. These estimates are therefore a lower-bound estimate.

[4] We use “Year 2” program costs for our cost estimate. The authors state that these are more representative of program costs because they do not include start-up costs.

Table A.14: Review of post-secondary interventions (continued)

Paper	Treatment	Setting	Findings	GPA & Costs
Oreopoulos and Petronijevic (2018)	A) Online exercise encouraging future-oriented thinking B) Treatment A as well as study advice and motivation via text messages C) Treatment A as well as one-on-one peer support	Field experiment with students at three campuses of the University of Toronto	A) Course grades: 0.143 (0.575) B) Course grades: 0.073 (0.505) C) Course grades: 4.897 (1.874) Credits: 0.501 (0.283)	Course grades: Table 3, Column 5[5] Costs: Bottom of page 323[6]
Park and Scott-Clayton (2018)	Extra Pell grant aid due to crossing threshold for $0 Expected Family Contribution	Regression discontinuity with community college students from 20+ institutions in a single state	GPA: 0.064 (0.082) Enrollment: 0.094 (0.034)	GPA: Table 5, Column 2 Costs: Table 5, Column 2
Patel and Valenzuela (2013)*	Financial incentives and advising for academic achievement as part of the Adalente Performance Award	Field experiment with Latino male students in Tucson, Arizona at Pima Community College	GPA: 0.01 (0.07) Credits: 0.07 (0.03)	GPA: From Lintner (2024) Costs: Page 47, Table 3.2
Richburg-Hayes et al. (2015)*	Financial incentives for academic achievement as part of Cash for College program	Field experiment with high school seniors in California	GPA: 0.02 (0.05) Credits: 0.04 (0.01)	GPA: From Lintner (2024) Costs: Page 98, Figure 5.1
Scott-Clayton (2011)	Free tuition as part of the West Virginia PROMISE program	Regression discontinuity with public university students in West Virginia	GPA: 0.066 (0.066) Credits: 1.572 (0.085) 4-year degree: 0.058 (0.004)	GPA: Table 3, Column 3 Costs: Middle of page 617
Sommo et al. (2014)*	Financial incentives and math tutoring for academic acheivement as part of MAPS program	Field experiment with students in Tampa, Florida at Hillsborough Community College	GPA: 0.04 (0.07) Credits: 0.04 (0.02)	GPA: From Lintner (2024) Costs: Page 72, Table 5.1

[5] Authors present course grades on a 0-100 scale. In figure 5, course grades have been divided by 25 for comparison with 4.0 GPA scale.
[6] Only treatment arm C is included in figure 5 because costs could not be calculated for A and B.

B. Instructions and Experimental Material

Immediate Incentives

PLEASE READ THROUGH THIS MESSAGE ENTIRELY
Starting this Sunday, and every weeknight (Sunday-Thursday) for the next four weeks, we encourage you to get 7 hours of sleep or more by 9 am the following morning.
Every time you meet this goal (i.e., sleep 7 hours by 9 am), you will earn a $4.75 PAYMENT via Venmo. Payments are redeemable only until 3 pm on the days you earn them, and you will receive the payment by 3 pm if you have redeemed by that time.

HOW IT WORKS
Every morning, you will receive feedback on your sleep. If you meet your goal, you will also receive the payment information via text message.

Next, we would like to ask you to pick your bedtime behavior – a behavior you would like to engage on right before going to sleep. Every weeknight, we will remind you of your bedtime behavior and we will encourage you to go to sleep early enough to meet your goal of sleeping at least 7 hours by 9 am. Please pick your bedtime behavior by texting back the number of your choice. If you choose other, please type 9, then the behavior you want to set as your bedtime behavior.

1. Turn off your phone
2. Turn your phone to silent
3. Turn off your computer
4. Turn off Netflix
5. Turn on bedtime music
6. Turn on meditation app
7. Turn on white noise
8. Turn on pink noise
9. Other

Delayed Incentives

PLEASE READ THROUGH THIS MESSAGE ENTIRELY
Starting this Sunday, and every weeknight (Sunday-Thursday) for the next four weeks, we encourage you to get 7 hours of sleep or more by 9 am the following morning.
Every time you meet this goal (i.e., sleep 7 hours by 9 am), you will earn a $4.75 PAYMENT via Venmo. Payments are redeemable only until 3 pm on the days you earn them, and the payment will be added to the amount of money you receive at THE END OF THE STUDY.

HOW IT WORKS
Every morning, you will receive feedback on your sleep. If you meet your goal, you will also receive the payment information via text message.

Next, we would like to ask you to pick your bedtime behavior – a behavior you would like to engage on right before going to sleep. Every weeknight, we will remind you of your bedtime behavior and we will encourage you to go to sleep early enough to meet your goal of sleeping at least 7 hours by 9 am. Please pick your bedtime behavior by texting back the number of your choice. If you choose other, please type 9, then the behavior you want to set as your bedtime behavior.

1. Turn off your phone
2. Turn your phone to silent
3. Turn off your computer
4. Turn off Netflix
5. Turn on bedtime music
6. Turn on meditation app
7. Turn on white noise
8. Turn on pink noise
9. Other

Cue / Feedback

PLEASE READ THROUGH THIS MESSAGE ENTIRELY
Starting this Sunday, and every weeknight (Sunday-Thursday) for the next four weeks, we encourage you to get 7 hours of sleep or more by 9 am the following morning.

HOW IT WORKS
Every morning, you will receive feedback on whether you met your goal.

Next, we would like to ask you to pick your bedtime behavior – a behavior you would like to engage on right before going to sleep. Every weeknight, we will remind you of your bedtime behavior and we will encourage you to go to sleep early enough to meet your goal of sleeping at least 7 hours by 9 am. Please pick your bedtime behavior by texting back the number of your choice. If you choose other, please type 9, then the behavior you want to set as your bedtime behavior.

1. Turn off your phone
2. Turn your phone to silent
3. Turn off your computer
4. Turn off Netflix
5. Turn on bedtime music
6. Turn on meditation app
7. Turn on white noise
8. Turn on pink noise
9. Other

Creativity Instructions (Example)

You will be asked to complete different short tasks over the course of the study. One of these tasks will be chosen for payment at the end of the study.

Today's task: Using some or all of the words below, write an interesting sentence. Your sentence will be rated based on its creativity from 1-5 points, where 5 is the most creative. If today's task is chosen for payment, your payment will be determined by how creative your sentence is. **You will receive $1 for each point your story is rated**.You will receive as little as $1 for completing this activity and up to $5 for the most creative sentences. You will receive your rating and your payment at the end of the study.

The words for you to use in your sentence are:
(Example) event, chocolate, system, indicate, article, emotion, possess, mom, poetry, reality

Math Instructions (Example)

You will be asked to complete different short tasks over the course of the study. One of these tasks will be chosen for payment at the end of the study.

Today's task: On the next page you will be asked to answer a math question. If today's task is chosen for payment, your payment will be determined by whether you answer the question correctly, and how quickly you answer. **You will receive $1 for answering the question correctly, and you will receive an additional $0-$4 depending on how quickly you answer the question**. You will receive as little as $1 for answering this question correctly and up to $5 for the quickest correct answers. You will receive your score and the payment at the end of the study.

Here is the question you are asked to answer:

It costs a manufacturer X dollars per component to make the first 1,000 components. All subsequent components cost $1 each. When X = $1.50 How much will it cost to manufacture 4,000 components?

o $3,500
o $3,000
o $4,000
o $3,250
o $4,500

App ScreenShots

Figure B.1: Bedtime Reminder

Notes: The Bedtime reminder included a personalized goal bedtime of approximately 1 hour before the baseline bedtime, with a latest possible time of 1 am. It also included a personalized bedtime behavior participants chose from before the beginning of the intervention, from a list containing "Turn off your phone", "Turn your phone to silent", "Turn off your computer", "Turn off Netflix", "Turn on bedtime music", "Turn on meditation app", "Turn on white noise", "Turn on pink noise", "Other". If participants selected "Other" they could specify a behavior of their choice.

Figure B.2: App Screenshots - Immediate Incentive Treatment

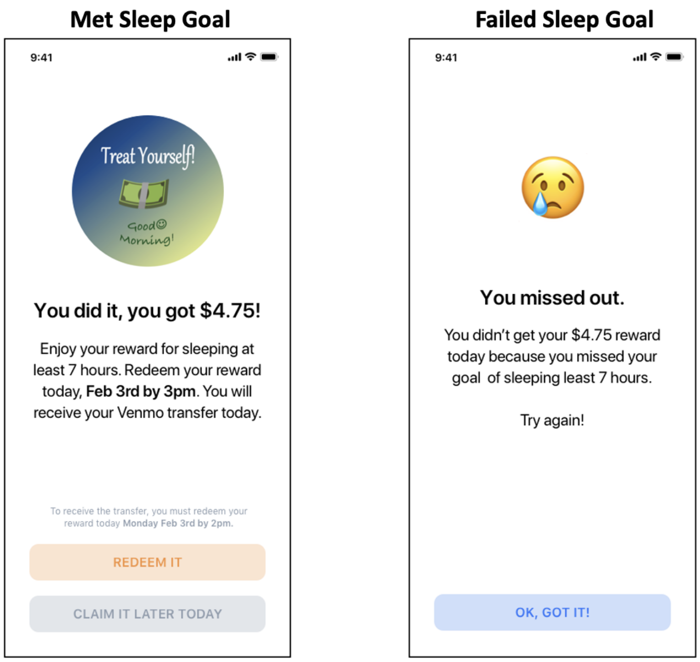


Figure B.3: App Screenshots - Delayed Incentive Treatment

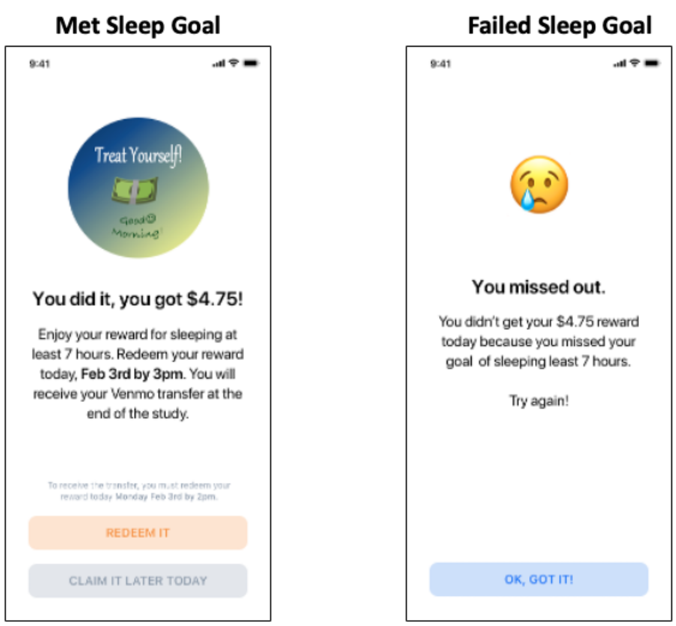


Figure B.4: App Screenshots - Cue/Feedback Treatment

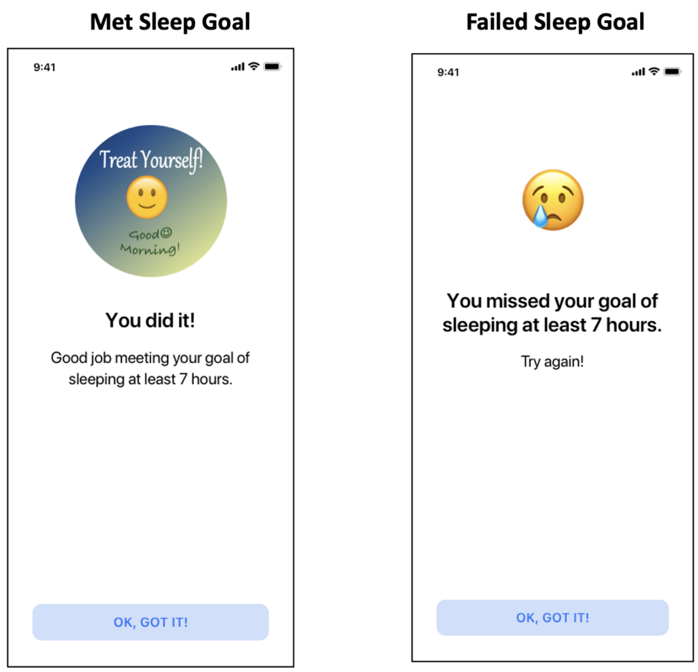


Figure B.5: Reminder to Sync - All Treatments

Appendix C: Pre-registration

We pre-registered the experiment on the AEA RCT registry (AEARCTR-0003235). The initial registration date was August 14, 2018, and the pre-analysis plan document was uploaded on March 1st, 2019. Data collection began in the Spring semester of 2019 and continued every semester until Spring 2022. Below, we describe the different sections of the pre-registration and note any deviation.

Experimental Design. In our pre-registered experimental design, we originally planned to have five, rather than four, treatments. In addition to the Financial Incentives with Reminders/Feedback, the Delayed Non-Financial Incentives with Reminders/Feedback, the Delayed Non-Financial Incentives with no Reminders/Feedback, and the Reminders/Feedback treatments (corresponding to Immediate Incentives, Delayed Incentives, Delayed Incentives with no cue/feedback, and Cue/Feedback treatments respectively), our original plan included a treatment involving Non-Financial Incentives with Reminders/Feedback.

This non-financial incentive treatment consisted of providing participants with a \$4.75 coupon for a breakfast treat at one of the University of Pittsburgh Einstein Coffee locations, delivered through the smartphone app we had developed for the other treatments. However, due to unforeseen logistical difficulties, we suspended this treatment after the first wave and excluded it from the following waves. Specifically, although personnel at all Einstein Coffee locations were instructed to accept the study coupon, high turnover among baristas and information loss led to frequent rejections of the coupons during the first wave. Consequently, we made the decision to discontinue this treatment, for which we had originally planned to collect 600 observations.

In addition, based on feedback we received during preliminary presentations of the project, in waves 5 and 7 we ran a variation of the Financial Incentive with Reminders/Feedback treatment (Immediate Incentives) that stopped the reminders and feedback in the post-treatment period. This was done to test the importance of continuing to provide reminders and feedback in the post-treatment phase. This variation of the main treatment was not pre-registered, and we treat the analysis of the impact of cues in the post-treatment period as exploratory.

Finally, the pre-registration explains that we would collect baseline data for 1-3 weeks, run our treatments for 6 weeks, and then continue to track participants. In practice, however, we immediately realized that to have any post-treatment data by the end of the semester, we had to reduce the treatment period to 4 weeks. This adjustment was necessary because the lab did not always allow us to recruit participants right when the semester began.

Key dependent variables. Table C.1 in this section describes the pre-registered primary and secondary outcomes, the additional exploratory measures we collected, and notes any deviations from the pre-registration.

Analysis - Primary Analyses. Following the pre-registration, we conducted intent-to-treat (ITT) regression analysis to measure the impact of the interventions on sleep, academic, and health outcomes. However, for academic and health outcomes, we also planned to conduct an instrumental variable (IV) regression, instrumenting for improved sleep to measure the impact of the interventions on these outcomes and the secondary outcomes. We did not conduct the IV analysis because our intervention may affect GPA and health through channels other than sleep–such as time allocation to other activities–potentially violating the IV exclusion restriction.

As pre-registered, we estimate the impact of Immediate Incentives both individually (reported in the main text) and pooled with Delayed Incentives (reported in Appendix Table A.3), column 7 for sleep, and Appendix Table A.6, column 8 for GPA). We also conduct analysis that includes all treatments once enrollment for the secondary treatments has ended (Table 8) and once enrollment for Control and Immediate Incentives has ended (Table 3). We correct for multiple hypothesis testing within families of secondary outcomes, summarizing the families in Table C.2 in this section. Additionally, we examine sensitivity to non-compliance (Table A.3, column 5).

We pre-registered that we would examine mechanisms other than improved sleep that may be driving the impact on outcomes. To that end, we focus on time use, cognitive performance, and health (Section 3.3). We do not examine the response to daylight saving in the current paper, but we will address this in a separate paper.

Analysis - Secondary Analyses. Our pre-registered secondary analysis examines 1) whether treatment effects are larger for people who are sleeping insufficiently at baseline (Table A.4), and 2) whether treatment effects vary by class time (Table 5, columns 2 - 4). In exploratory analyses, we examine the impact by class type (STEM vs. non-STEM, Table 5, columns 5 and 6), which we did not pre-register.

Sample Size. Our pre-registered sample size was N=2100, with N=600 for the control group, Financial Incentives with Reminders/Feedback, and Non-Financial Incentives with Reminders/Feedback treatments, and approximately N=100 for the remaining treatments (Delayed Incentives with Reminders/Feedback, Delayed Incentives with no Reminders/Feedback, and Reminders/Feedback). We intended to run the study across multiple semesters to reach the full sample size, assigning participants approximately equally across all treatments until the secondary treatment sample size was met. Subsequently, we planned to randomize participants across the three remaining treatments.

As discussed above, we discontinued the Non-Financial Incentive treatment in the first wave. In addition, we deviated from the targeted sample size for Control and Financial Incentives with Reminders/Feedback due to the COVID-19 pandemic. Specifically, by wave 3 (Spring 2020), we had achieved a sample size of approximately N=100 for the secondary treatments. At this point, following university protocols, we shifted our experiment from the lab to Zoom to avoid contact and mailed Fitbits to participants' homes. Participants were then instructed to mail them back to us at the end of the study. This change substantially increased our study expenses, preventing us from reaching our target of N=600 participants in each of the Control and Treatment groups before our funding ended, leaving us with N=380 and N=468, respectively.

Table C.1: Outcomes and families

Secondary outcome variables	Family of Outcomes for MHT	Secondary outcome variables	Family of Outcomes for MHT
Sleep hours	Sleep	Time use: Study	Time use
Bedtime	Sleep	Time use: Social	Time use
Wake-up time	Sleep	Time use: Work	Time use
Regularity: Daily SRI	Sleep	Time use: Eating & Prep Food	Time use
Regularity: Sleep Hours	Sleep	Time use: Exercise	Time use
Regularity: Bedtime	Sleep	Time use: House errands	Time use
Regularity: Wake-up time	Sleep	Time use: Personal Care	Time use
Sleep 7-9 hours	Sleep	Time use: Screen- combined	Time use
Quality: Efficiency	Sleep	Time use: Screen- TV & other videos	Time use
Quality: REM	Sleep	Time use: Screen- Internet	Time use
Quality: Deep Sleep	Sleep	Time use: Screen- Games	Time use
Sleep $\geq$ 6 hours	Sleep	Time use: Other	Time use
		Cognitive Performance - Math	Cog Performance
Term +1 GPA	Education	Cognitive Performance - Creativity	Cog Performance
Term +2 GPA	Education	Health: Resting Heart Rate	Health
Grade (Y/N)	Education	Health: Activity- Steps	Health
Withdrawn	Education	Health: Activity - Active Minutes	Health
Failed	Education	Mental health- CESD	Health
Passed	Education	Mental health- GAD-7	Health
Credit Completed	Education	Mental health - Mood	Health
Time use: Sleep	Time use	Mental Health - Resilience	Health
Time use: Sleep $\geq$7	Time use	Mental Health - Stress	Health

Table C.2: Familily of Outcomes

Primary (pre-reg)	**Secondary (pre-reg)**
Panel A: Sleep	
Sleep ≥ 7 during and after the intervention, measured by Fitbit	Sleeping at least 7 hours on weeknights and weekends
	Sleeping 7-9 hours on weeknights and weekends
	Hours per night
	Sleep including naps
	Bedtime
	Wakeup time
	Regularity
	Sleep quality
Panel B: Academic Achievement	
GPA (administrative record)	GPA after the study
	Corse completion
	Grades conditional on completion
	Credits
	Major
	All the measures were obtained using administrative records
Panel C: Health	
Activity, measured via Fitbit	BMI
Heart Rate, measured via Fitbit	Blood pressure
Mental Health, measured by CES-D, GAD-7	Survey measures of health behavior
Panel D: Additional measures	
	Time use
	Cognitive function (surveys & cognitive tasks)
	Actual behavior compared to predicted & desired behavior
	Willingness to pay (WTP) for incentives to sleep

Notes:
Panel A: For regularity, we consider: 1) The daily sleep regularity index (SRI), the day-by-day similarity in sleep and wake patterns; and 2) the within-individual standard deviation of total sleep hour, bedtime, and wakeup time at the week level. For sleep quality, we consider sleep efficiency, REM sleep and deep sleep. In the paper we also report exploratory analysis of sleep ≥ 6 hours.
Panel B: For GPA after the study, we look at it up to 2 terms after study conclusion. We could not obtain administrative data on major, attainment, and academic behavior, though we have self-reported information on major.
Panel C: After running a few participants in wave 1, we discontinued data collection for BMI and blood pressure, as enrollment took too long. We also collected exploratory measures of mood, stress resilience measured weekly via text-messaging.
Panel D:In the current paper, we do not report the survey measures of health behavior, the actual vs. predicted behavior & the WTP measure. We will report these measure in a separate paper. For cognitive function, we measured performance in math questions from the GRE and creativity following Charness and Grieco (2019).