Sleep: Educational Impact and Habit Formation*

Osea Giuntella
University of Pittsburgh
IZA and NBER

Silvia Saccardo
Carnegie Mellon University

Sally Sadoff
UC San Diego

December 5, 2023
(Click Here for the Latest Version)

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 automatic habit formation, our primary intervention couples personalized bedtime reminders to go to sleep with morning feedback and immediate rewards for sleeping at least seven hours on weeknights. The intervention increases the share of nights with at least 7 hours of sleep by 28 percent and average weeknight sleep by an estimated 19 minutes during a four-week treatment period, with persistent effects of about 9 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: J10, I10, I23

Keywords: Sleep, Human Capital, Education, Habit Formation, Health Behaviors

*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 Universitat in Berlin, UCSD Spring School on Behavioral Economics, University of Chicago, Harvard University, RAND, University of Southern California, the UCLA Anderson School of Management, and the New York Federal Reserve. We are grateful to Mallory Avery, Daniel Banko-Ferran, Kelly Hyde, 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 benefited from comments and discussion with Michele Belot, Alison Buttenheim, Gretchen Chapman, David Dickinson, Kareem Haggag David Huffman, George Loewenstein, Frank Schilbach, Heather Schofield, Peter Schwardmann, Kevin Volpp. 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). A large body of work shows that sleep affects cognitive function, productivity, earnings and academic performance, as well as, physical and mental health (see e.g., Lindquist and Sadoff, 2023, for a review). At the same time, statistics suggest that 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). Roenneberg (2013) referred to sleep deprivation as the most prevalent high-risk behavior in modern societies. Yet, little is know about whether interventions targeting sleep can improve productivity and performance.

In recent work, Bessone et al. (2021) implemented a randomized field experiment in India that increased sleep through sleep aids and incentives. They find that their intervention increases nighttime sleep by an average of 27 minutes over a twenty day treatment period. But there is no meaningful impact on cognition or productivity. The intervention could even decrease earnings because the increased time spent in bed decreased time spent working. That increased sleep did not improve labor market performance could be due in part to the study's context: participants were severely sleep deprived at baseline and the environmental conditions led to poor sleep quality. Consistent with this, they find evidence that high quality sleep via office naps can increase productivity. The study cannot disentangle whether the differential effects of naps compared to nighttime sleep are due to the differences in the timing of sleep or the quality. Further, it is possible that to observe impact of sleep on cognition or productivity requires increases in sleep that are sustained over a longer time horizon. It therefore remains an open question whether sustained exogenous increases in nighttime sleep can improve productivity and performance in the U.S., where average sleep is of high quality and closer to recommended guidelines.

Our study implements a randomized field experiment among U.S. university students to examine the impact that interventions targeting sleep have on sleep habits and academic performance. We ran the experiment over seven semester-long waves from Spring 2019 to Spring 2022. The 1149 participants wore tracking devices (Fitbits) that measure 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, cognitive performance 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.

Our primary intervention aims to build persistent sleep habits. To do so, we build on theories of cue-based habit formation, which highlight the role of context cues, repetition, and rewards in habit development (Verplanken and Wood, 2006; Wood and Neal, 2007; Wood and Rünger, 2016). These theories suggest that repeatedly taking a desired action in response to a consistent cue can gradually create an association between the cue and the reward that follows 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). Following this framework, our intervention repeatedly paired daily cues to go to sleep with immediate rewards for meeting sleep goals throughout the four-week intervention period.

Specifically, in our treatment groups, we set a goal for participants to sleep at least 7 hours by 9 am on weeknights (Sunday - Thursday), following recommended guidelines (Panel et al., 2015). During the intervention period, participants in our primary treatment received personalized bedtime reminders every weeknight, prompting them to follow a self-selected bedtime routine and get at least 7 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. In the post-intervention period, we stopped the financial reward but continued to cue participants to sleep via bedtime reminders and provide morning feedback to reinforce the behavior. This allows us to investigate the persistence of behavior change in response to the cue once the reward is removed. In secondary treatments, we tested variants that provided rewards with a delay rather than immediately; and, that turned off either the rewards or the cue and feedback.

Our primary analysis compares a 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 nights. During the treatment period, the intervention increases the rate of sleeping at least seven hours on weeknights by an estimated 12 percentage points ($p < 0.001$), 28 percent higher than baseline. The treatment effects persist into the post-intervention period but are smaller: an estimated 5.6 percentage points ($p < 0.001$), a 13 percent increase compared to baseline. We estimate that average weeknight sleep increases by 19 minutes during the treatment period and 9 minutes during the post-treatment period ($p < 0.001$).[2]

[1]The primary *Immediate Incentives* group pools two sub-treatments that received cues, feedback, and immediate incentives: one that continued to receive reminders and feedback in the post-treatment period and a secondary treatment group that did not receive reminders and feedback in the post-treatment period, which allows us to examine the importance of providing context cues for the persistence of behavior after the reward is removed. We do not find significant differences in the post-treatment effects of the two groups and pool them for our primary analysis.

[2]Our focus on weeknight sleep is in line with prior work that examines the impact of school and class

To understand the extent to which our results reflect sustained changes in sleep habits, we examine treatment effects on sleep behaviors. As discussed above, we designed our intervention to establish the habit of earlier bedtimes, triggered by the nighttime reminders. We find that the intervention initially leads to earlier bedtimes (and directionally earlier wake up times) but these behaviors do not persist. On average, bedtime returns to baseline levels and wake-up time becomes slightly later. While our intervention does not establish early bedtime habits on average, we find that within-individual bedtime and wake up time become more regular in both the treatment and post-treatment period. These results suggest that the intervention led participants to establish more stable routines independent of the external cue to go to bed earlier.

We further explore mechanisms of habit formation by comparing our primary intervention to three secondary treatments: (1) *Delayed Incentives*, which is identical to Immediate Incentives but with the payout delayed to the end of the study (about a month after the treatment period); (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 reminders and feedback with no rewards. During the treatment period, we find that Immediate Incentives significantly outperform the secondary treatments. The effects of Immediate Incentives are 60 to 85 percent higher than Delayed Incentives and three to four times larger than Cue/Feedback alone. During the post-treatment period, the estimated effects of Immediate Incentives are generally larger than the secondary treatment but are not statistically distinguishable. Our results suggest that combining cues with immediate rewards has large impacts while incentives are being offered but may not be particularly effective at enhancing the persistence of habits.

We then turn to the educational impact of incentives to sleep. Immediate Incentives increase semester course performance by an estimated 0.075 - 0.088 grade points ($p = 0.044$ and $p = 0.035$, respectively).[3] We find evidence of similar sized treatment effects on grade point average in the semester following the intervention, but no effect two semesters after the intervention. We examine heterogeneity by course type and time of day and find that treatment effects are largest in STEM courses and classes that take place midday.

We benchmark our effects in comparison to prior work linking sleep to academic performance. Estimates from natural experiments in the U.S. suggest that a one hour later shift in sunrise or class start time increases sleep by an average of 6 - 35 minutes and has either

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.

[3]Effects are similar but with slightly smaller magnitudes if we pool all incentive treatments (immediate and delayed).

no discernible impact on academic achievement or can increase grades and test scores by 0.06 - 0.16 standard deviations (SD) (Carrell et al., 2011; Heissel and Norris, 2018; Groen and Pabilonia, 2019). By comparison, our intervention increases weeknight sleep by an estimated 19 minutes during treatment and 9 minutes during post-treatment; and, grades by 0.10 - 0.11 SD.

We consider three primary channels through which sleep could affect academic performance: lifestyle, cognition, and well being. To that end, we examine the impact of our intervention on time use, performance in math and creativity tasks, and measures of physical and mental well-being. 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 impacts on our measures of cognitive performance: math and creativity. We also 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. 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.

Our study is the first to show that an intervention targeting sleep can improve academic performance. These findings contribute to the growing literature on the economics of sleep. Seminal work on sleep finds a negative association in time use surveys between sleep and work hours, but is not able to identify causality (Biddle and Hamermesh, 1990; Basner et al., 2007). Related work finds a positive correlation between sleep and health outcomes (Cappuccio et al., 2010). In a separate line of research, sleep lab studies find that short-term severe sleep deprivation worsens attention, memory, cognition and mood, but are not able to estimate the effect of moderate longer-term changes in sleep in natural contexts (Banks and Dinges, 2007; Killgore, 2010). Other studies find that sleep deprivation affects decision making, ethical behavior, and social decisions (Dickinson and McElroy, 2017; Dickinson and Masclet, 2023; McKenna et al., 2007).

More recently, studies using naturally occurring data find that later sunset times are correlated with lower cognitive performance and earnings, and worse physical and mental health, arguing that the channel is via reduced sleep (Giuntella et al., 2017; Gibson and Shrader, 2018; Giuntella and Mazzonna, 2019; Jin and Ziebarth, 2020). Related work finds that earlier class times are correlated with less sleep, lower academic performance at both the K - 12 and post-secondary levels (Carrell et al., 2011; Heissel and Norris, 2018; Jagnani, 2021; Groen and Pabilonia, 2019), and voting behavior (Holbein et al., 2019). In addition, recent research among U.S. university students finds a strong positive correlation between freshmen's academic performance and sleep, particularly in the first half of the term (Creswell et al., 2023). However, Lusher et al. (2019) find no evidence of significant effects of sleep regularity on academic outcomes in a large study at a Vietnamese university and little impact of delayed start times. None of these studies exogenously vary sleep or test policies to improve sleep habits.

Recent work using field experiments has tested the impact of interventions targeting sleep. In addition to Bessone et al. (2021) discussed above, two additional field experiments have examined interventions to improve sleep. Avery et al. (2022) find that incentives to early bedtime and sleep duration increased sleep and identify a demand for commitment devices to improve sleep habits. Barnes et al. (2017) test the effect of treating insomnia with internet-based cognitive behavior therapy and find beneficial effects on negative affect, job satisfaction, and self-control.

Our findings also contribute to the large literature on improving academic performance, particularly among college students. Our intervention is highly cost effective compared to previously examined policies, including financial aid, mentoring and support services, and performance-based incentives. Our results suggest that it may be more cost effective to improve academic achievement via sleep rather than to incentivize performance directly. This finding is akin to recent work showing that incentives for exercise can improve educational achievement (Cappelen et al., 2017).

Lastly, we contribute to the literature on habit formation by examining the impact of interventions that pair cues with feedback and rewards on short and longer term habits. Prior work has largely examined these separately. For example, Wellsjo (2021) focuses on the role of cues for generating automatic habits in the context of handwashing; and, Byrne et al. (2022) test the impact of repeated feedback on sustained behavior change in the context of water conservation. A large prior literature offers rewards for repeated engagement in desirable behaviors in the context of smoking, weight loss, exercise, and handwashing, always distributing rewards with a delay from the time of the incentivized behavior (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 during the intervention period, which adds to prior findings that immediate rewards outperform delayed rewards in the context of one-time behaviors (Levitt et al., 2016). Finally, our examination of habit formation suggests that combining cues with immediate incentives does not improve the persistence of habit in our setting and that individuals establish their own sleep routines rather than following prescribed bedtime cues. These results are in line with prior work showing that interventions that require set exercise routines are less effective than those that allow for individual flexibility (Beshears et al., 2021).

In the remainder of the paper, we describe the experimental design, data and analysis in Section 2. Section 3 presents the results and Section 4 concludes with a discussion of the cost-effectiveness of our intervention.

2 Experimental design and data

We conducted our experiment in seven semester long waves from Spring 2019 - 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 (for sleep and physical activity), survey measures (time use, cognitive performance 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 personal wearable devices (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 demonstrated the accuracy of these heart rate-enabled Fitbits compared to actigraphy, a commonly used method for outpatient sleep screening, suggesting their suitability for population-based sleep research (Haghayegh et al., 2019).

One source of concern with studies that rely on wearable trackers is that the devices require continued engagement via daily syncing. To ensure high sync levels for our experiment 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 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, and to deliver feedback and immediate rewards to participants upon meeting their sleep goals.

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 registry (RCT ID AEARCTR-0003235).

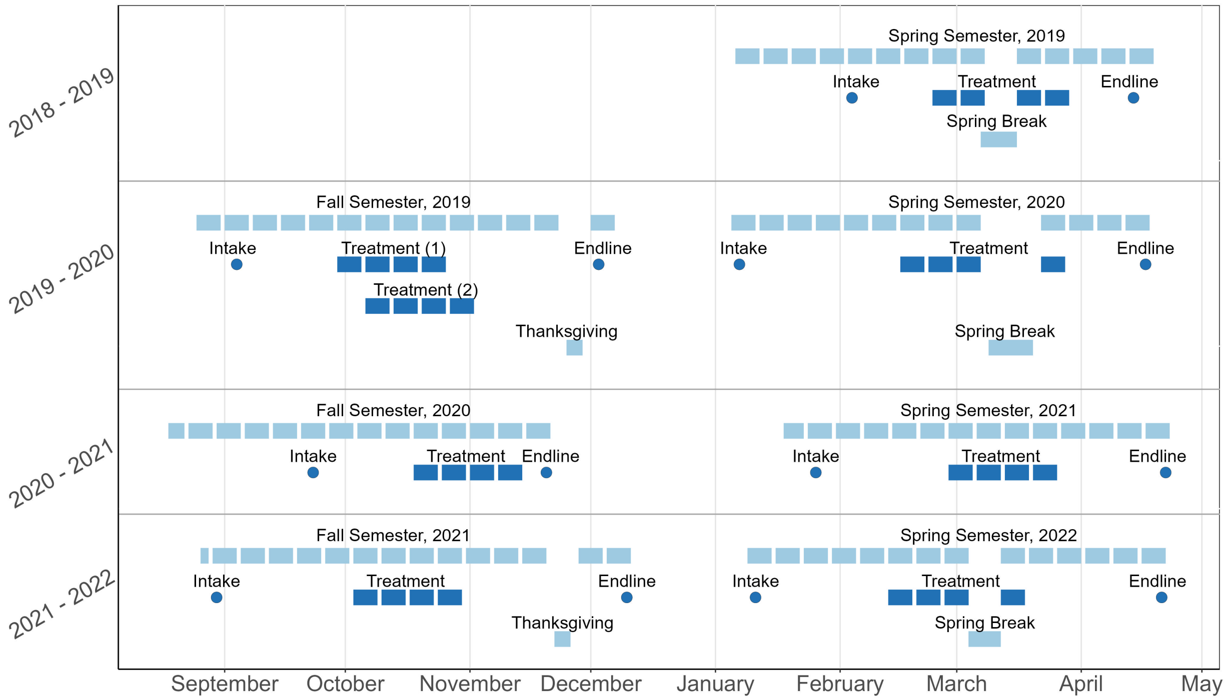
We recruited participants through the Pittsburgh Experimental Economics Laboratory (PEEL) and invited them to participate in a semester-long study on wellness for a guaranteed minimum payment of $30 and the opportunity to receive additional earnings based on luck as well as their choices during the experiment. 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-size 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 1149 participants.

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. 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 4 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 study always ended during the last week of classes, before the final exams. In the different waves, the start of the recruitment period depended on lab availability. The timing of the treatment period and the length of the post-treatment period varied in each wave depending on when we were able to start recruiting, how quickly we enrolled participants, and the semester schedule. The timeline of the experiment for each of the seven waves is depicted in Figure 1.[4]

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,

[4]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 1. In Spring 2020, the semester schedule was changed by the university closure prompted by the onset of the COVID-19 pandemic. Students enrolled in 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.

Figure 1: Timeline of the experiment



received the Fitbit and installed our custom smartphone app.[5] 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. To mitigate potential experimenter demand effects, we did not specifically disclose to participants that our interest was to study sleep behavior. Instead, we broadly explained that we were interested in wellness. 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 collected information on socio-demographic characteristics and baseline measures of physical and mental health.

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

[5]From Fall 2020 onwards, 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.

2.3 Treatments

In total, 1219 individuals completed an enrollment survey. In order to be randomized, participants had to have at least a day of Fitbit data in the baseline period. In total, we randomized 1149 participants to treatments.[6]

After the baseline period, we randomized participants into treatment groups, which are displayed in Table 1.[7] Participants in the *Control* group (n=373, waves 1-7), received no intervention and continued to wear their wearable trackers 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 7 hours per night by 9am on weeknights (Sunday through Thursday).

Table 1: Treatments

	N	**Waves**	**Treatment**		**Post-Treatment**
			Reminders & Feedback	Rewards	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	–
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.

Drawing on the habit formation framework outlined earlier, our *IImmediate Incentives* intervention (468 participants, waves 1-7) leverages cues, rewards, and repetition to establish persistent sleep habits.

[6]We mistakenly assigned eight participants to treatments who did not have any baseline Fitbit data.

[7]We made two deviations in the treatments from the pre-registered experimental design. First, our original plan included an incentive treatment where participants would receive a $4.75 coupon for a breakfast treat at one of the University of Pittsburgh Einstein Coffee locations. However, due to unforeseen logistical difficulties, we suspended this treatment after the first few weeks of the first wave, and exclude it from the following waves. Second, the COVID-19 pandemic prevented us from meeting the pre-registered sample size for our main treatments (Control and Immediate Incentives), as we incurred additional costs for mailing Fitbits to participants.

To provide participants with a consistent cue, we sent them reminders to meet their target goal of sleeping 7 hours per night by 9 am every weeknight (Sunday-Thursday). These reminders had two major components. First, they included a personalized target bedtime calculated at the individual level based on each participants' baseline sleep patterns. We set a goal bedtime of about an hour earlier than their bedtime in the baseline period and sent them reminders to go to bed half an hour before their goal bedtime.

Second, as the cue-based framework emphasizes the importance of a 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".[8] Appendix Figure C.1 displays the bedtime reminder.

Next, to link sleeping behavior with a pleasurable 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 7 hours via the custom-made 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 Appendix Figure C.2). Participants in this treatment received a monetary reward of $4.75 through a Venmo transfer on the same day.[9] Participants who fell short of their 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 7 hours was discouraged (see e.g., Schultz et al., 2007), and an 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.

At the end of the intervention period, we discontinued the financial rewards.[10] In our main variation of the Immediate Incentives treatment, *Immediate Incentives – Post Cue/Feedback* (N=356, waves 1-7), we continued to send bedtime reminders (i.e., the cue) and morning feedback on whether they had achieved their sleep goal throughout the post-intervention

[8]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.

[9]For logistical reasons, the payment was received after 3pm each day, which introduced a 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.

[10]We notified participants at the end of the intervention period via text message.

period, which lasted through the end of classes. The feedback was identical to the intervention period, except that we removed mention of the financial reward, as displayed in Appendix Figure C.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.

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. Instead of receiving a $4.75 Venmo transfer each day they met their sleep goal, participants redeemed the incentive every day but only received one single transfer with the total payment at the end of the study period, one to five weeks after the intervention ended. Appendix Figure C.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 relative to 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 7 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.

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 Appendix Figure C.4). The participants in the Cue/Feedback treatment continued to receive the same reminders and feedback throughout the post-intervention period.

2.4 Outcome measures

2.4.1 Sleep

Our primary pre-registered sleep outcome is the share of weeknights (Sunday—Thursday) participants sleep at least 7 hours. Our secondary measures of sleep include sleep hours per night, sleeping between 7 and 9 hours on weeknights, sleeping at least 6 hours, bedtime, wake up time, sleep regularity, and sleep quality as measured by the Fitbit.

We define sleep regularity as the sleep variability across the week, as measured by the within-person standard deviation in the outcome. 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 a polysomnography (PSG).

Sync rates throughout the study were relatively high. On average participants synced their devices for 88% of the days. As shown in Appendix Table A.1, 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). 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 results do not change (Appendix Table A.3).

2.4.2 Educational outcomes

Our primary pre-registered educational outcome is term Grade Point Average (GPA), measured using administrative data. We obtained course data from the University of Pittsburgh Registrar Office for each semester from Fall 2018 through Spring 2023. The data allow us to have course information for our experimental participants in the term prior to our intervention, the term of our intervention and at least two terms following our intervention if a student was enrolled in any class for a grade.

Term GPA includes all courses in which a student received a letter grade, A+ through F, which we convert to a 4-point scale (see Appendix Table A.1 for the grading system). Term GPA averages the course grade points weighted by the number of credits for each course. GPA grades are conditional on course completion. In secondary analysis, we examine treatment effects on course completion and credits, as well as withdrawals, course failure and course pass rates.[11]

The course data allow us to classify courses by class type and start time. Class types include lectures, seminars, credit laboratories, praticum, workshops, independent studies, directed studies, internship, and laboratories. Lectures comprise 79.95% of the classes. As shown in Appendix Figure A.3, 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, where 67% of students receive the highest possible grade (compared to 47% of students in lecture courses). This raises concerns that the grading system in non-lecture courses leaves little scope for treatment effects. We therefore report estimated treatment effects for all course types and for lectures alone (we report the effects for non-lecture courses in Appendix Table A.6). We also classify each course as STEM or not STEM using the Department of Homeland Security (DHS) 2023 list of STEM designated CIP codes.[12]

Of 1149 participants in our experiment, 1128 have at least one course grade for the term of the intervention (98.1% 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

[11] Appendix Table A.1 details how each outcomes is defined. Our pre-registered secondary outcomes did not include withdrawals, course failure and course pass rates; and did include major, attainment, and academic behaviors if the data were available. Data were not available on major, attainment and additional academic behaviors.

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

analysis includes all available grades data. We also obtained data from the Registrar on participants' high school GPA, for which we could match 1049 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 1123 participants with baseline GPA (97.7% 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 the 848 participants, we match 833 to course grades in the term of the intervention (98.2%), 784 to high school GPA (92.4%) and an additional 41 to prior term GPA so that 825 participants have a baseline GPA (97.3%). As shown in Appendix Table A.1, 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 (Appendix Table A.7). There is no difference between the groups in the likelihood of having course grades, which is our primary outcome measure (Appendix Table A.2).

2.4.3 Additional outcomes[13]

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 service, 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.

Cognitive performance. We measured cognitive performance through math and creativity questions. We drew the math questions from the math section of the Graduate Record Examination Test (GRE) and we measured creativity using an adapted version of the task

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

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 one multiple choice math question or one creativity task. Both tasks were incentivized (see instructions in Appendix C). 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 4 undergraduate research assistants of 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 2 raters; the median number of ratings per sentence is 8.

Physical health. From the wearable trackers (Fitbits), we collected data on resting heart rate, 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).[14]

Well-being. To measure well-being, we collected measures of mood as well as measures of stress and ability to cope with stress (resilience). 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. Every week we alternated between mood and the stress/resilience questions. These measures were collected via text message and, every week, participants were randomly assigned to receive the text message at different times of the day (11am, 4pm, 9pm). We also collected measures of mental health in the intake survey (upon enrollment) and in the endline survey at the end of the semester. We measured 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 all of the times). An overall depression score is calculated by summing answers to all 20 items. We also measured

[14]In practice, this measure sums the lightly active, fairly active and very active minutes collected by the Fitbit. Our pre-registered secondary measures of health also include Body Mass Index (BMI) and blood pressure, which we did not collect due to logistical constraints.

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.

2.5 Randomization and baseline characteristics

The 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 7 hours (above vs below median).[15] 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, 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 (columns 2). Students in the control group are on average 19.4 years old, with a large share of freshmen (52%). Sophomore, junior, and senior and above students make up respectively 12%, 23% and 12% of the control group. Female students and Asian students are over-represented compared to the full-time Pitt (U.S. college) population. Approximately 56% (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%) 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.[16] 25% of the students in the control group report their parents' highest degree is less than a college education (either a high school degree and no college or a high school

[15] We did not balance the randomization on baseline GPA because GPA data was not available at the time of the randomization; we received GPA data once the study was completed.

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

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	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)
Observations	380	468	848

Notes: The sample is restricted to individuals in the Immediate Incentives treatment and individuals in the Control group. Early class sessions are classes starting at 10am 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

,

degree and some 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-college degree.

In the Control group, students slept at least 7 hours on only 44% of weeknights at baseline. On 29% of nights, they slept less than 6 hours. The average Control group bedtime was fifteen minutes past 1am and the average wake up time was 3 minutes before 8am.

About 58% of participants in the Control group are STEM majors. They are enrolled in an average of 5.2 courses with an average of 1.52 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 the University of Pittsburgh (Pitt) in 2022 had a weighted average GPA ranging from 3.91 to 4.42.[17]

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 compared to the Immediate Incentives group. We also estimate Treatment-Control differences for each treatment group separately in Appendix Table A.2 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 slept at least 7 hours),[18] high-school GPA (in quartiles), previous term GPA (in quartiles) if high-school GPA is missing, indicators for the number of classes starting before 10am in a week (0-5), and demographic

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

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

controls for gender, age in years (dummies), race/ethnicity (Asian, Black, Hispanic, White, other), and indicators for parents' highest education (less than college–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. ρ_t, w_t, μ_t, d_t are a set of fixed effects for, respectively, wave of the experiment, week of the experiment, month of the year, and day of the week. 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 primary treatment, the Immediate Incentives group. As pre-registered, we also present the analysis for the primary outcomes comparing the Control group to the pooled incentives treatments (see Appendix Tables A.4 and A.7).[19]

We report both unadjusted p-values and statistical significance adjusted for multiple hypothesis testing (MHT) within pre-registered families of secondary measures, using the method described in Anderson (2008).[20] Our primary outcome measures are sleeping at least 7 hours and course grades during the intervention term. The families of secondary outcomes we consider are: effects of secondary treatments, secondary sleep outcomes, secondary educational outcomes, time use, cognitive performance, physical health, and mental well-being. The secondary treatments include tests of each secondary treatment compared to Control; and each secondary treatment compared to Immediate Incentives. The secondary sleep measures include: sleep hours, bedtime, wake up time, sleeping more than 6 hours per night, sleeping between 7-9 hours per night, sleep regularity, bedtime regularity, wake up time regularity, sleep efficiency, REM sleep and Deep sleep. The secondary educational outcomes include course completion, credits, withdrawals, course failure and course pass rates. Time use includes all individual time use categories. Cognitive performance includes performance on the math and creativity tasks. Physical health includes heart rate, steps

[19] In a deviation from the pre-registered analysis we do not include instrumental variables (IV) analysis for GPA, instrumenting sleep. Our intervention may affect GPA through channels other than sleep – such as time allocation to other activities – and thus the IV exclusion restriction may be violated.

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

and physical activity. Mental well-being includes our measures of mood, stress, resilience to stress, anxiety (GAD - 7) and depression (CES - D). We compute the adjustment separately for treatment and post-treatment. 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. For outcomes measured by the Fitbit, we exclude eight participants who have no Fitbit data. For the remaining participants, we replace missing days of Fitbit data with an individual's average in the baseline period. In the Appendix, we report analyses that do not replace missing observations.

3 Results

We first examine treatment effects on sleep habits. We then turn to the impact of our intervention on educational outcomes. Finally, to explore potential mechanisms for our effects, we analyze time use, cognitive performance and physical activity and mental well-being.

3.1 Sleep

Baseline. Data from the baseline (pre-intervention) period reveals that a considerable portion of college students in our sample are sleep-deprived. Participants meet the recommendation of sleeping at least 7 hours on approximately 43% of the nights; sleep less than 6 hours on approximately 28% of the nights; and less than five hours more than 10% of the nights. About half of our participants have an average bedtime after 1 am, and about a quarter go to bed after 2 am on average. These data suggest that sleep deprivation is prevalent in our sample and is in line with a recent report by the National Institutes of Health indicating that more than 70% of college students sleep less than eight hours a day (Hershner and Chervin, 2014). In our sample participants sleep less than 8 hours approximately 84% of the nights.

In the intake survey, we asked participants how many hours they slept on a typical weeknight and how many hours of sleep they considered optimal. Consistent with previous studies, participants tend to report longer sleep than what is measured using Fitbit data (Avery et al., 2022; Lauderdale et al., 2008). 69% of the participants report sleeping at least 7 hours on a typical weeknight of the term (vs. 43% of nights as measured by the Fitbit). Compared to their (over-estimated) self-reported sleep, 81% of the participants state a longer optimal sleep duration during the week. On average, participants' stated optimal sleep time is an hour longer than what they report as their typical sleep duration. 97% of the students in our sample report that their optimal sleep on a weeknight would be at least 7 hours, and 81% report an optimal sleep time during the week longer than 8 hours. These results suggest there may be 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 7 hours on weeknights (Sunday - Thursday). This outcome variable only includes nighttime sleep, and excludes weekends, holiday and naps (defined as episodes of sleep that start between 7am-8pm). Figure 2 plots the estimated difference in the rate of sleeping at least 7 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).[21] 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 are 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).

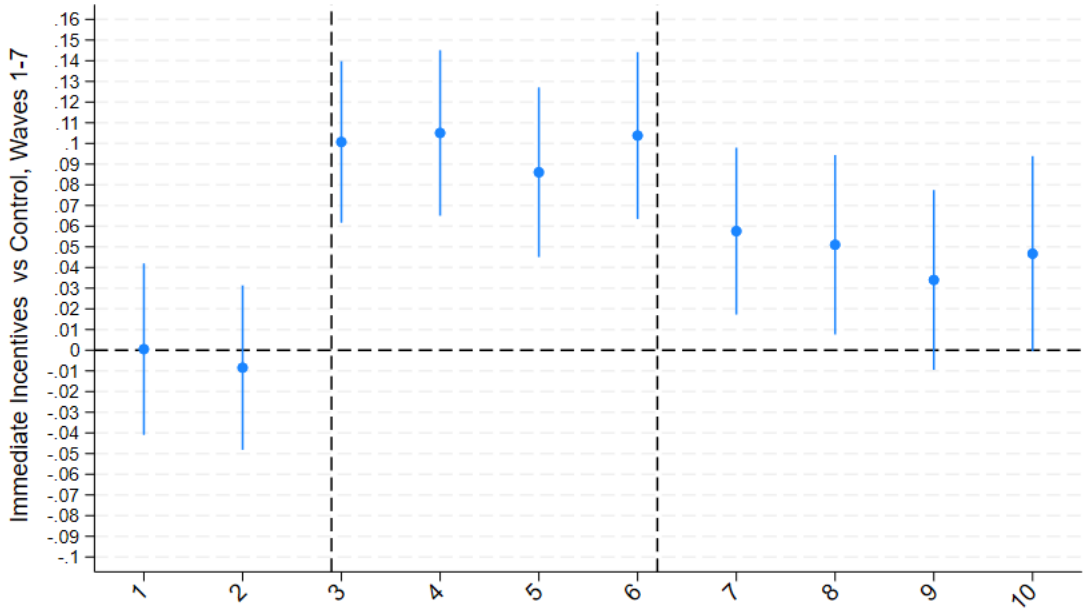
In the first two columns of Table 3 Panel A, we present regression estimates of the treatment and post-treatment impacts of Immediate Incentives on sleeping at least 7 hours on weeknights and treatment effects on weeknight sleep hours, following the specification described in equation 1. At baseline, participants meet the goal of sleeping at least seven hours on approximately 43% of the nights.[22] During the intervention period, Immediate Incentives increase the rate of sleeping at least seven hours by an estimated 11.8 percentage points, a 28% increase. The treatment effects persist into the post-intervention period but are about 53 percent smaller: an estimated 5.6 percentage points, 13% higher than baseline.[23]

[21]We present the analogous figures for sleep hours in Appendix Figure A.2 and distributions of sleep hours in Appendix Figure A.3.

[22]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.

[23]We conduct the following sensitivity analyses in Appendix Table A.4: 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, and exclude wave 3 (onset of COVID-19). The results do not change. We estimate treatment effects of 11.3-12.7 ppts and post-treatment effects of 5.4-6.4 ppts.

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.

We estimate that total sleep hours increase an estimated 19 minutes on average during the intervention period and an estimated nine minutes during the post-intervention period. All estimates are significant at the $p < 0.001$ level and are robust to adjusting for multiple hypothesis testing.

The effects at the mean reflect shifts throughout the distribution of sleep, as measured by sleep hours and share of nights sleeping at least 7 hours (Appendix Figure A.3). In Appendix Table A.5, we estimate treatment effects by baseline quartile of sleep (share of nights sleep at least 7 hours in panel A and sleep hours in panel B). We find similar effects across quartiles during the intervention period; and some evidence of larger post-intervention effects among participants with lower levels of sleep at baseline. These results suggest that our intervention has the most persistent impact on those who are most sleep deprived.

As shown in Appendix Table A.6 (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

Table 3: Immediate Incentives and sleep

	(1)	(2)	(3)	(4)
Panel A:				
Daily level	Sleep $\geq$ 7 hrs	Sleep hours	Bedtime	Wake-up time
Treatment	0.1186***	0.3203***	-0.3146***	-0.0471
	(0.013)	(0.035)	(0.036)	(0.034)
Post-Treatment	0.0551***	0.1420***	-0.0271	0.0712*
	(0.015)	(0.036)	(0.037)	(0.038)
Observations	46,989	46,989	46,989	46,989
Mean of dep. var.	0.429	6.647	25.22	7.939
Std. dev.	0.495	1.279	1.457	1.438
Number of individuals	840	840	840	840
Panel B:				
Regularity				
(within-individual weekly s.d.)		Sleep hours	Bedtime	Wake-up time
Treatment		-0.1084***	-0.0500**	-0.0701**
		(0.032)	(0.023)	(0.031)
Post-Treatment		(-0.0515	-0.0561**	-0.0930***
		(0.035)	(0.025)	(0.036)
Observations		8,631	8,631	8,631
Mean of dep. var.		1.171	0.896	0.937
Std. dev.		0.614	0.435	0.534
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. In Panel A, observations are the dependent variable at the daily level. In Panel B, observations are the standard deviation of the dependent variable at the weekly level. Panel B includes all fixed effects and controls listed above, except the day of the week fixed effects. 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.

increase in the likelihood of sleeping at least 7 hours during weekends in the post-intervention period.[24]

In Panel B of Appendix Table A.6, 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 7-9 hours, with persistent but smaller impacts in the post-treatment period.[25] 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. We note that, in our sample, baseline efficiency is high: participants are asleep an estimated 94% of the time they are in bed. By comparison, Bessone et al. (2021) estimate efficiency of 70% among their experimental participants in India.

3.1.1 Drivers of short and long-term sleep habits

Bedtime and wake up time. In the last two columns of Table 3 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 ($p < .001$) with directionally earlier wake up times ($p = .16$). 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 ($p = .065$). As shown by Appendix Figure A.2, both bedtime and wake up time get progressively later over the course of the intervention period and stabilize during the post-intervention period. These results suggest that combining bedtime reminders with incentives initially induces participants to go to bed earlier, but does not establish a sustained habit linked to the bedtime cue.

Sleep regularity. Panel B of Table 3 estimates treatment and post-treatment effects on sleep regularity. To do so, we examine the within-individual standard deviation of total sleep hours, bedtime and wake up time at the week level (the level of observation is individual-week).[26] We find significant decreases in sleep variability across the week, equivalent to about

[24]Including naps, holidays and weekends, we estimate the intervention increased the share of nights with at least 7 hours of sleep by 6.9 percentage points in treatment and 4.2 percentage points in post-treatment, and increased total sleep hours by an estimated 13 minutes in the treatment period and 6 minutes in the post-treatment period ($p < 0.01$ for all estimates).

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

[26]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.

10 percent of baseline, or 0.1 standard deviations. The magnitude of the effects during the intervention and post-intervention periods are of similar size. These findings show that, while treated participants do not on average sustain earlier bedtimes after the intervention ends, they do develop more regular bedtime and wake up time habits. That the habits persist into the post-intervention period suggests treated participants found personal bedtimes and wake up times they were able to maintain. Such 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; Trockel et al., 2000; Smarr, 2015).

Secondary treatments. As discussed in Section 2 and summarized in Table 1, our three pre-registered secondary treatments vary elements of our primary Immediate Incentives treatment in order to investigate the importance of cues and immediate rewards: (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 4 estimates the effects of our primary and secondary treatments on sleep hours and sleeping at least 7 hours on weeknights. We restrict the analysis to waves 1-3 of the experiment when the secondary treatments were conducted. As shown in columns 1-2, the effects of Immediate Incentives during treatment are about 53% to 86% percent higher than the effect of Delayed Incentives (with or without reminders and feedback); and about three to four times higher than the effects of reminders and feedback alone. The differences between the estimated impact of Immediate Incentives and each of the secondary treatments are all significant at the $p < 0.05$ level after adjusting for multiple hypothesis testing. During the post-intervention period, the estimated effects of Immediate Incentives are generally larger than those of the secondary treatments. However the effects are statistically indistinguishable.

We next focus on the role of the cue and feedback for developing and sustaining habits in combination with rewards. 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 Delayed Incentives No Cue/Feedback treatment has more persistent effects in the post-intervention period. In Appendix Table A.4, columns 8 and 9, we estimate the post-treatment effects of the Immediate Incentives intervention separately for the subgroup of participants who continued to receive personalized bedtime reminders and feedback in

Table 4: Secondary treatments

	(1) Sleep ≥ 7	(2) Sleep hours
Treatment		
Immediate Incentives	0.1433***	0.3270***
	(0.021)	(0.054)
Delayed Incentives	0.0945***	0.1781***
	(0.022)	(0.058)
Delayed Incentives, No Cue/Feedback	0.0890***	0.1806***
	(0.022)	(0.059)
Cue/Feedback Only	0.0364*	0.1145**
	(0.020)	(0.052)
Post-Treatment		
Post: Immediate Incentives	0.0487*	0.2083***
	(0.025)	(0.067)
Post: Delayed Incentives	0.0143	0.0403
	(0.027)	(0.071)
Post: Delayed Incentives, No Cue/Feedback	0.0551**	0.0932
	(0.028)	(0.076)
Post: Cue/Feedback Only	0.0144	0.0778
	(0.027)	(0.073)
Observations	34,954	34,954
Mean of dep. var.	0.434	6.696
Std. dev.	0.496	1.536
Number of individuals	589	589

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 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.

the post-treatment period (Immediate Incentive with Cue/Feedback) and for participants who stopped receiving them at the end of the intervention-period (Immediate Incentive No Cue/Feedback). We restrict the analysis to waves 5 and 7 in which we ran both variants. Our estimates reveal no significant differences between these two subgroups, suggesting that receiving bedtime cues after incentives stopped did not further help sustain the routines developed during the intervention period.

Collectively, the results presented in this section do not provide strong evidence that individuals built automatic habits as a result of our external cue. In the post-intervention period average bedtime reverted to baseline despite the bedtime cue. We also find little evidence that the external cue enhanced the impact of incentives during the intervention period or the persistence of habits during the post-intervention period. Nonetheless, the observed increase in sleep regularity (i.e., the reduced variability in sleep hours, bedtime and wakeup) persists in post-treatment. This suggests that the intervention facilitated the establishment of more dependable sleep routines, irrespective of the external cues provided.

3.2 Educational outcomes

Next, we investigate the impact of our primary treatment on educational outcomes.

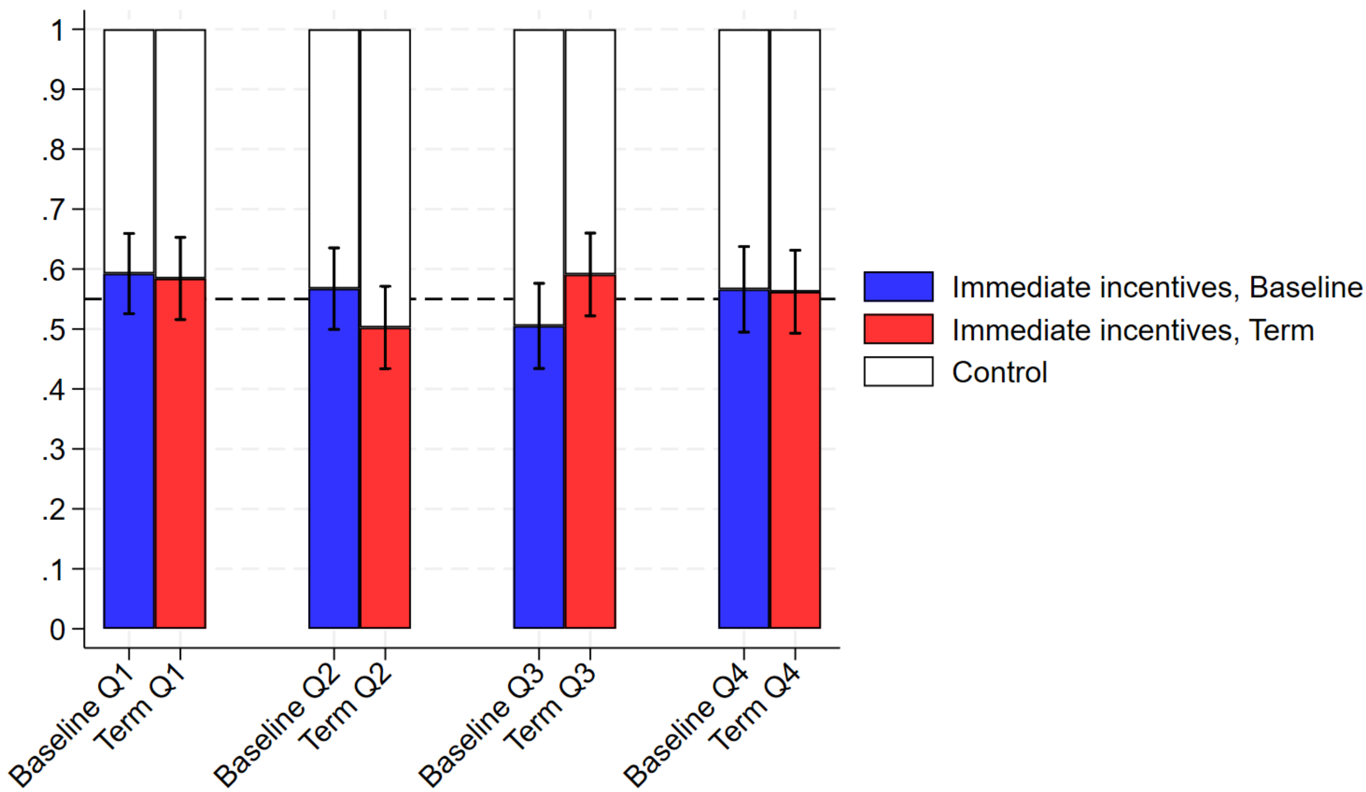
Figure 3 displays the share of individuals in the Immediate Incentives treatment and in the control group who are in each quartile of the GPA distribution at baseline and at the end of the intervention term.[27] The figure highlights that, as compared to baseline GPA, the share of treated participants in the bottom two quartiles declines and the share in top two quartiles increases, with the biggest shift in the middle two quartiles. Table 5 presents regression estimates of the impact of the intervention on semester GPA and secondary educational outcomes.[28] The regressions follow the specification for 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.

As shown in columns 1 and 2, we estimate that Immediate Incentives improved average course performance by 0.075 grades points in all classes ($p = 0.044$) and 0.088 grade points

[27] We use high school GPA for baseline and only include participants with high school GPA in the figure.

[28] As discussed in Section 2.6, we are missing GPA for 1.9 percent of our participants. We examine differential attrition on the GPA measure in Appendix Table A.1 and find no evidence for differential attrition on term GPA.

Figure 3: Immediate Incentives and GPA



Notes - The figure reports the share of individuals in each quartile of the GPA distribution for both baseline GPA (the high-school GPA) and the term GPA during the intervention for the Immediate Incentive treatment. Bars indicate 95% confidence intervals.

in lecture classes ($p = 0.035$).[29]

We estimate a treatment effect of similar magnitude on course performance in the semester following the intervention, providing suggestive evidence of persistent effects (columns 3 ($p = 0.066$) and 4, $p = 0.027$). However, we do not find treatment effects in the semester two terms after the intervention (columns 5 and 6).

[29]We conduct the following sensitivity analyses in Appendix Table A.7: limit the covariates to wave fixed effects, gender, baseline sleep and baseline GPA; limit the sample to participants who have baseline GPA (high school or baseline term GPA); limit the sample to participants who have post-intervention term grades; limit the sample to participants who have sleep data at baseline; and, exclude wave 3 (onset of COVID-19). The estimated effects are slightly smaller when limit the covariates or the sample, 0.060-0.067 grades points for all classes and 0.076-0.079 for lectures. Our estimated impacts increase when we exclude wave 3. Estimates are instead slightly smaller when reweighing the sample with respect to gender to make it representative of the gender composition of US college students. Following our pre-registration, we also report results where we conduct our main analysis pooling all incentives treatments. The estimates are similar with slightly smaller average impacts, as shown in Table A.4 for sleep and Table A.7 for grades.

Table 5: 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 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.

The estimated GPA impacts of 0.075 - 0.088 grade points during the intervention semester are equivalent to a 0.10 - 0.11 standard deviation (SD) increase in grades. In Appendix Table A.7 we report various sensitivity checks on course grade. Interestingly, when we exclude the Spring 2020 semester (wave 3), which experienced the onset of the COVID-19 pandemic leading to the abrupt closure of the university in the middle of the semester and disruptions in sleep and other lifestyle habits (Giuntella et al., 2021), our estimated treatment impact on term GPA is slightly higher, an estimated 0.09 grade points or 0.11 SD and 0.105 grade points or 0.13 SD when restricting to lectures. Panel C of Table A.7 documents how our intervention had no effect in non-lecture classes (i.e. seminar, labs, internships, directed studies). It is worth noting that the median grade in these classes is A, and the lowest quartile is A-, leaving little room to improve grades in these classes (see Figure A.1).

Turning to heterogeneity, Table 6 shows that the results are not driven by performance in early-morning classes but rather the largest effects are in late morning/early afternoon classes that occur between 10am and 2pm, followed by afternoon/evening classes (after 2pm). This is in line with the findings from Carrell et al. (2011), who find that early class start time affect performance in all classes, not just classes taking place early in the morning. Further, these results are driven by STEM courses: on average our intervention leads to a .13 grade points increased on grades in STEM courses, which corresponds to a 0.15 SD increase in grades. By contrast, we estimate small increases of 0.018 grade points in non-STEM courses.

Table 6: 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 10am, 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.

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 (over 57%), another possibility is that improved sleep could enhance performance in courses important for a given major. In exploratory analysis, we find larger effects among women. Further, first-term freshmen students exhibit substantially larger effect, (Appendix Table A.8), consistent with the finding that point estimates for both treatment and post-treatment on sleep are larger among these students (Panel A). 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). Point-estimates on sleep are larger among female, first-term, and STEM-major students.

We also explore the effects of our intervention on other academic outcomes (Table A.9). Interestingly, students in the treatment group are less likely to receive a grade (column 1). This is mostly reflecting the increase in the likelihood of withdrawing from a class (column 2). Students in treatment group are marginally less likely to fail a course (column 3). As a result, there are no significant differences in the likelihood of passing a class nor in the number of credits completed in a term.[30]

To benchmark our intervention, 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 an hour later school start time increases sleep by 35 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 0.081 SD increase in math scores and a non-significant 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.[31] Our impacts of 19 minute average increase in weeknight sleep during treatment, a 9 minute average increases in post-treatment

[30] Table A.1 describes the grading system.

[31] 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 0.1 SD and leads to 0.14 more years of schooling.

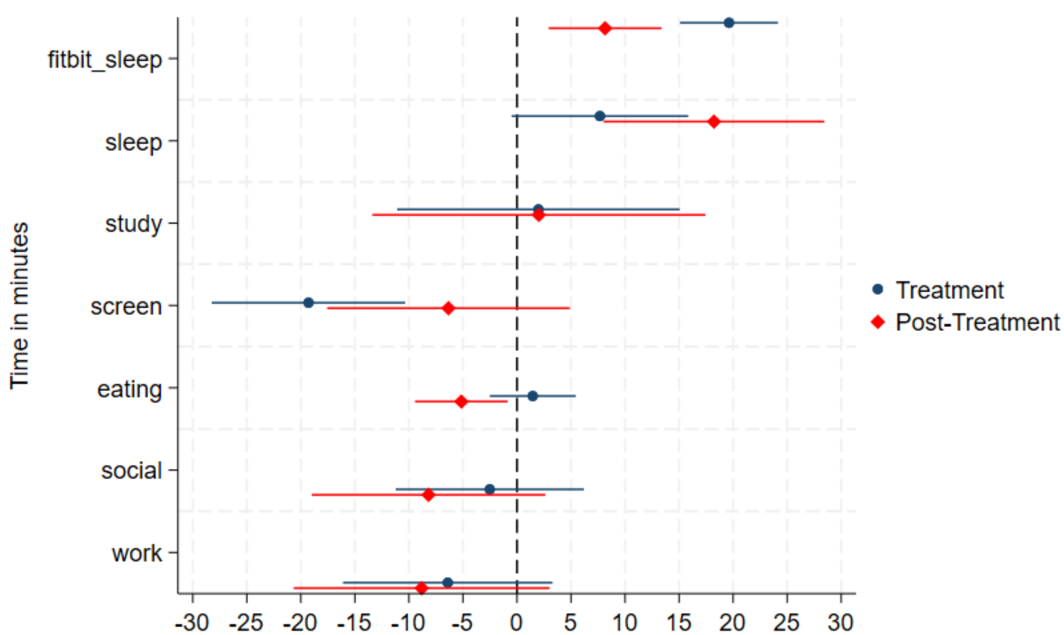
and a 0.10 - 0.11 SD improvement in grades falls within the range of the estimates in prior work on the causal relationship between shifts in sleep and changes in academic performance.

3.3 Additional measures

To help make sense of our results, throughout the study we collected measures of time use, cognitive performance in math and creativity tasks, physical health via the Fitbit and well-being.

Lifestyle. We next focus on our survey measures of time use that we asked weekly throughout the study (see Section 2 for details). Figure 4 shows estimated treatment and post-

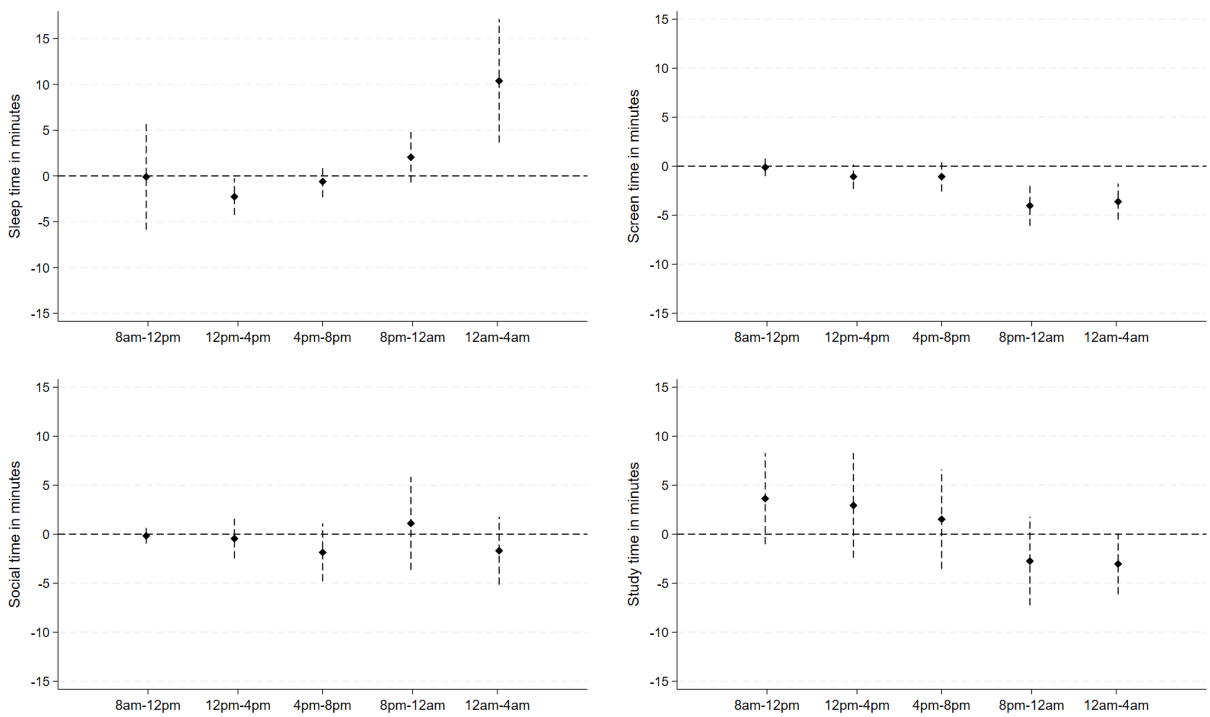
Figure 4: Incentives to sleep and time use (minutes)



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 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 was missing. Standard errors are clustered at the individual level. Bars indicate 95% confidence intervals.

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 9 minutes after the removal of the incentives. Immediate Incentives directionally increase self-reported sleep in both the intervention and post-intervention periods by about 7 - 18 minutes on average per day (Table A.10). We also find that subjects were 6.6 (7.4) percentage points more likely to report at least 7 hours of sleep during the intervention (in the post-intervention) period.

Figure 5: 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 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 was missing. Standard errors are clustered at the individual level. Bars indicate 95% confidence intervals.

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 19 minutes per day ($p < .01$). We estimate smaller and not statistically significant treatment effects on screen time during the post-intervention period. These estimates are robust to adjusting for multiple hypothesis testing ($p < .01$ for share nights with less than 7 hours of sleep; $p < .01$ for screen time). We do not find evidence of meaningful changes in time spent studying, socializing, eating or working.[32] In Appendix Table A.10 we report estimates on all the time use categories.

In Figure 5, 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 (8pm - 4am). Interestingly, while total study time does not increase, we observe a reallocation of study time from the evening/night (8pm—4am) to the morning (8am - 12pm), 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 also estimate treatment effects during the post-intervention period and find a similar, but weaker, patterns (Appendix Figure A.4).

Cognitive performance. To examine cognitive performance directly, we collected measures of performance in a math and creativity tasks on alternating weeks throughout the study. We do not find any impact of the intervention on these proxies for cognitive performance (see Appendix Table A.11, columns 1 and 2), which could be due to the intervention not affecting cognitive performance or to our measures not being able to capture the impact of performance on cognition.

Well-being and physical health.

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 wellbeing and physical health (Giuntella and Mazzonna, 2019; Giuntella et al., 2017; Jin and Ziebarth, 2020).

To investigate the impact of the intervention on well-being by employing 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 Epidemiologic Studies Depression (CES-D) scale to gauge depression levels. These scales were administered at baseline and endline only and 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

[32] We estimate treatment effects separately for internet, TV/videos and games in Appendix 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 (486 minutes), study (326 minutes), screen (170 minutes), eating and preparing food (94 minutes), social (104 minutes), work (94 minutes).

Table 7: Immediate Incentives and well-being

	(1) Happiness	(2) Stress	(3) Resilience	(4) CES-D	(5) GAD-7
Treatment	-0.0655	0.0631	0.1503***		
	(0.108)	(0.059)	(0.055)		
Post-Treatment	-0.0569	-0.0411	0.0640	0.3997	0.0818
	(0.111)	(0.069)	(0.063)	(0.886)	(0.404)
Observations	4,166	3,629	3,558	1,462	1,462
Mean of dep. var.	6.404	3.115	2.993	15.78	6.832
Std. dev.	1.646	1.116	0.997	10.23	4.864
Number of individuals	794	800	794	834	834

Notes:
Estimates in columns 1-3 include day of the week, week of the experiment, 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 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. 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

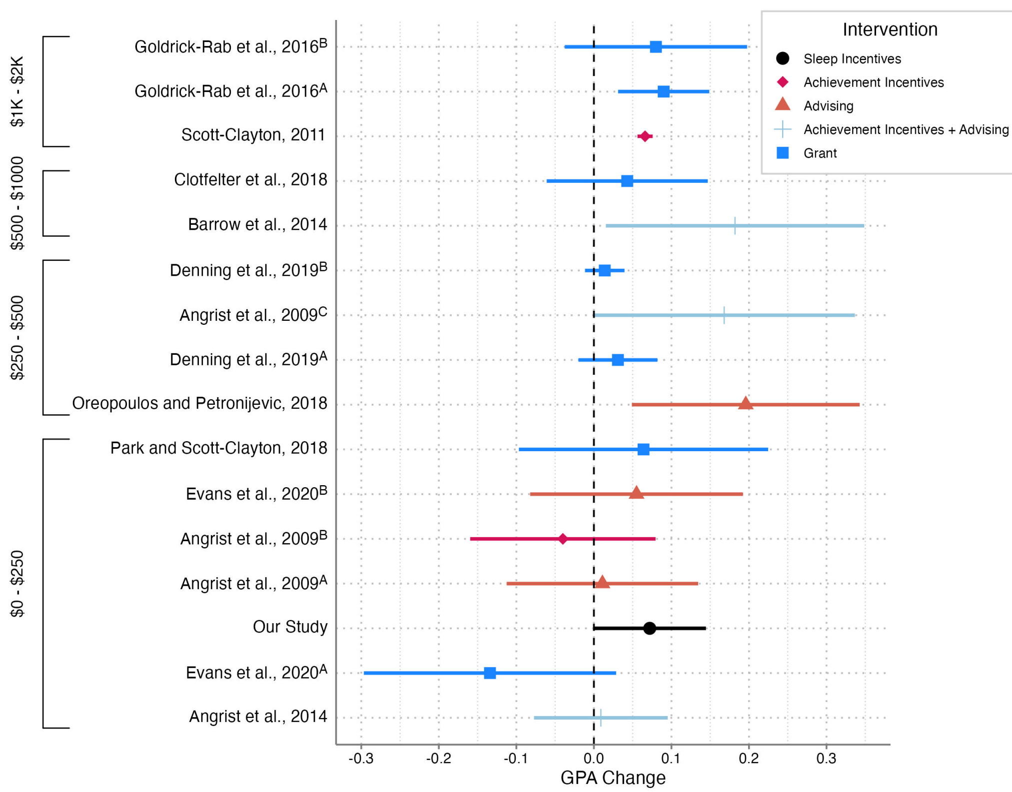
is significant at the 10% level after adjusting for multiple hypothesis testing. On the other hand, the intervention did not show any significant effects on post-treatment measures of depression and anxiety (columns 4 and 5), with point estimates being small in magnitude and lacking statistical significance.

As discussed in Section 2, we use the Fitbit to measure participants' heart rate, daily steps, 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 and 5).

4 Discussion and conclusion

In this paper, we show that an intervention targeting sleep habits improves academic performance. To benchmark our results, we compare the cost effectiveness of our intervention to prior work examining policies aimed at improve 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 $1250 per student per semester. By comparison, we estimate that incentives to sleep increase semester GPA by 0.075 grade points and cost approximately $110 per participant for the semester and would cost about $160 per participant per year. This includes $60 for the cost 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 6: Cost-Effectiveness



Notes - The figure compares our main effects 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 GPA effects.

In Figure 6, we report the estimated effects on GPA from prior work examining the impact of achievement incentives, advising, and grants, ordered from most to least costly.[33] As depicted in the figure, our intervention is characterized by relatively 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 fivefold higher cost per participant. Our results demonstrate that focusing on sleep can be a cost-effective approach to improving educational outcomes.

We explore mechanisms for the impact of our intervention on both sleep habits and academic performance. Inspired by cue-based theories of habit formation, our Immediate Incentives intervention aimed to establish automatic habits through repeated exposure to recurring cues coupled with immediate rewards. The intervention increases sleep during the treatment period with smaller persistent effects in the post-treatment period. Our results show that Immediate Incentives can enhance habit formation during the intervention period, compared to variants with delayed or no rewards. However, we find little evidence that immediate incentives generate automatic habits triggered by the external cue. Instead, our results point to participants developing their own routines that persist into the post-treatment period. This could reflect some combination of treated participants acquiring a taste for sleep (i.e., increased benefits) and also finding sleep behaviors that are easier to sustain (i.e., lower costs). Future research could develop intervention designs that separately identify mechanisms of habit formation, including, automaticity, learning about benefits and lowering costs (Volpp and Loewenstein, 2020).

We then examine channels through which sleep may influence academic performance, including cognitive function, lifestyle factors, and overall well-being. While we do not detect an impact of our intervention on performance in math questions or creativity, sleep could have influenced cognition through channels like attention or memory consolidation, which were not captured by our measures (Diekelmann and Born, 2010). Examining lifestyle, our intervention led to a decrease in screen time and a reallocation of study time to morning hours, when students are potentially more alert and able to focus. Finally, we find evidence of a positive impact on students' ability to cope with stress, which may in turn have affected their academic performance. Further investigation of these mechanisms in future research can provide a deeper understanding of the multifaceted contributions of sleep to educational outcomes.

Taken together, our results show that offering incentives in the middle of the semester can improve term GPA. This result is consistent with recent evidence from Liu et al. (2022)

[33] Achievement incentives include performance-based incentives and merit aid. Advising includes advising and support services, see A.12. 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.

who show that engagement interventions are more effective in the middle of the term. Future work could examine targeting the intervention (for example to first-term freshmen), the role of the timing of the intervention, and the impact of longer (or shorter) interventions in order to understand how to cost-effectively sustain effects on academic performance across multiple terms.

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.

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.

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.

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.

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 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.

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.

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.

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.

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.

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.

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.

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*.

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.

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.

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.

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.

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

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.

Trockel, M. T., Barnes, M. D., and Egget, D. L. (2000). Health-related variables and academic performance among first-year college students: Implications for sleep and other behaviors. *Journal of American College Health*, 49(3):125–131.

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.

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.

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

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.

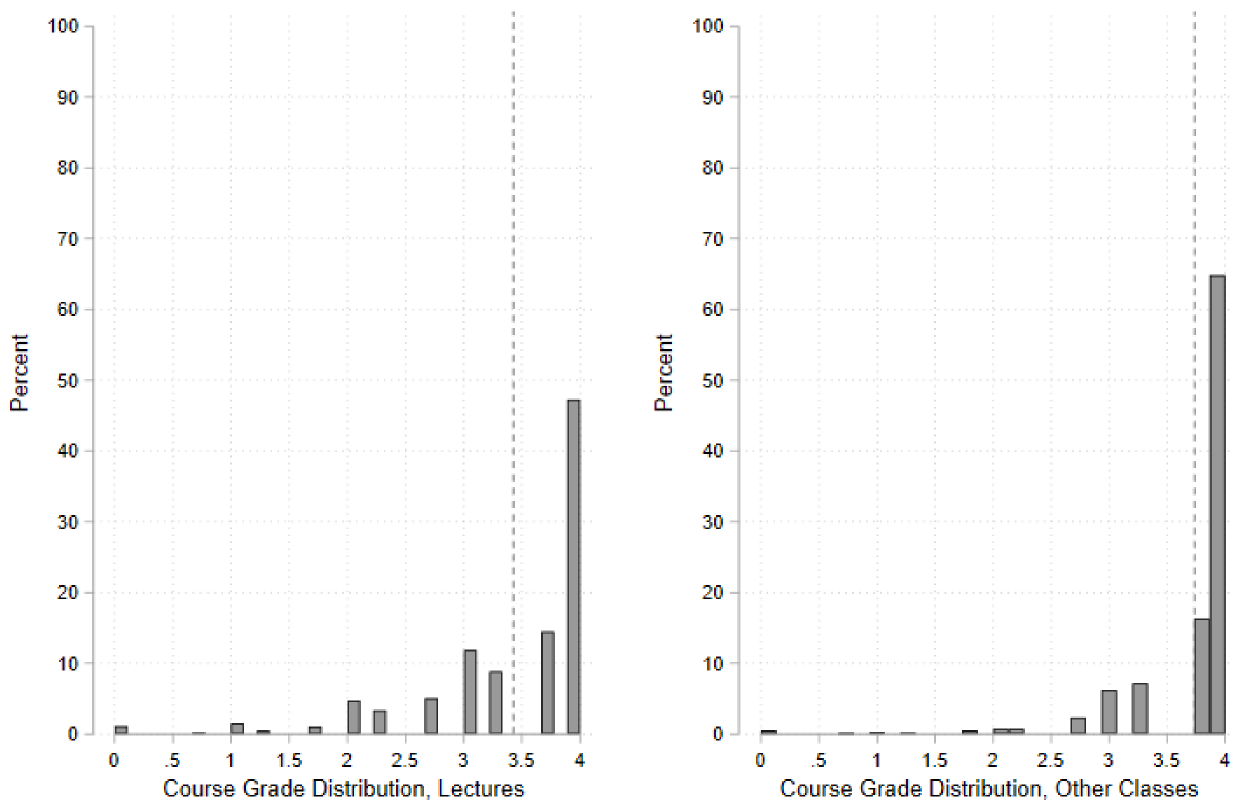
Wright, S. P., Hall Brown, T. S., Collier, S. R., and Sandberg, K. (2017). How consumer physical activity monitors could transform human physiology research. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 312(3):R358–R367.

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

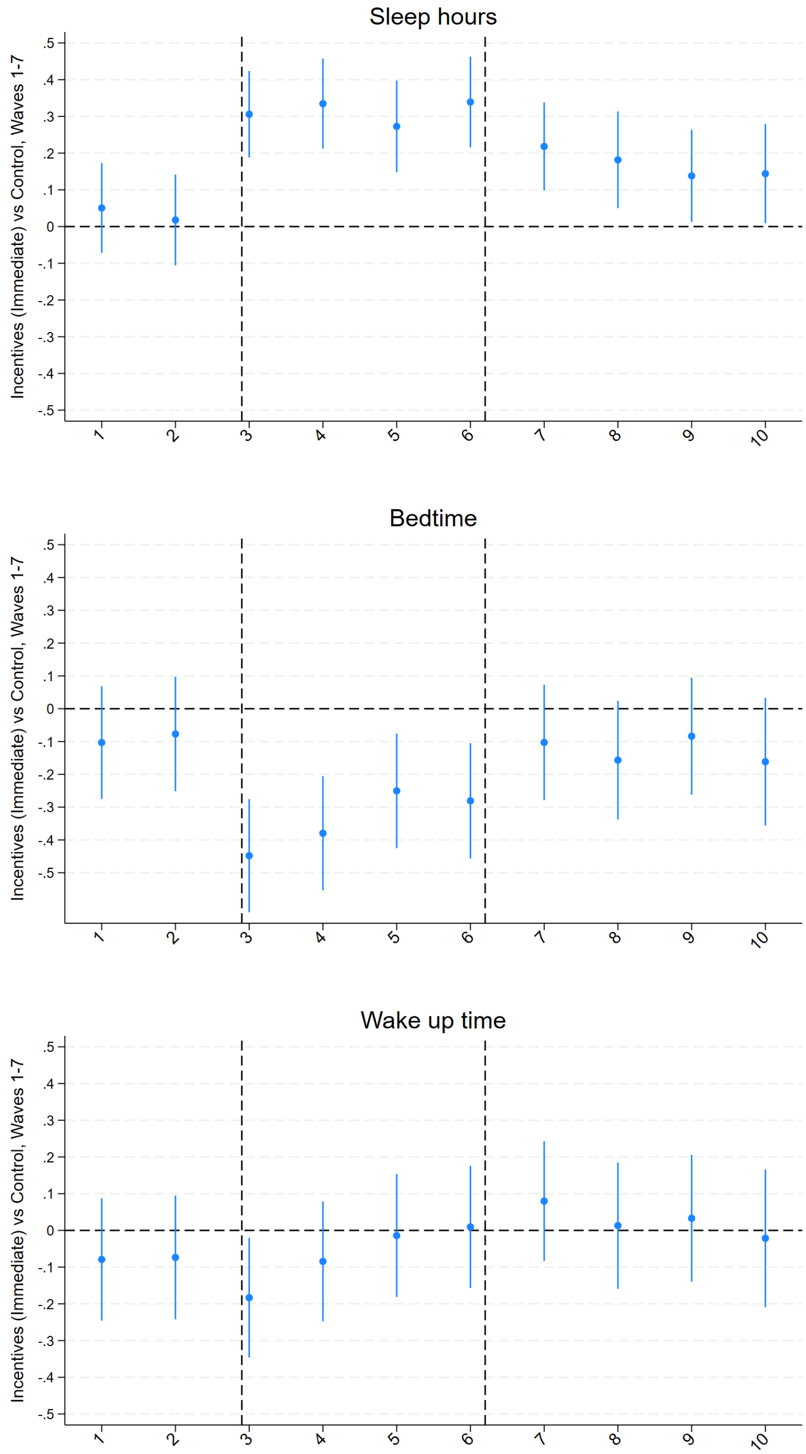
A. Figures and tables

Figure A.1: Grades distribution



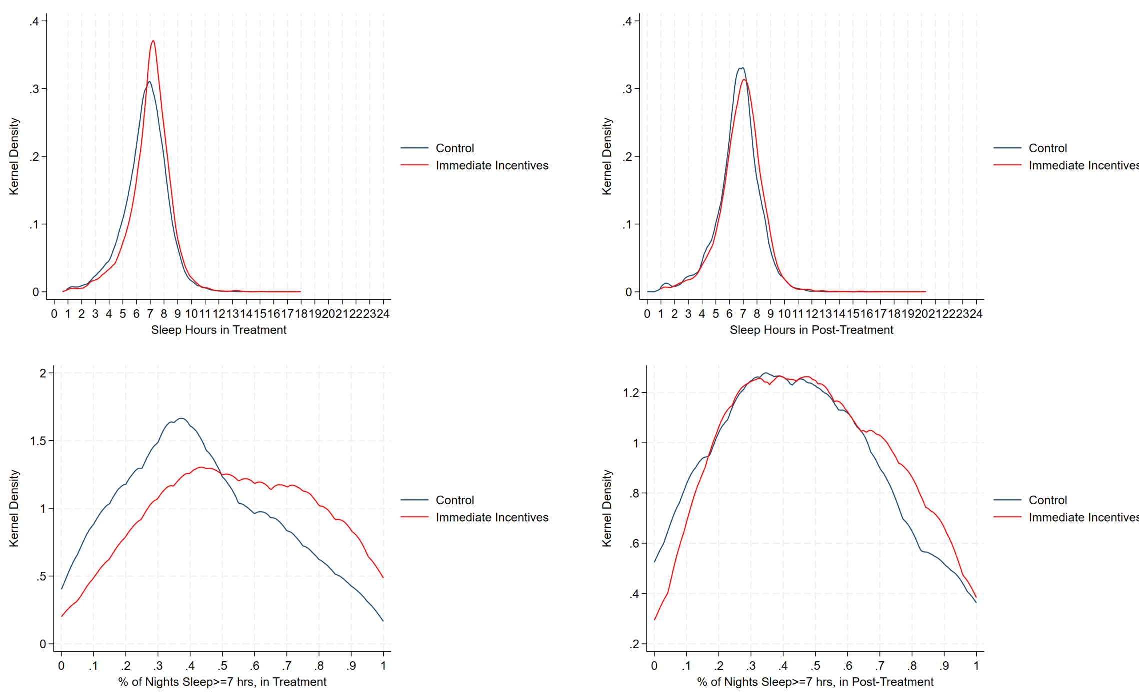
Notes: The figure reports the distribution of grades in lectures and other classes. The dashed vertical line identifies the average grade in these class types

Figure A.2: Immediate Incentives, sleep hours, bedtime and wake-up time



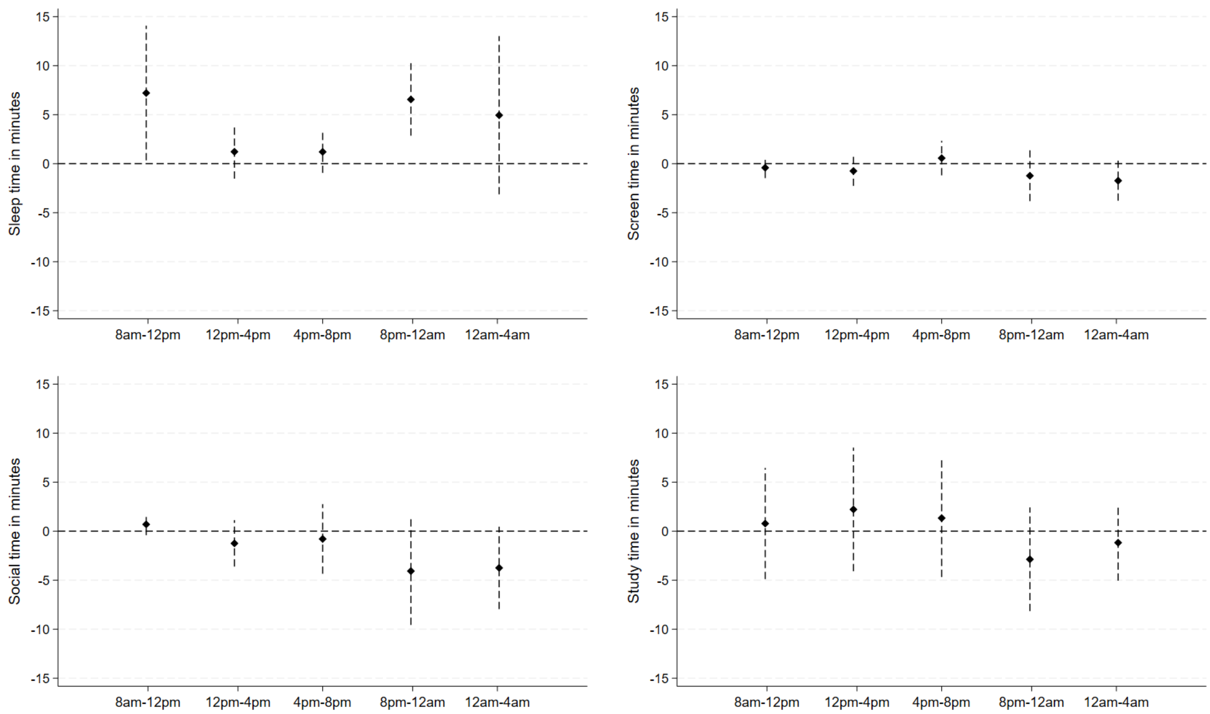
Notes - The sample is restricted to weekdays (Sunday-Thursday nights). On the horizontal axis we report time in weeks since the study started (week=3 is the first week of treatment, week=6 is the last week of treatment). The coefficient reports the differences in average sleep hours, bedtime, and wake 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.3: Immediate incentives and distribution of sleep during and after the intervention



Notes: The figure reports the distribution of sleep hours during (top panel) and after the intervention (bottom panel) for participants in Control group (navy) and Immediate Incentives treatment (red).

Figure A.4: 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 during the intervention throughout the day. All the coefficients are obtained from regressions including 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 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) Course grade	(4) Quality points	(5) Has a grade	(6) Withdrawn	(7) Passed	(8) Credit completed
A+	YES	YES	4	YES	NO	YES	YES
A	YES	YES	4	YES	NO	YES	YES
A-	YES	YES	3.75	YES	NO	YES	YES
B+	YES	YES	3.25	YES	NO	YES	YES
B	YES	YES	3	YES	NO	YES	YES
B-	YES	YES	2.75	YES	NO	YES	YES
C+	YES	YES	2.25	YES	NO	YES	YES
C	YES	YES	2	YES	NO	YES	YES
C-	YES	YES	1.75	YES	NO	YES	YES
D+	YES	YES	1.25	YES	NO	YES	YES
D	YES	YES	1	YES	NO	YES	YES
D-	YES	YES	0.75	YES	NO	YES	YES
F	YES	YES	0	YES	NO	NO	NO
G	NO	NO	0	NO	NO	NO	NO
H	NO	NO	0	YES	NO	YES	YES
HS	NO	NO	0	YES	NO	YES	YES
I	NO	NO	0	NO	NO	NO	NO
N	NO	NO	0	NO	NO	NO	NO
NC	NO	NO	0	NO	NO	NO	NO
NG	NO	NO	0	NO	NO	NO	NO
R	NO	NO	0	NO	NO	NO	NO
S	NO	NO	0	YES	NO	YES	YES
U	NO	NO	0	YES	NO	NO	NO
W	NO	NO	0	NO	YES	NO	NO

Notes -
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			
	# Fitbit days before intervention	# Fitbit days during intervention	# Fitbit days post intervention	Has HS GPA	Has baseline GPA	Has course grades
Immediate Incentives	0.145	0.779**	-0.096	0.017	0.046*	0.008
	(0.260)	(0.321)	(0.406)	(0.020)	(0.025)	(0.010)
Observations	840	840	840	840	840	840
Mean of dep. var.	11.89	18.17	13.03	0.895	0.782	0.981
Std. dev.	6.021	4.100	6.428	0.307	0.413	0.137
			Panel B			
	Has time use	Has math task	Has creativity task	Has mood survey	Has resilience survey	Has mental health
Immediate Incentives	0.006	0.010	0.010	0.005	-0.003	-0.001
	(0.007)	(0.013)	(0.015)	(0.011)	(0.012)	(0.006)
Observations	840	840	840	840	840	840
Mean of dep. var.	0.988	0.872	0.951	0.967	0.959	0.991
Std. dev.	0.110	0.334	0.216	0.178	0.198	0.0931

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. *** p<0.01, ** p<0.05, * p<0.1

Table A.3: Differences in baseline characteristics: Secondary treatments

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Female	Age	Asian	Black	Hispanic	White	Other	
Delayed Incentives	0.0111	0.1893	-0.0640	0.0126	-0.0104	0.0698	-0.0080	
	(0.060)	(0.492)	(0.059)	(0.037)	(0.021)	(0.065)	(0.022)	
Delayed Incentives, No Feedback	-0.0164	-0.0607	-0.0290	-0.0020	0.0110	0.0160	0.0041	
	(0.062)	(0.390)	(0.061)	(0.034)	(0.027)	(0.066)	(0.026)	
Feedback Only	0.0051	-0.0994	-0.1195**	0.0636	0.0194	0.0143	0.0222	
	(0.060)	(0.385)	(0.056)	(0.041)	(0.028)	(0.066)	(0.027)	
Observations	442	442	442	442	442	442	442	
Mean of dep. var.	0.686	19.83	0.271	0.0973	0.0362	0.554	0.0407	
Std. dev.	0.465	3.270	0.445	0.297	0.187	0.498	0.198	
Immediate Incentives	0.0139	-0.0302	0.0582	-0.0344	0.0190	-0.0083	-0.0344	
No Cue/Feedback in Post	(0.056)	(0.252)	(0.058)	(0.034)	(0.033)	(0.066)	(0.034)	
Observations	224	224	224	224	224	224	224	
Mean of dep. var.	0.777	19.31	0.259	0.0714	0.0625	0.536	0.0714	
Std. dev.	0.417	1.927	0.439	0.258	0.243	0.500	0.258	
	Sleep ≥ 7 hours	Sleep ≥ 6 hours	Sleep hours	Bedtime	Wake-up time	HS GPA	Baseline GPA	
Delayed Incentives	-0.0338	-0.0197	0.0081	-0.2647*	-0.3047**	0.0633	-2.1743	
	(0.034)	(0.029)	(0.120)	(0.146)	(0.142)	(0.053)	(4.442)	
Delayed Incentives, No Feedback	-0.0442	-0.0246	-0.0731	0.0056	-0.0501	0.0952*	-3.8859	
	(0.033)	(0.030)	(0.108)	(0.150)	(0.149)	(0.057)	(4.570)	
Feedback Only	-0.0210	-0.0092	-0.0199	-0.2074	-0.2445*	0.1036*	-5.5754	
	(0.033)	(0.030)	(0.101)	(0.143)	(0.144)	(0.059)	(4.293)	
Observations	444	444	444	444	444	385	437	
Mean of dep. var.	0.447	0.733	6.754	24.99	7.774	4.126	22.16	
Std. dev.	0.259	0.230	0.841	1.128	1.116	0.399	38.08	
Immediate Incentives	-0.0142	-0.0380	-0.1351	0.0074	-0.0388	-0.0161	-0.0693	
No Cue/Feedback in Post	(0.036)	(0.034)	(0.117)	(0.175)	(0.161)	(0.060)	(1.703)	
Observations	224	224	224	224	224	215	223	
Mean of dep. var.	0.457	0.737	6.758	25.23	8.064	4.138	5.212	
Std. dev.	0.267	0.254	0.886	1.307	1.222	0.444	12.71	
	Freshman	Sophomore	Junior	Senior	STEM major	Less than college	College	More than college
Delayed Incentives	-0.0242	-0.0466	0.0479	0.0354	0.1369**	-0.0274	0.0350	-0.0077
	(0.063)	(0.045)	(0.056)	(0.049)	(0.062)	(0.059)	(0.062)	(0.064)
Delayed Incentives, No Feedback	-0.0327	-0.0361	0.0218	0.0595	0.0051	0.0196	0.0565	-0.0760
	(0.065)	(0.048)	(0.056)	(0.052)	(0.067)	(0.061)	(0.063)	(0.065)
Feedback Only	-0.0303	0.0075	-0.0164	0.0516	0.1016	-0.0123	-0.0069	0.0192
	(0.062)	(0.049)	(0.053)	(0.051)	(0.064)	(0.060)	(0.060)	(0.065)
Observations	444	444	444	444	442	442	442	442
Mean of dep. var.	0.432	0.160	0.223	0.180	0.620	0.287	0.312	0.400
Std. dev.	0.496	0.367	0.417	0.385	0.486	0.453	0.464	0.491
Immediate Incentives	0.1002*	-0.0253	-0.0412	-0.0337	0.0718	0.0321	0.1034*	-0.1355**
No Cue/Feedback in Post	(0.060)	(0.037)	(0.051)	(0.042)	(0.067)	(0.062)	(0.061)	(0.066)
Observations	224	224	224	224	224	224	224	224
Mean of dep. var.	0.607	0.0848	0.192	0.116	0.522	0.299	0.290	0.411
Std. dev.	0.489	0.279	0.395	0.321	0.501	0.459	0.455	0.493

Notes: All estimates include wave fixed effects. Robust standard errors are in parenthesis. *** p<0.01, ** p<0.05, * p<0.1

Table A.4: Immediate incentives and sleep: Sensitivity analyis

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Primary Specification	Basic Controls	Has term GPA	Excludes missing nights	Excludes wave 3	Weighted by gender	Incentives, pooled	No Cue in Post
Treatment	0.1186***	0.1134***	0.1172***	0.1270***	0.1199***	0.1139***	0.1094***	0.1193***
	(0.013)	(0.011)	(0.013)	(0.012)	(0.014)	(0.014)	(0.012)	(0.021)
Post-Treatment	0.0551***	0.0607***	0.0546***	0.0539***	0.0635***	0.0565***	0.0530***	0.0626**
	(0.015)	(0.011)	(0.015)	(0.014)	(0.016)	(0.015)	(0.014)	(0.025)
Immediate Cash*								0.0097
No Cue in Post-Treatment								(0.027)
Observations	46,989	46,989	46,146	35,182	41,753	46,989	58,833	19,932
Mean of dep. var.	0.429	0.429	0.427	0.432	0.434	0.429	0.435	6.742
Std. dev.	0.495	0.495	0.495	0.495	0.496	0.495	0.496	0.498
Number of individuals	840	840	825	840	763	840	1040	357

Notes: The sample is restricted to individuals in the Immediate Incentive treatment and individuals in the Control group. Individuals in the Cue/Feedback treatment were not included in this analysis. All estimates except those in 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 10am 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 includes 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). Column 4 does not replace missing nights with baseline data as in our main analysis. 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.5: Immediate Incentives and sleep: By quartiles of sleep at baseline

	(1)	(2)	(3)	(4)
	Q1	Q2	Q3	Q4
	Sleep≥7			
Treatment	0.1495***	0.1111***	0.1507***	0.0977***
	(0.023)	(0.024)	(0.022)	(0.026)
Post-Treatment	0.0952***	0.0367	0.0410	0.0278
	(0.021)	(0.028)	(0.029)	(0.030)
Observations	11,999	11,907	8,763	8,729
Mean of dep. var.	0.116	0.320	0.516	0.791
Std. dev.	0.320	0.467	0.500	0.407
Number of individuals	213	211	206	210
	Sleep hours			
Treatment	0.3569***	0.3469***	0.4110***	0.2763***
	(0.072)	(0.077)	(0.070)	(0.083)
Post-Treatment	0.2075***	0.1928**	0.2096**	0.0820
	(0.072)	(0.086)	(0.092)	(0.090)
Observations	11,999	8,906	8,763	8,729
Mean of dep. var.	5.773	6.394	6.919	7.627
Std. dev.	1.378	1.444	1.466	1.150
Number of individuals	213	211	206	210

Notes: The sample is restricted to individuals in the Immediate Incentive 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 10am 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.6: Immediate incentives and sleep: Additional outcomes

	(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	0.0022	0.0143	0.0096	0.0687***	0.2062***
	(0.004)	(0.015)	(0.016)	(0.011)	(0.031)
Post-Treatment	-0.0050	0.0481***	0.0458***	0.0409***	0.0940***
	(0.004)	(0.016)	(0.016)	(0.012)	(0.031)
Observations	69,937	18,100	22,948	69,937	69,937
Mean of dep. var.	0.0513	0.478	0.478	0.568	7.213
Std. dev.	0.188	0.500	0.500	0.495	1.425
Number of individuals	840	840	840	840	840
	Sleep ≥ 6	Sleep 7-9	Efficiency	REM sleep	Deep sleep
Treatment	0.0805***	0.1099***	0.2497*	2.4852***	0.4040
	(0.010)	(0.011)	(0.143)	(0.769)	(0.648)
Post-Treatment	0.0368***	0.0491***	0.0694	1.7137**	0.1183
	(0.010)	(0.012)	(0.182)	(0.872)	(0.695)
Observations	46,989	46,989	46,989	43,168	43,168
Mean of dep. var.	0.714	0.382	93.52	84.12	74.52
Std. dev.	0.376	0.402	5.170	28.32	22.01
Number of individuals	840	840	840	798	798

Notes: The sample is restricted to individuals in the Immediate Incentive 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 10am 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.7: 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 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). 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.

Table A.8: 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	0.0951***	0.1312***	0.1534***	0.1119***	0.1129***	0.1275***
	(0.025)	(0.015)	(0.031)	(0.014)	(0.019)	(0.017)
Post-Treatment	0.0642**	0.0549***	0.1272***	0.0450***	0.0489**	0.0642***
	(0.027)	(0.018)	(0.036)	(0.017)	(0.024)	(0.020)
Observations	12,848	33,909	8,120	38,869	19,827	26,930
Mean of dep. var.	0.344	0.463	0.424	0.430	0.475	0.397
Std. dev.	0.475	0.499	0.494	0.495	0.499	0.489
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 Incentive 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 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.

Table A.9: 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. Vvar.	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 Incentive 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 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.

Table A.10: Immediate 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	10.1637***	0.0689***	2.9172	-2.3196	-6.4537	1.7756	1.3890
	(3.701)	(0.016)	(6.492)	(4.454)	(4.960)	(2.013)	(1.628)
Post-Treatment	17.7161***	0.0748***	2.9074	-7.4170	-8.5724	-4.9861**	-3.0683
	(4.708)	(0.019)	(7.764)	(5.550)	(6.079)	(2.194)	(1.881)
Observations	5,704	5,704	5,704	5,704	5,704	5,704	5,704
Mean of dep. var.	491.5	0.733	323.4	101.5	92.56	95.47	22.85
Std. dev.	102.4	0.443	191.7	123	144.2	51	42.98
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	-1.8404	1.0106	-19.3525***	-12.1082***	-6.4515**	-0.3582	-6.5624
	(1.305)	(1.465)	(4.554)	(3.005)	(2.993)	(1.977)	(6.156)
Post-Treatment	-1.7483	-2.3535	-5.8788	-3.2536	-4.3923	2.1519	-4.9530
	(1.388)	(1.881)	(5.717)	(3.553)	(3.525)	(2.825)	(7.685)
Observations	5,704	5,704	5,704	5,704	5,704	5,704	5,704
Mean of dep. var.	18.25	54.29	171.6	70.37	79.17	22.11	455.5
Std. dev.	36.95	40.25	136.7	91.96	90.27	63.70	171
Number of individuals	836	836	836	836	836	836	836

Notes:
The sample is restricted to individuals in any of the cash incentive treatments and individuals in the Control group. Individuals in the Cue/Feedback treatment were not included in this analysis. All the estimates include controls for month fixed effects, indicators for the number of classes starting before 10am, 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 10am, 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.

Table A.11: Immediate Incentives, cognitive performance and physical health

	(1) Correct math answer	(2) Creativity score	(3) RHR	(4) # steps	(5) Active minutes
Treatment	0.0024	0.0025	-0.2529	35.7563	-1.0652
	(0.023)	(0.057)	(0.186)	(154.475)	(3.866)
Post-Treatment	-0.0308	0.0002	-0.1902	-73.3478	-7.5864
	(0.032)	(0.059)	(0.205)	(237.144)	(6.125)
Observations	3,181	3,243	46,542	46,989	46,989
Mean of dep. var.	0.363	3.307	65.67	7161	191.9
Std. dev.	0.481	0.717	8.314	5756	140.8
Number of individuals	809	803	832	840	840

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 (RHR). 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 Incentive 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 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. 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: 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 - Column 4 reports average treatment effects with standard errors in parentheses. 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. When per-person treatment costs were not reported, they were calculated by dividing overall program costs by the ITT treatment group size.

Table A.12: Review of post-secondary interventions (continued)

Paper	Treatment	Setting	Findings	GPA & Costs
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
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[1] Costs: Bottom of page 323[2]
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
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

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

C. Instructions and Experimental Material

Immediate Incentives

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

HOW IT WORKS
Every morning, you will receive feedback on whetehr 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 9am. 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

Cost-Effectiveness Analyisis

To ensure comparability with our results, GPAs reported in Figure 6 are non-cumulative, either at the semester or year-level. All coefficients included in the figure are OLS estimates of program impacts. Whenever applicable, we use authors' baseline estimate for program impacts. To calculate per subject per semester costs, we divide total program costs by the ITT sample size of the program. See Table A.12 for more information on each study included.

App ScreenShots

Figure C.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 1am. 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”m they could specify a behavior of their choice.

Figure C.2: App Screenshots - Immediate Incentive Treatment

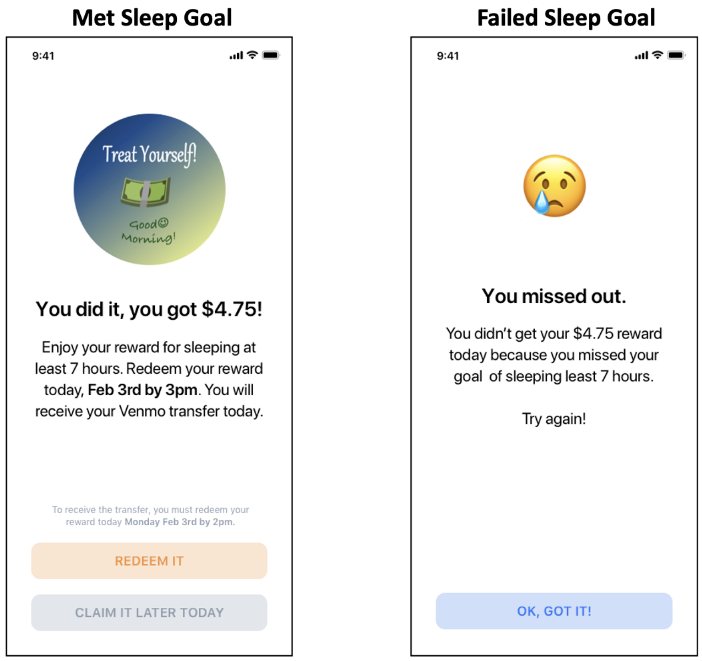


Figure C.3: App Screenshots - Delayed Incentive Treatment

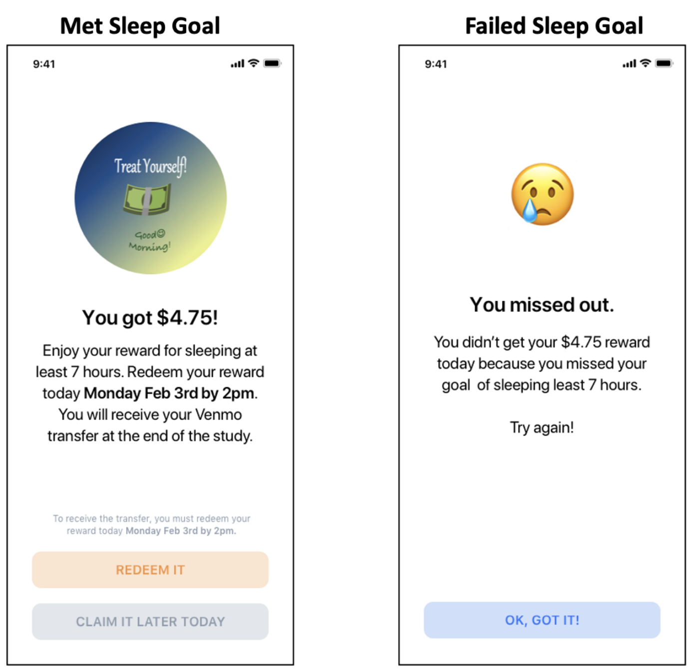


Figure C.4: App Screenshots - Cue/Feedback Treatment

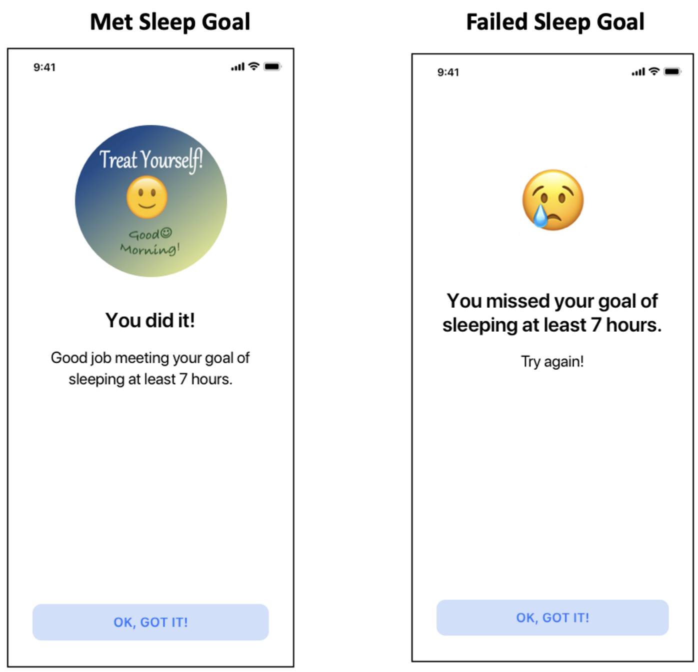


Figure C.5: Reminder to Sync - All Treatments