Encouraging Digital Wellness at Scale: Experimental Evidence From 13 Million Social Media Users

Peter Hickman
Harvard University

Ruru Hoong
Harvard University

Yulu Tang
Harvard University

This version: January 7, 2025
Click here for latest version

Abstract

We document the demand for, and effectiveness of, tools aimed at supporting user's digital wellness through a large-scale natural field experiment orders of magnitude larger than any previous study. We partner with a leading social media platform to prompt 13 million active social media users to set Digital Wellness alarms—*soft commitment devices* in the literature—and estimate a 4.8% take-up rate among this broad population. We consider and present evidence for why take-up may differ in this study from earlier small-scale studies in the literature: the prevalence of perceived self-control problems, the effectiveness of the alarm, and sensitivity to the timing and nature of the invitation to take up commitment. Moreover, we find evidence that higher demand in earlier studies is a consequence of lab experiments selecting participants with higher-than-average demand for commitment.

*Contact: ruruhoong@g.harvard.edu. We thank Hunt Allcott, John Beshears, Katie Coffman, David Laibson, Christine Exley, Matthew Gentzkow, Gautam Rao, Jesse Shapiro, Winnie Van Dijk, and workshop participants in the Harvard Behavioral Lunch for helpful comments. This study was approved by the Institutional Review Board at Harvard (IRB21-1291). A pre-analysis plan can be found on the AEA RCT Registry (Trial #9143). Hickman gratefully acknowledges support from the NSF Graduate Research Fellowship Program.

1 Introduction

Public discussion over how to support digital wellness in an age of digital devices and social media is widespread, accompanied by ongoing debate over who is best positioned to address these challenges. Many companies have taken steps to provide users with tools that support their digital well-being. Social media platforms like Facebook and Instagram have introduced in-app features such as "Your Time" which help users monitor and control the time they spend on these platforms. Third-party apps like Freedom enable users to set "Focus" to minimize distractions, while phone developers like Apple and Android offer features such as Screen Time and Digital Wellbeing, which allow users to set alarms to remind them when they have reached their preset usage limits. These tools that help people constrain their own choice set in the future—which in the literature have been called *commitment devices*[1]—proliferate in the social media space, often as a way to empower individuals to take greater control over their screen time and foster healthier digital habits.

Yet this presents a puzzle—despite the wide availability of a range of both third-party and in-app commitment devices, they are used by very few in practice.[2] Why do people neglect these tools when given the choice to adopt them? One possibility is that users might not be aware of their existence. Alternatively, even when users know about these tools, the effort required to set up a commitment or download a third-party app may deter adoption. Another explanation is that people might believe they don't need or want such commitments, underestimating their own behavioral tendencies.

Recognizing the importance of understanding how to encourage adoption, we partnered with a leading social media platform to explore how best to promote digital wellness among users. We conducted a large-scale experiment inviting their users to utilize an existing in-app commitment device—hereafter referred to as the Digital Wellness Alarm. In an experiment several orders of magnitude larger than any other study of commitment, we randomly assign a subset of 13 million active users invitations to set up the Digital Wellness Alarm. The invitation was salient and eliminated informational frictions, covering the screen with a

[1] See Bryan et al. (2010) for a review on commitment.

[2] E.g., Allcott et al. (2022) find that only 5% of their recruited sample of participants recruited over Facebook already use apps to limit their own phone use. Natural use of commitment in the field is also low in other contexts, as surveyed in Laibson (2015).

prominent message of encouragement to set up the alarm: "You have watched for a long time. Set an alarm to rest. Set a rest time and begin a healthy lifestyle." Users were required to engage by pressing either "Cancel" or "Set reminder to rest" in order to continue using the platform. The tool was expedient, allowing users to customize their commitment via a simple slider interface, eliminating the need to navigate to separate settings pages and reducing setup frictions. The commitment itself takes the form of an alarm that goes off after a self-set amount of daily usage or at a particular time of day. Upon hitting the alarm, the user's screen blacks out, and they see a message encouraging them to rest. Importantly, the Digital Wellness Alarm maintained user autonomy by being "soft": users can choose to snooze the alarm when it goes off and can update their alarm settings at any time.

Selected participants are randomized into treatment groups that vary (i) when—if at all—in the day they are invited to set up a Digital Wellness alarm and (ii) what setting the suggested Digital Wellness alarm is defaulted to. After randomizing individuals into their experimental treatments, we then use the administrative data available from the platform to study a host of measurement outcomes, including participants' usage and uninstallment patterns. We complement our analysis with a survey—administered to 63,459 participants within the platform through two questions that pop up on the screen—that asks about their predicted and ideal usage for the following week.

We first find that adoption of digital wellness tools remains a challenge, and is limited when scaling beyond controlled environments. While laboratory studies often report high uptake rates for commitment devices, our study reveals a more complex reality when engaging a diverse, real-world user base. In our sample of over 13 million users, just 0.5% already had an alarm implemented, despite the fact that the functionality had been available in their settings for half a year. However, upon introducing invitations to set up alarms, adoption rose modestly: 2.0% opt to do so upon first invitation, with eventual take-up reaching 4.8% after repeated nudges over a two-week experimental period for those users who had previously declined the invitation. Contingent on seeing an invitation, treatment users are invited an average of 5 times over the two-week experimental period. Interestingly, slight differences emerged among users with varying app usage patterns: those who had uninstalled the app at least once in the prior six months (around 50% of the 13 million users) showed marginally higher adoption rates (4.9% for those who have uninstalled vs. 4.6% for those who have not).

The observed adoption rates might appear modest when viewed in the context of previous experiments that find high demand for commitment. Previous literature prompting social media users to adopt online commitment devices exhibit higher take-up rates ranging from 25% to over 75%, despite the fact that they often involve higher set-up costs, like downloading a third-party app or taking multiple steps to set up the functionality.[3] However, those studies often rely on voluntary participation through targeted recruitment, such as Facebook ads or Amazon Mechanical Turk (MTurk), which may naturally select for participants with a preexisting interest in commitment devices. Within our study, users who completed a brief two-question survey–a subset likely more engaged or self-aware—exhibited an 8% take-up rate, over 70% higher than the broader experimental population. This suggests that selection effects may play a significant role in adoption outcomes.

How, then, does the population in our study compare to that of previous experiments? It is hard to compare them directly: our users are based outside the US and are using a highly dynamic, video-focused social media platform that differs fundamentally from other multimedia social networks previously studied. Perceived self-control challenges, which can be a driving force behind demand for commitment devices, are also less prevalent in our sample: only 49% of surveyed users reported self-control concerns, compared to over 90% in studies of other platforms.[4] This suggests either that previous studies might be disproportionately selecting respondents with higher perceived self-control problems, or people might have lower perceived self-control issues on this particular platform than in other experimental settings.

To better understand the adoption patterns, we examine three key factors: (i) the prevalence of perceived self-control issues, (ii) the effectiveness of the alarm, and (iii) sensitivity to features of the choice environment, such as timing and defaults. First, as seen from the subset of participants that complete our two-question survey, about 49% of users exhibit *perceived* self-control problems; that is, their predicted daily usage is higher than their ideal usage. This perception is critical, as those who do not view their usage as excessive are

[3] E.g., Acland and Chow 2018; Hoong 2021; Allcott et al. 2022.

[4] E.g., Hoong 2021.

unlikely to see the need for commitment tools.

Second, the treatment-on-treated effects of our intervention are modest—for participants that adopted commitment, the alarm reduced their average daily use in the week after the experimental period by 5 minutes, or on average about 9% of daily usage.[5] The intent-to-treat effect was smaller, at just 0.2 minutes per day, reflecting the low overall take-up. These results indicate that while the tool can have meaningful impacts for adopters, the modest effect size may partially influence perceptions of its necessity. This is consistent with previous findings that the "hardness" choices of commitment contracts matter a lot towards take-up, though harder commitment is often hard for platforms to offer due to desires to respect user autonomy and flexibility.[6]

Third, we also find that those who receive an invitation after spending more time on the platform that day are less likely to take up the treatment, suggesting that invitation timing matters. Users are in the midst of watching videos on the platform when they are invited to set up alarms. Users who received invitations earlier in their daily usage (e.g., after 20 minutes) were more likely to set up alarms than those nudged later (e.g., after 120 minutes), with take-up rates of 5.8% versus 3.5%, respectively.[7]

However, the difference in take-up based on invitation timing is unlikely to explain the low take-up overall; we suggest that the first two reasons hold more credence. The former likely explains the majority of the lack of demand for commitment—51% of the population do not have perceived self-control problems. The effectiveness of the offered device also likely matters, though we are unable to quantify to what degree as we do not vary commitment effectiveness, nor do we elicit ex-ante perceived effectiveness of the device.

We then present a model of quasi-hyperbolic discounting agents to unify our findings and provide deeper insights into user behavior. This model highlights how (i) the absence of perceived self-control issues, (ii) the effectiveness of commitment, and (iii) the timing of invitations contribute to variations in take-up. Importantly, it also explains why more

[5]Because different participants potentially adopt treatment on different days, the "experimental period" can correspond to 1-21 days after first adoption, depending on the user.

[6]In the context of retirement savings, there is some evidence that higher penalties attract more commitment account deposits, e.g., Beshears et al. (2020).

[7]There are multiple reasons for this discrepancy, including selection of higher usage users into actual treatment adoption in the latter treatment group. We discuss this in the results section.

sophisticated users—who are more aware of their self-control challenges—may exhibit lower take-up compared to naifs, and why individuals who either overestimate their usage or fall short of ideal usage benchmarks may show greater engagement with these tools.

Understanding how individuals engage with their digital habits is essential for designing tools that help them better align their online activities with their personal goals. While much of the economic literature has explored time-sensitive decision-making in contexts like retirement savings or health behaviors,[8] our study shifts the focus to the unique and dynamic environment of social media, specifically on one of the world's largest and most engaging platforms. By examining how users interact with platform-integrated features aimed at promoting mindful engagement, we offer new insights into the interplay between user behavior and digital wellness tools.

Our contributions are three-fold: first, we extend the literature of experimental papers studying the impacts of social media use on a range of outcomes, including subjective well-being, work productivity, political engagement, and more.[9] Unlike much of the earlier work that focuses on correlations between social media habits and various characteristics,[10] our study examines causal impacts in a naturalistic setting. By analyzing a sample of 13 million users, we shed light on how platform-integrated tools influence behavior at scale, offering a broader and more nuanced perspective than previous studies that typically examine smaller, more targeted populations.

Second, we contribute to the growing body of research on the effects of digital wellness

[8]E.g., Thaler and Shefrin 1981; Becker and Murphy 1988; Laibson et al. 1998; Thaler and Benartzi 2004; Kan 2007; Giné et al. 2010; see Ericson and Laibson 2019 for a review.

[9]E.g., Sagioglou and Greitemeyer 2014; Tromholt 2016; Hunt et al. 2018; Vanman et al. 2018; Mosquera et al. 2020; Allcott et al. 2020; Collis and Eggers 2022. It is also worth mentioning a large preceding psychology literature that documents social media use (e.g, Griffiths 2005; Lenhart and Madden 2007; Ko et al. 2006)—ranging from psychological measurements to assess and predict habit formation (Bianchi and Phillips 2005; Andreassen et al. 2012; Andreassen et al. 2013; Schou Andreassen and Pallesen 2014) to studies investigating the relationship of social media use and welfare outcomes (Pelling and White 2009; Karaiskos et al. 2010; Wilson et al. 2010; Kuss and Griffiths 2011; Elphinston and Noller 2011; Koc and Gulyagci 2013; Hong et al. 2014; Ryan et al. 2014; Vanman et al. 2018).

[10]More recently, Guo (2022) uses adoption of high-speed internet as an instrument for increased social media use to show that visual social media increases teen girls' severe mental health conditions relative to teen boys'.

tools, particularly those designed to help users align their platform use with personal goals.[11] These are most related in that they run experimental studies of such tools in the realm of smartphone and social media use.Existing studies in this space often rely on experimental recruitment methods, such as targeted advertisements that advertise researcher-designed interventions. In contrast, our study evaluates the adoption and effectiveness of a commitment tool embedded within the social media platform, providing insights into user behavior in a naturalistic and large-scale environment—13 million users is a far larger population of users than in all of these studies.

Third, we contribute to the literature on demand for commitment and the factors that contribute to it. One large reason for take-up in prior studies could be experiment selection and experimenter demand effects. Unlike in previous studies, we offer commitment directly in the capacity of the platform, minimizing the likelihood that our participants would have strong beliefs that the "experimenter" desires to reduce their usage.[12] This approach allows us to observe genuine engagement patterns without participants feeling external pressure to adopt the tool. Interestingly, the modest adoption rates suggest that, in a naturalistic setting, demand for commitment tools is nuanced and context-dependent, often leading to lower adoption rates.[13]

The rest of the paper proceeds as follows: Section 2 details experiment design. Section 3 lays out our empirical strategy, and Section 4 presents results. Section 5 presents a model to unify these findings, and Section 6 concludes.

[11] E.g., Marotta and Acquisti 2017; Acland and Chow 2018; Hoong 2021; Allcott et al. 2022. There is also a large human–computer interaction literature that looks at the effects of different interventions on behavior change, including Klasnja et al. 2011; Lee et al. 2014; Lee et al. 2014; Ko et al. 2015; Choi et al. 2016; Okeke et al. 2018; Kim et al. 2017; Kim et al. 2019; Genç and Coskun 2020.

[12] For discussion of experimenter demand effects, see e.g., De Quidt et al. (2018).

[13] Additionally, we corroborate findings in previous studies—e.g., Carrera et al. (2022)—that more sophisticated users are less likely to take up commitment. This is related to the set of studies that have documented the correlations between demand for commitment and indicators of time inconsistency—ranging from positive (Augenblick et al. 2015; Kaur et al. 2015) to weak or negative associations (Ashraf et al. 2006; Sadoff et al. 2020; John 2020; Carrera et al. 2022), with the last study finding that an exogenous shock to sophistication reduces demand for commitment contracts.

2 Experimental Design

Our study is entirely administered within the video-sharing social media app, with the experimental period spanning two weeks between July 6th and July 20th, 2022. We partner closely with the platform's engineering team to design the experiment—which they thereafter implement—and use administrative data on the platform behavior prior to and subsequent to the experimental period to analyze the effects of our interventions.

2.1 Administrative Data

The platform stores a rich dataset detailing each user's demographics and history of behavior on the platform. In this analysis, we focus on the following variables.

Demographics. We predominantly use participants' gender, age, and locational characteristics to conduct balance checks and heterogeneity analyses. Gender and age (binned into 18-23, 24-29, 30-39, and 40-49 years old) are predicted rather than self-reported, although the platform's internal validation has found thse predictions to have over 80% accuracy. The platform also directly collects locational information and classifies users as rural or urban, along with whether they are in a tier 1 city.

Behavior. Our main outcome of interest is daily usage—the number of minutes spent on the platform each day. We observe this before, during and after the experiment. We also track which days users are active, their uninstallment and reinstallment behavior, and the number of videos they viewed on any given day. Finally, we observe if and when participants set any Digital Wellness Alarm features.

Experimental invitations. We track how many times the participant sees an invitation to use the Digital Wellness alarm during the experiment, when they see those invitations, whether they ever opt in, and the specific parameters (e.g., daytime vs. nighttime) for the alarm they set.

2.2 Participant Selection

Prior to July 6, we select 13,048,771 users aged above 18 years on the platform to participate in our study. Our focus is on active users, so we require an average daily duration of at least one hour of use over the two weeks before the experiment (June 20 to July 4). Furthermore, we require that half of the participants must have uninstalled the app at least once in the 6 months before July 4. This requirement stems from our interest in uninstallment as a potential expression of users attempting to moderate their own usage. Of course, these users necessarily reinstalled the platform, thus meeting the requirements to be included in our sample.

2.3 Invitation to Set Up Commitment

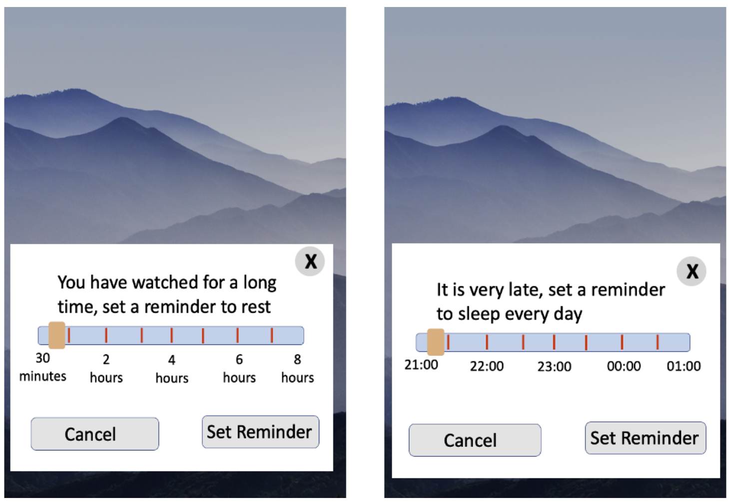
We randomize participants into a control group that is never shown an invitation to set up any alarm, and a treatment group that sees some variation of an invitation to set up future alarms. Treatment group participants see one of two versions of the invitation in Figure 1:

- **Daytime version:** Appearing before 9pm, this version invites users to set the Digital Wellness alarm for a chosen number of hours into their future daily use.
- **Nighttime version:** Appearing after 9pm, this version reminds users to rest by a chosen time at night.

In both versions, a message appears in the bottom half of the screen containing a slider to set the alarm, and options to confirm or cancel. In the daytime version, participants can select from 1, 2, 4, 6, or 8 hours. The nighttime version allows participants to select a reminder time of 9pm, 10pm, 11pm, 12am, or 1am.

There are 6 treatment sub-groups in a 3 $\times$ 2 factorial design, as seen in Figure 2. The first element of randomization regards timing: users are assigned to see the invitations at 20, 60, or 120 minutes into their daily use. Note that due to imprecision in real-time estimates of participant use on the platform's end, the invitation ends up occurring within fairly coarse bounds as opposed to a very precise time—between 0-45 minutes for the first group, 45-90 minutes for the second group, or over 90 minutes for the last group. This

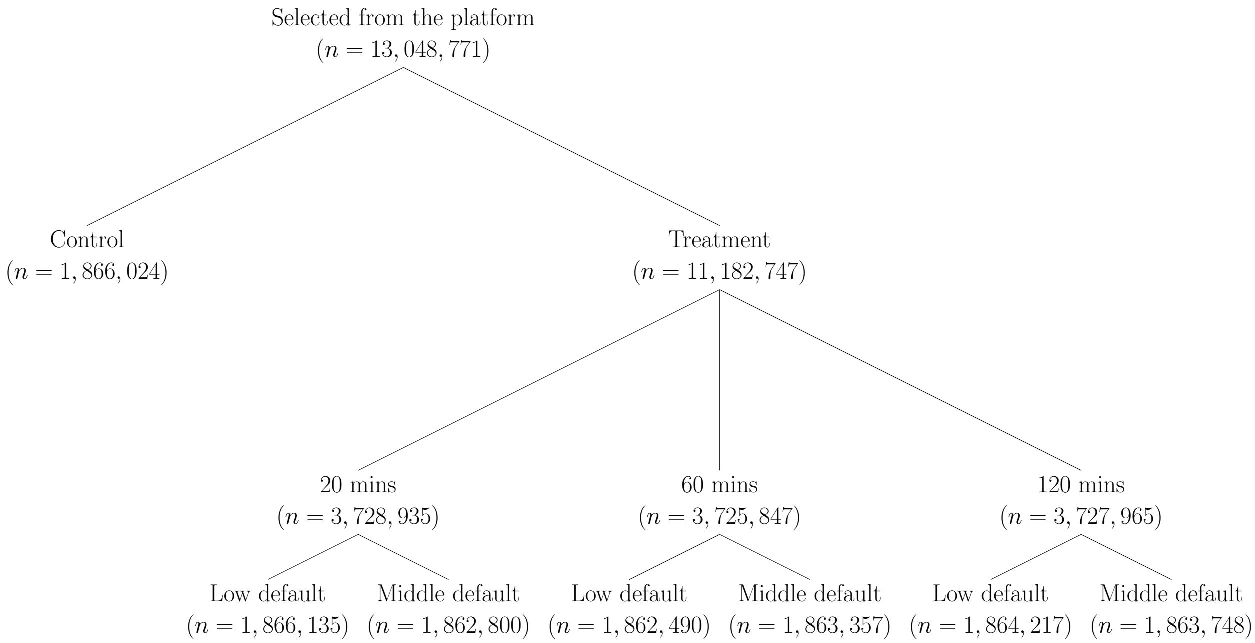
Figure 1: Pop-Ups (Invitation to Take Up)



Notes: This figure presents a mock-up of the pop-up message inviting users to take up the commitment device. There were two versions of the invitation. The left image shows the "daytime" version of the invitation, which was shown to users who received the invitation before 9pm. The right image shows the "nighttime" version for users who received invitation after 9pm. In the "daytime" version, users are invited to set a reminder to rest and can choose from among several discrete points on the slider. In the "nighttime" version, users are invited to set a reminder to sleep at certain clock time.

imprecision results in a very small number of participants not seeing any invitation despite being assigned to treatment; we discuss the implications of this later in the results section. Second, we randomize the default suggested alarm to either the tightest or middle limit available. This corresponds to 1 or 4 hours of use in the "daytime" version, and 9pm or 11pm in the "nighttime" version.

Figure 2: Randomization of Participants Into Treatment Groups

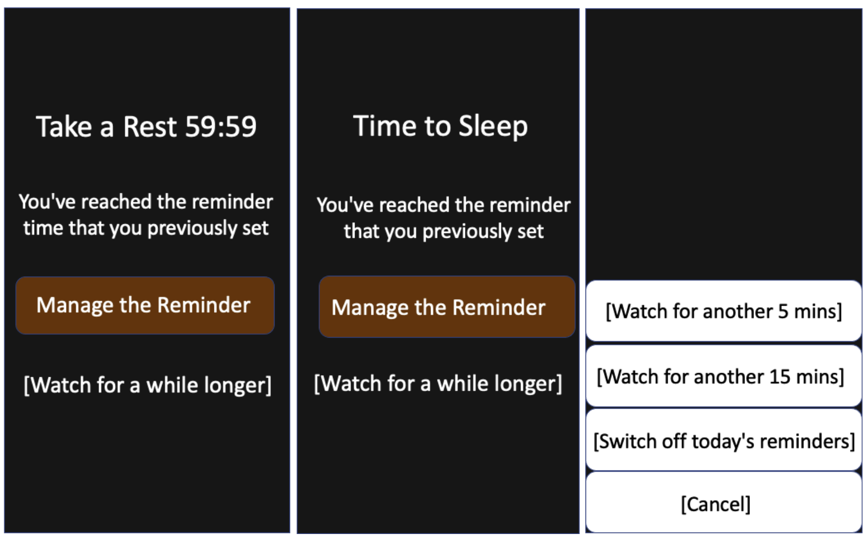


Notes: This figure illustrates the randomization procedure. Over 13 million users were selected as the experimental sample. Nearly 2 million were allocated to a control group and the rest to a treatment. Treatment users were cross-randomized into 2 default groups and 3 treatment timing groups of roughly equal size.

they see the interface shown in Figure 3 whenever they reach their self-selected usage limit. Once the alarm is triggered, the phone transitions into a gentle, full-screen prompt encouraging users to take a break. For the daytime version, it reads, "Rest a while. You've reached the reminder time that you previously set." For the nighttime version, the message is "Time to sleep. You've reached the sleep reminder that you previously set."

In both scenarios, users can click on "Manage reminder timings" to edit their alarm or "Watch for a while longer" if they feel like continuing. If they choose to continue viewing videos, another menu appears with four options: (i) "Watch for another 5 minutes," (ii) "Watch for another 15 minutes", (iii) "Switch off today's reminders", or (iv) "Cancel" (to return to the previous screen). Both control and treatment participants could adjust their alarm settings at any time during the experiment. This design underscores the platform's desire to ensure that while users are reminded about their intended limits, they retain flexibility and control over how they balance app engagement with personal well-being.

Figure 3: Alarm Features



Notes: This figure presents a mock-up of the Digital Wellness alarm in action, after the user has hit their chosen usage limit or sleep time. The left image shows the "daytime" version of the alarm, informing users they reached the time they previously set. The right image shows the "nighttime" version. The far right image shows the options avaiable if users click "watch for a while longer": users could watch for 5 or 15 more minutes before getting another alarm, or they could switch off the reminders for the day.

Columns (2) and (3) of Table 1 show the balance between the control group and participants that were assigned to any treatment at baseline. Our participants spend an average of 2.5 hours on the platform daily, average around 33 years in age, and are 55% male, and are balanced on all of these characteristics between control and treatment. Participants are slightly older, more active, and slightly more male than the platform as a whole, which makes sense as we purposely selected participants who were recently active and excluded any users below the age of 18 for ethical purposes.

Table 1: Balance Table

	(1) Whole Platform	(2) Control	(3) Any Treatment	(4) p-val	(5) sd
Percent Male	51.4	54.7	54.7	.376	-
Age	24.5	32.5	32.5	.765	8.6
Percent Urban	66.8	66.4	66.4	.735	-
Percent Tier 1 City	25.8	26.4	26.3	.65	-
Price of Phone (USD)	503.9	519.5	519.3	.52	337.8
Percent of Days Active (Prior 28 Days)	85.4	91.4	91.4	.606	22.3
Average Daily Minutes (Prior 28 Days)	122.9	152.1	152.2	.328	90.4
Days Active Since Signup	439.2	417.0	417.1	.484	324.8
Percent Ever Uninstalled App	4.1	54.1	54.1	.771	-
Mean Uninstalls	0.05	1.3	1.3	.944	3.4

Notes: Column (1) reports means for the population of the platform as a whole. Columns (2) and (3) report means for control and treatment groups. Column (4) reports the t-test p-value between control and treatment in our sample, and Column (5) reports the standard deviation.

2.4 Additional Survey Data

To complement the administrative dataset, we invite a subset of participants to take an in-app survey about their ideal use and predicted use of the platform. We obtained permission from the platform to send two questions to roughly 800,000 participants at the start of the experimental period.[14] These questions popped up on their screen whilst they were using the platform and were designed to be multiple-choice so as to increase the probability of participants responding. The two survey questions were: "In the next 7 days, how much time would you predict you would want to spend on the platform per day?" and "In the next 7 days, how much time would you ideally want to spend on the platform per day?" Each question had the same five discrete options: 0-30 minutes, 30-60 minutes, 1-2 hours, 2-3 hours, and 3 or more hours.

[14]Though we scheduled the survey to take place at the start of the experimental period, because of delayed implementation for some users, some of the treatment participants answered the survey after they had received an invite to set up an alarm. This leaves open the possibility that the survey answers were influenced by the treatment. We check for this by estimating the effect of treatment on survey answers, and do not detect any effects of the treatment on ideal or predicted use, i.e., the control group and treatment group have statistically identical distributions of survey answers. This suggests some confidence that the survey answers are roughly similar to what we would have seen if we had collected them strictly before the experiment began. Moreover, we double check that the reverse effect is not present: being invited to respond to the survey does not affect behavior on the platform. Note also that a balance table comparing control and treatment baseline characteristics for only those surveyed can be found in Table 7 of the Appendix.

Participation in the survey was optional: users could simply swipe away if they chose not to respond. Our survey achieved an 8% response rate, generating 63,459 observations of predicted and ideal daily usage. This participation rate is fairly typical for short in-app surveys on the platform.

Table 2 reveals selection into response: those who answered tend to be younger, more likely male, more urban, and more frequent users compared to those who did not respond. They are also more likely to have uninstalled the app in the past—potentially reflecting a stronger interest in managing their usage or exploring new platform features.

Table 2: Comparison of Baseline Characteristics between Surveyed and Non-Surveyed Users

	(1) No Survey	(2) Survey	(3) p-val
Percent Male	54.7	58.3	0.000
Age	32.5	27.7	0.000
Percent Urban	66.4	67.8	0.000
Percent Tier 1 City	26.3	28.4	0.000
Price of Phone (USD)	519.1	566.1	0.000
Percent of Days Active 28 Days Before Experiment	91.4	92.0	0.000
Daily Minutes 28 Days Before Experiment	152.1	168.5	0.000
Days Active Since Signup	416.6	514.6	0.000
Percent Ever Uninstalled App	54.1	57.0	0.000
Mean Uninstalls	1.3	1.4	0.329
n (Number of Participants)	12,985,312	63,459	-

Notes: Columns (1) and (2) report means for surveyed and non-surveyed users, respectively, and Column (3) shows the t-test p-value for differences.

We categorize survey respondents as having *perceived* self-control issues if their predicted daily usage surpasses their ideal usage. Roughly 49% of respondents meet this definition. As shown in Figure 12, the median predicted use (over 3 hours) is notably higher than the median ideal use (1–2 hours), highlighting that some users perceive a significant gap between how much they'd *like* to spend on the platform and how much they believe they *will* watch.

To further understand actual behavior, we see if respondents' real usage in the next 7 days exceeds the ideal usage they reported. Around 56% of surveyed participants end up using the platform more than they had said was ideal, as seen in Table 13 of the Appendix.

Of those, 53% *underestimate* their future time spent (actual > predicted), while 31% predict it accurately, and the remaining 16% *overestimate* their future use. This distribution shows the varying degrees to which personal expectations, preferences, and real-world behaviors align or diverge on the platform.

3 Empirical Strategy

Our main causal estimate of interest is the first stage: how invitations affect take-up of commitment. We estimate:

$$X_{it} = \alpha_1 T_i \times post_t + \lambda_i + \delta_t + \varepsilon_{it} \tag{1}$$

where $X_{it} \in \{0, 1\}$ is an indicator for whether participant i set the alarm at time t. $T_i \in \{0, 1\}$ indicates if user i is assigned to any treatment group, and $post_t \in \{0, 1\}$ indicates that period t occurs on or after the day the experiment begins. λ_i are participant fixed effects, δ_t are time fixed effects, and α_1 is the coefficient of interest. In the main specification, we define $X_{it} = 1$ if a participant accepted an invitation at any point during the 2-week experimental period, and zero otherwise. Note that the 0.5% of users in the control group that already had existing app limits (from prior to the beginning of the experiment) retained their treatment assignment value of 0—a robustness check assigning a treatment value of 1 to participants that already have an app limit prior to experiment can be found in Table 14 of the Appendix.

The second stage concerns how this impacts a host of behaviors on the platform: time spent on the platform, days active on the platform, and uninstallment probabilities. We organize our results into intent-to-treat (ITT) effects and treatment-on-treated (TOT) effects. The ITT effect measures the effect of sending invitations on app usage behaviors, while the TOT effect estimates the effect of sending invitations on app usage behaviors *among the participants who took up the alarm*, a group generally termed "compliers" in the literature (e.g., Angrist and Pischke 2009).

For the ITT effects, our specification is:

$$Y_{it} = \beta_1 T_i \times post_t + \mu_i + \zeta_t + \eta_{it} \tag{2}$$

This is analogous to Equation (1) but sets Y_{it} as the outcome variable; we focus on the average daily usage over the period of a week as our main outcome. For the TOT effects, we estimate:

$$Y_{it} = \gamma_1 \hat{X}_{it} + \theta_i + \tau_t + \nu_{it} \tag{3}$$

where $\hat{X}_{it}$ are the predictions from Equation (1). Our identification is established by the random assignment of T_i. We leverage the time and user fixed effects to increase precision. Individual fixed effects absorb any time-invariant variable, so we omit control variables that are meant to capture differences in demographic characteristics. All standard errors are clustered at the individual level to account for intra-person correlation in daily platform time usage.

3.1 Treatment Sub-Groups

In our main analysis, we pool all 6 treatment sub-groups together and estimate the aggregate effect. Here, we estimate the effects of each treatment sub-group separately. For example, the ITT is specified by:

$$Y_{it} = \sum_v \beta_v T_i^v \times post_t + \kappa_i + \psi_t + \epsilon_{it} \tag{4}$$

where β_v is the ITT effect of treatment assignment to sub-group v and $T_i^v \in \{0, 1\}$ indicates that participant i is in treatment sub-group v. Identification is again established by random assignment of the treatment sub-groups. We can test whether $\beta_v = \beta_w$ to see if treatment sub-groups v and w had the same effect. ITT effects may differ across treatment variants because of either different first stages or different TOT effects. Moreover, TOT effects may differ across treatment variants because the treatment is inherently more effective on the compliers or because it generates a different group of compliers.

3.2 Heterogeneous Treatment Effects (HTEs)

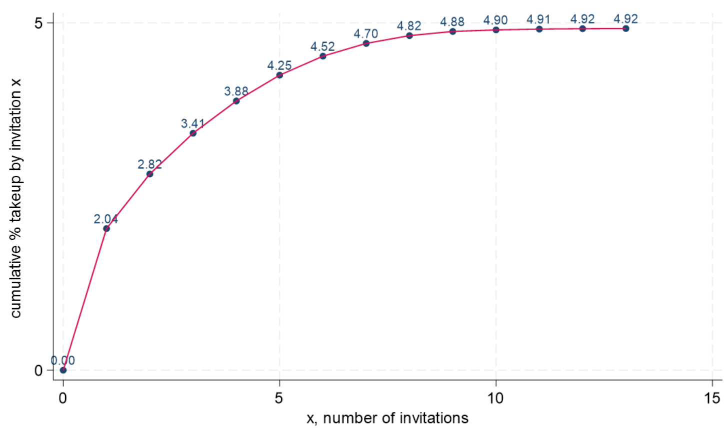
A given treatment may be more effective for some demographic groups than others. Our large sample allows us to segment participants by demographic while retaining enough power to precisely estimate take-up rates. We estimate separate regressions using Equations (1)-(3), splitting the sample by demographic group (e.g., men vs women).

4 Results

4.1 Detailing Take-Up: First Stage Effects of Treatment

Figure 4 shows the cumulative share of participants who adopt the Digital Wellness alarm after each invitation. Overall, 2.0% accept right after their first invitation. Additional reminders further nudge small increments of users to adopt, leading to about 4.9% take-up among those who eventually see at least one invitation.[15]

Figure 4: Take-Up by Invitation Number



Notes: The x-axis is the number of Digital Wellness alarm invitations received by treated users. The y-axis is the cumulative fraction of these treated users who have adopted the alarm at or before that invitation.

The first stage relationship—with adoption value equalling 1 if the user agreed at any point over the 2 week experimental period to set an alarm—can be found in Table 3. Here, we see a slightly lower percentage of individuals taking up commitment: 4.8% of those *assigned to treatment* eventually set up commitment. [16]

[15] Table 8 in the Appendix compares the baseline traits of participants who adopt at different invitation numbers.

[16] Adding controls for demographic characteristics does not diminish the strength of the significant relationship between the instrument and limit adoption; furthermore, it does not change the size of the coefficient on the instrument.

Table 3: First-Stage Between Treatment Assignment and Commitment Adoption

	(1) No Controls	(2) With Controls
Limit Nudge	0.0480***	0.0480***
	(0.000057)	(0.000057)
Female		0.00438***
		(0.000057)
Age		0.00266***
		(0.000058)
Android (vs iOS)		-0.00230***
		(0.000105)
Phone Price		0.000315***
		(0.000037)
Baseline Daily Average Usage (Mins)		-6.22e-08***
		(6.83e-10)
Tier 1 City		-0.000828***
		(0.000065)
Urban		-0.00135***
		(0.000061)
Observations	26,097,542	24,265,550

Notes: Column (1) presents the first-stage relationship between treatment assignment and limit adoption without controls, whilst Column (2) presents the relationship when controls are added. Standard errors in parentheses, $^{***}p < 0.01,^{**} p < 0.05,^{*} p < 0.1$.

The discrepancy between the 4.8% in the first-stage regression and the previous 4.9% take-up rate amongst those who see an invitation can be attributed to the small subset of treatment participants that do not see an invitation. This occurs for two main reasons: (i) the imprecision in real-time estimates of participant use on the platform's end, and (ii) a small but present likelihood of never seeing the invitation for those who use the platform for a shorter duration on any given day, especially for those in the 120 minute treatment group. However, these participants constitute less than 4% of the population.[17] Even if we conservatively estimate that these participants exhibit double the take-up rates than the rest of the treated population, the take-up rates would remain under 5% should they all have been shown an invitation to take-up the treatment. To maintain the randomization of treatment assignment, we keep these participants in our analysis and do not drop them, but it is worth mentioning that their exclusion would—if anything—augment our ITT effects.

Among those who do adopt an alarm, the average daytime limit set is around 2.2 hours, while the average nighttime reminder is about 10:39pm. [18] 98% of adopters choose limits that would have been binding on at least some of their usage in their most recent week, indicating genuine interest in moderating viewing sessions. On average, participants set a limit that would've been binding for half of the week. To give a sense of the stringency of the limits, we define absolute tightness in a manner analogous to the *limit tightness* variable in Allcott et al. (2022)—the number of minutes by which a user's limits would have hypothetically reduced screen time if fully effective, i.e., had they stopped using the platform upon seeing the alarm. The limits are fairly tight on average (around 50 minutes), though the standard deviation is high.[19]

4.2 Why Is Take-Up Low?

We identify three main factors that may explain why $\tilde{5}$% of invited users ultimately opt into the Digital Wellness alarm: (i) perceived self-control, (ii) the effectiveness of the Digital Wellness alarm, and (iii) features of the choice environment that affect demand.

[17]Table 9 of the Appendix shows the share of treatment-assigned participants who saw any invitation over the experimental period, by treatment group.

[18]Figure 8 in the Appendix details the full distribution of the endogenously chosen limits.

[19]Table 10 of the Appendix provides more details on limit tightness.

4.2.1 Perceived Self-Control Problems

Only those who feel they spend more time than intended on the app have a reason to adopt. We show this logic more formally later in a model, but the logic is straightforward: those whose predicted daily usage falls short of their ideal—either due to naivete or time-consistency—do not want commitment as they do not think there is any behavior they want to limit.

Survey data suggest that 49% of those who responded perceive themselves as overshooting their ideal daily use. If our survey is representative, this suggests that 51% of our wider population can be automatically precluded from demanding commitment. Across the broader user population, this proportion might be even higher. For reasons we outline in prior sections, the populations that select into the survey are disproportionately active, have demonstrated uninstallment behavior before, and are more likely to take-up commitment. Our wider population is therefore likely to have even lower perceived self-control problems than our survey statistics suggest.[20]

4.2.2 Effectiveness of Digital Wellness alarm

The Digital Wellness alarm, once adopted, reduces daily usage among compliers by about 5 minutes on average—around 9% of their baseline. Though we do not ask users *ex ante* about how effective they expect the device to be, it is reasonable to assume that perceived effectiveness would be highly correlated with the reality. If users correctly perceive that commitment will only be moderately effective, that could affect their willingness to set it up: again this is more formally established in the predictions of a model later on.

Table 4 reports the full intent-to-treat and treatment-on-treated impacts on usage and uninstallment outcomes, following Equations (2) and (3). Assignment to invitation reduces

[20] Note that perceived self-control problems should be distinguished from overuse and underestimation. We look at heterogeneous take-up by these classifications in Figure 6 and find take-up is lowest for those who overuse the platform and those who correctly estimate their use. Though these facts may be counterintuitive, we later show how they can be consistent with a model of quasi-hyperbolic discounting.

time spent on the platform by about 14 seconds per day, or about 0.4% of daily usage.[21] Even among compliers, adopting alarms had no more than a modest effect on usage. In our treatment-on-treated results, we find that for the participants that adopted any alarm at any point over the experimental period, the alarm was effective in reducing their daily use by 5 minutes, around 9% of average usage.

Disaggregating the treatment effects into the extensive margin (number of days active) and the intensive margin (the extent of usage on any given day), we find that the main effects are on the intensive margin. Most of the reduction in usage comes from reduced use across days, as opposed to reduced the number of days spent on the platform. It is unsurprising that usage is impacted mainly on the intensive margin: the alarm is a tool for users to enjoy shorter sessions, not for blocking out full days of use.

Table 4: Effects on Usage Behavior (All Treatments)

	(1) Daily usage	(2) Log daily usage	(3) Share days active	(4) Daily usage if active	(5) Uninstall prob
ITT	-0.237	-0.004	-0.001	-0.235	0.000
	(0.079)***	(0.001)***	(0.000)**	(0.079)***	(0.000)
TOT	-4.936	-0.086	-0.011	-4.888	0.004
	(1.637)***	(0.027)***	(0.005)**	(1.651)***	(0.004)
User & Time FE	Y	Y	Y	Y	Y
Observations	26,097,542	26,097,542	26,097,542	26,097,542	26,097,542
Users	13,048,771	13,048,771	13,048,771	13,048,771	13,048,771
Control Mean	145.3	4.4	0.9	144.0	0.0
ITT R^2	0.002	0.011	0.007	0.002	0.001

Notes: This table presents ITT and TOT effects. The pre-period is defined as the 7 days before the experiment, i.e. June 29th to July 5th. The post-period is defined as the 7 day period between July 21st to July 27th. Column (1) details results on average daily usage (in minutes) over the 7 days, Column (2) shows this in log time. Column (3) shows results on the share of the 7 days the user was active (the extensive margin), and Column (4) on the average daily usage over the active days (in minutes). Column (5) shows the effect on uninstallment probability over a 2-week post period. Standard errors in parentheses, $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$.

Heterogeneous treatment effects (HTEs). Though the device is limited in effectiveness for most populations, we consider heterogeneous treatment effects by (i) timing of invi-

[21] Figure 10 in the Appendix plots point estimates and confidence intervals from an estimation of Equation (2) that uses a day-level panel. These point estimates are noisier because the outcome variable—a single day's usage—is noisier than a weekly average. Nonetheless, we see a clear pattern of negative ITT effects of around a quarter minute starting a few days into the experiment—after users have had a chance to see the invitation—that persists through the end of the experiment.

tation, (ii) defaults, (iii) baseline usage and uninstallment, (iv) overuse and underestimation issues, and (v) baseline demographics (age, gender, phone, and locational characteristics).

Panels (a) and (b) of Figure 5 show differential results by treatment sub-groups. We interpret the differential treatment effects by invitation timing with caution: we cannot interpret these effects purely as the difference between treating someone at 20 minutes vs. 60 minutes or 120 minutes, since users in these treatment groups have different likelihoods of getting treatment on a given day. In any case, there is no significant difference in any of the ITT or TOT treatment effects by invitation timing. There is also little difference in effectiveness by suggested default.

Panel (c) explores HTEs by baseline usage. We find that light users have 75% *smaller* ITT effects than heavy users, despite being more likely to set the alarm. This is unsurprising given that they have lower baseline usage to begin with. The log specification in the Appendix suggests similar ITT and TOT effects for both groups in percentage terms. Panel (d) finds that users who had ever uninstalled the platform do not show significant differences in device effectiveness.

Figure 6 considers heterogeneity by overuse and underestimation. We cannot establish much about differences in treatment effects, especially given the survey sample is an order of magnitude smaller than the main sample.

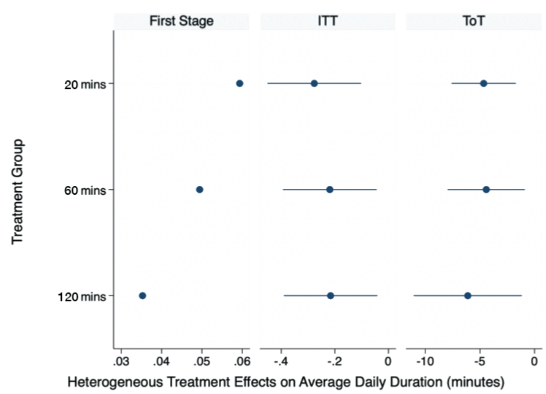
Figure 7 examines other baseline demographics. Though women are over 50% more likely to adopt the alarm than men, their TOT effects are lower, suggesting the device may be less effective for them. We cannot reject the null that the two ITT effects are the same. For age group, phone type, phone price, city tier, and rural/urban, we similarly cannot establish anything about heterogeneous treatment effects.

4.2.3 Features of the Choice Environment: Sensitivity to Timing and Defaults

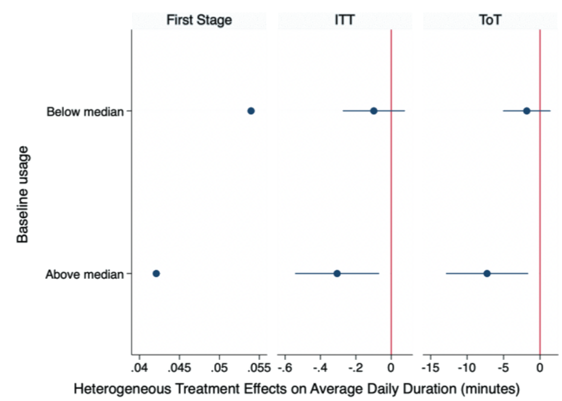
Looking closely at the invitation process, participants invited earlier in their sessions are more likely to adopt. Panel (a) of Figure 5 shows that take-up is highest for the 20 minute treatment group, followed by the 60 minute treatment group, and is lowest for the 120 minute treatment group. This is consistent with the idea that once people are deeply absorbed in content, they may be less open to setting future alarms. If we assume that this monotone

Figure 5: Heterogeneous Treatment Effects by Sub-Treatment Groups, Baseline Usage and Uninstallment

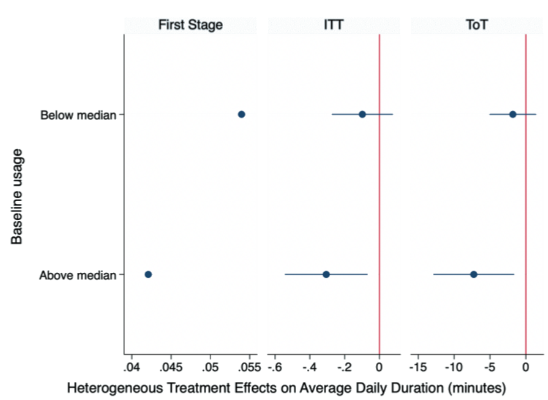
(a) Timing Treatment Groups



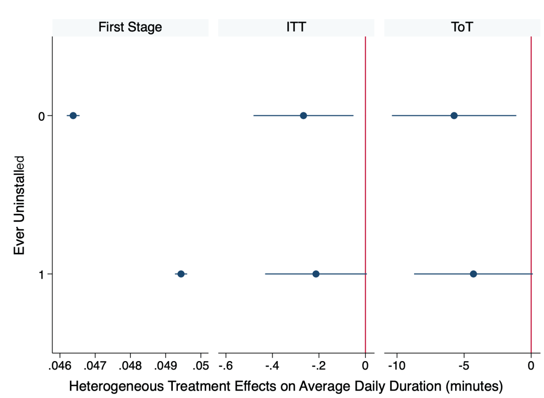
(b) Default Treatment Groups



(c) Baseline Usage (average Daily Use Over Pre-Period Week)



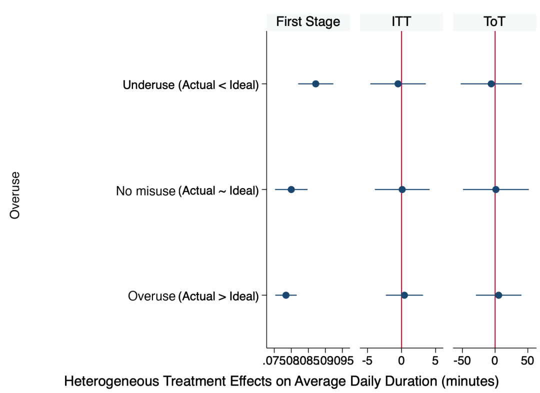
(d) Baseline Uninstallment (ever Uninstalled in History of Using the Platform Prior to Experiment)



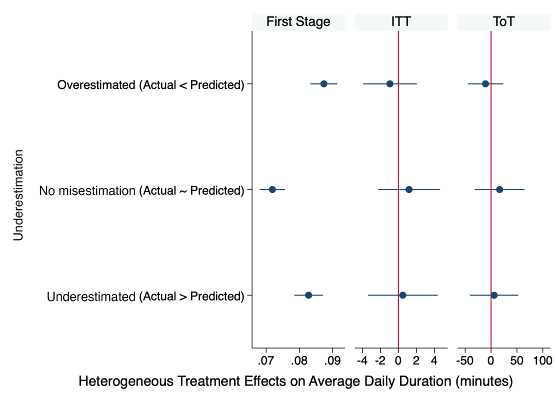
Notes: This figure plots point estimates and 95% confidence intervals from first stage, ITT, and TOT regressions. In each panel, the left sub-panel plots the first stage, the middle sub-panel plots the ITT, and the right sub-panel plots the TOT. The outcome variable for the ITT and TOT is minutes of platform usage. Panel (a) estimates serparate effects for each randomly assigned treatment timing group, and panel (b) estimates separate effects for each randomly assigned default treatment group. Panel (c) estimates separate effects for users with below or above median baseline usage, and panel (d) estimates separate effects for users who ever had or had not uninstalled the platform.

Figure 6: Heterogeneous Treatment Effects by Overuse and Underestimation

(a) Overuse



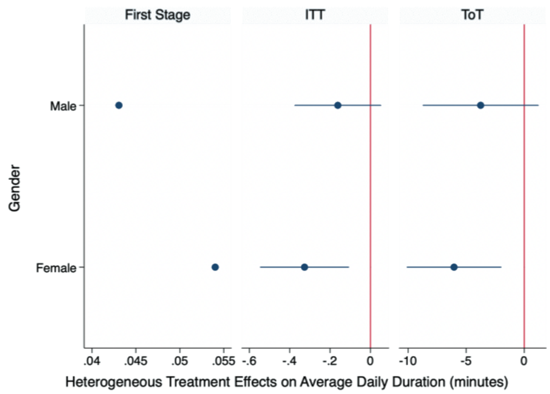
(b) Underestimation



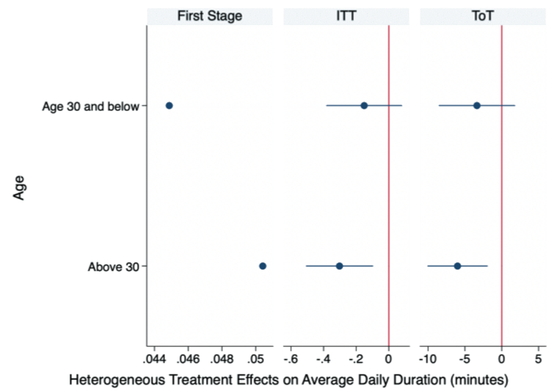
Notes: This figure plots point estimates and 95% confidence intervals from first stage, ITT, and TOT regressions. In each panel, the left sub-panel plots the first stage, the middle sub-panel plots the ITT, and the right sub-panel plots the TOT. The outcome variable for the ITT and TOT is minutes of platform usage. Panel (a) estimates separate effects for users who under-used, had no misuse, or overused the platform, and panel (b) estimates separate effects for users who ever had overestimated, accurately estimated, or underestimated their platform usage.

Figure 7: Heterogeneous Treatment Effects by Demographics

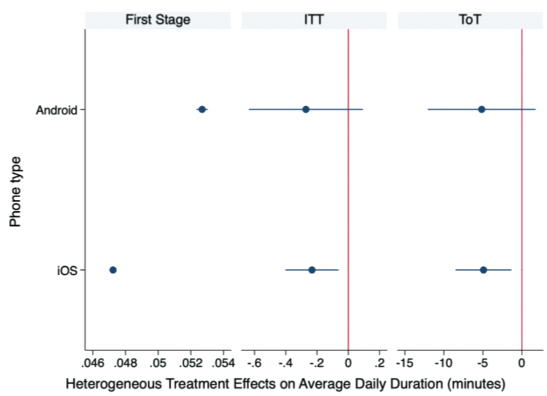
(a) Gender



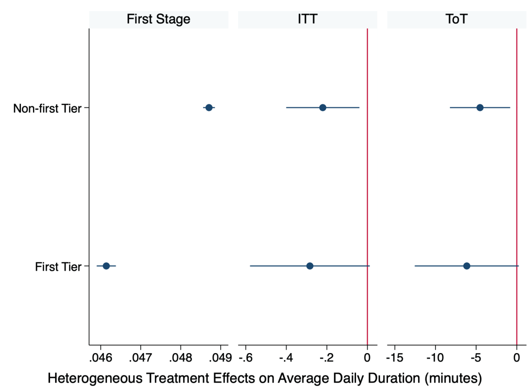
(b) Age Group



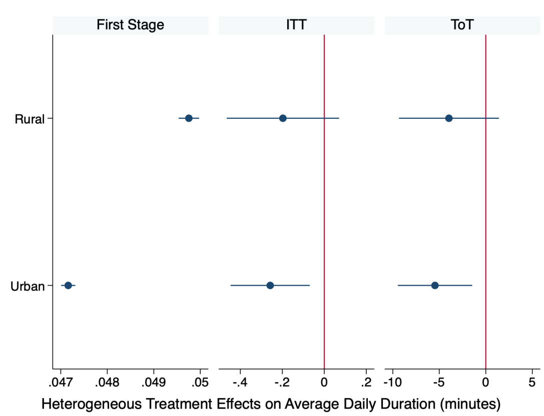
(c) Phone Type



(d) City Tier



(e) Urban or Rural



Notes: This figure plots point estimates and 95% confidence intervals from first stage, ITT, and TOT regressions. In each panel, the left sub-panel plots the first stage, the middle sub-panel plots the ITT, and the right sub-panel plots the TOT. The outcome variable for the ITT and TOT is minutes of platform usage. Panel (a) estimates separate effects by gender, panel (b) by age group, (c) by phone type, (d) by city tier, and (e) by urban vs. rural.

trend holds, take-up might be even higher if participants were asked to set up an alarm before they started using the platform that day, instead of being in a "hot" state in the midst of use.

We address some concerns over the valid comparison across these treatment groups. Users assigned to different invitation timings might face: (i) a different number of invites over the experimental period, and (ii) differential selection into seeing treatment arising from different likelihoods of seeing an invitation on any given day. To ameliorate the first of these concerns, observe first that the percentage who take-up treatment on *first* invite is still much lower for the 120 min treatment group, as seen in Table 9 of the appendix.

One might still be worried about differential selection of the 91.3% of participants who see at least one invitation in the 120 minute treatment group, versus the near 100% of participants invited in the 20 minute group. Those in the former group might mechanically have a higher average daily usage, since only those who spend more than 120 minutes on the platform would be able to receive the treatment. To address this, we perform a simple comparative calculation. For the take-up rate to be the same amongst the 20 and 120 minute treatment group, we would need the remaining participants in the 120 minute treatment group who never seen the invitation to have a 8.5% take-up rate, *almost 5 times the take-up rate for the rest of the people in that treatment group.* This seems highly unlikely. As such, we conclude that users likely exhibit higher take-up rates of commitment if they are asked to do so earlier on in their usage.

Another feature of the choice environment also matters: defaults. Those who are shown a looser default alarm suggestion have higher take up rates at 4.9% as opposed to 4.7%, as seen in 5. However, we note that in both cases—sensitivity to timing and defaulted alarm suggestions—the differences are small and therefore are unlikely to explain the low take-up overall.

4.3 Why Is Take-Up Low Relative to Previous Studies?

Several earlier experiments on social media usage found higher adoption rates for similar digital wellness or "commitment" tools, despite the fact that they often involve downloading a third-party app or taking multiple steps to set up the functionality. For example, Allcott et al. (2022) find that 78% of their participants set binding limits on their phone usage when offered a proprietary app that would help them limit their future use, while Hoong (2021) finds a 66% take-up rate for participants prompted to set up Screen Time limits on their Apple devices.

We purport that the discrepancy between these studies and ours is potentially due to the high selection of participants into online experiments. People who respond to external advertisements or research calls often have a stronger interest in reducing app time. We focus on a large sample of a platform's in-app users; previous work recruits participants on a voluntary basis through Facebook ads or through MTurk. There is one exception to this—Acland and Chow (2018) recruit around 10,000 participants by working directly with online game Wordplay and therefore may be much less subject to self-selection. As is consistent with what we would expect, their take-up rate is lower than the aforementioned studies—25% of players of online game Wordplay choose to set in-app commitments that limit their gameplay.

Those willing to take the time to consent to participating in an experimental study might be fairly different from those on the platform as a whole. We see some evidence of this even amongst our surveyed participants, despite the fact that our survey took the form of a pop-up that was only 2 questions long. Take-up of the reminders amongst our surveyed respondents average around 8%, over 70% higher than take-up in the wider experimental population.

As seen previously in Table 2, although the survey was shown to a random subset of our 13 million participants, there was clear selection of participants into answering the survey: of the 8% of participants that responded to the survey, respondents were disproportionately male, younger in age, from more urbanized and higher-tier areas, were more active and exhibited higher daily baseline usage, and were more likely to have ever uninstalled the platform previously. This is interesting when viewed in context of the heterogeneous take-up of our wider population: we find that those with lower baseline usage, are female, older than age 30, and come from more rural areas are more likely to take up treatment. In other words, though the people that select into our survey on the surface come from demographics that are *less* likely to take-up commitment, the take-up rates of survey participants are nevertheless much higher than those of the wider population. This suggests that demographic targeting in experiments may even further drive up take-up rates. Experiments tend to aim to recruit a balanced sample based on observable demographics like gender, age, and location—in order to achieve that balance, they may be unintentionally further funneling in populations that are more likely to take-up the device.

Does this selection effect result in a substantially different population that accounts for the different take-up rates? Though it is hard to make direct comparisons between our population and those of previous experiments—our users are based outside the US, and are using a video-based social media platform as opposed to multi-media social networks like Facebook—there is cause to believe that they might be fairly different along important dimensions. Perceived self-control issues seem to be less prevalent than in previous experimental studies: 49% of our users exhibit perceived self-control problems, lower than in other studies—e.g., Hoong (2021) finds that over 90% of individuals in their sample has perceived self-control issues. This suggests either that previous studies might be disproportionately selecting respondents with higher perceived self-control problems, or people might have lower perceived self-control issues on this particular platform than in other studied settings.

Our TOT effects show a 9% daily reduction among adopters. While meaningful, it is smaller than the 33% reduction observed by Hoong (2021) for certain "Screen Time" alarms. This can be driven by differences in (i) tools, (ii) the contextual setting (e.g., this platform vs. Facebook), and (iii) the population, which is wider and less selected than in other studies. The alarm feature in our study is very similar to the "Screen Time" limits—soft commitment alarms that pop up after X minutes of use with the option to extend usage by 15 minutes at a time. Even when viewed on the most favourable heterogeneous dimensions, the treatment-on-treated effects of our intervention are low by comparison. This suggests that our users might be substantially different in nature from those in the previous study: analogous to the take-up selection story, those for whom commitment is highly effective might voluntarily self-select into experiments.

Lastly, unlike in other studies, our invitations appear *during* usage sessions, at which point participants may be less inclined to interrupt or limit their session. By contrast, studies that prompt participants *before* launching the platform may boost acceptance and subsequent alarm adherence.

5 A Model of Quasi-Hyperbolic Discounting

The previous sections highlight several findings related to the Digital Wellness alarm. Adoption of this wellness tool is moderate and appears driven by whether users perceive a need to manage their usage, how strongly they believe the alarm will help, and when they are prompted to adopt it. Additionally, individuals who are sophisticates demonstrate lower take-up than naifs. Moreover, those who overestimate their future consumption—or who end up using the platform *less* than their stated ideal—display a higher propensity to adopt. To unify these observations, we develop and present a three-period model of quasi-hyperbolic discounting based on Carrera et al. (2022).

5.1 Model Setup

We consider a timeline with three periods $\{0, 1, 2\}$. In period 0, individuals choose whether or not to set up commitment—i.e., future alarms. This is denoted by $m \in \{0, 1\}$, where $m = 0$ if an individual chooses not to set an alarm, and $m = 1$ if they choose to set one.

In period 1, individuals choose their usage of the platform corresponding to action $a \in \{0, 1\}$, where $a = 0$ represents using the platform *more* than they would ideally desire, whereas $a = 1$ represents the task of controlling their own usage of their platform and *not* overusing the platform more than they desire. Taking this action generates immediate stochastic costs and delayed deterministic benefits. The cost of choosing $a = 1$ in this period is denoted c and realized in period 1.

In period 2, the deterministic benefits b of taking the action are realized, where $b \geq 0$. For example, this could be the benefits of spending more time writing up a draft of a paper rather than scrolling through EconTwitter. Individuals also incur a delayed penalty p in this period if they chose to set an alarm in period 0 (i.e., $m = 1$) and exceeded their contract in period 1 (i.e., $a = 0$). p can be thought of as the psychological or regret cost of hitting an alarm, where $p \geq 0$.

Importantly, we assume that for a given individual, the cost c is stochastic. In period 0, the individual correctly perceives that the possible realizations of the cost are distributed as

$c \sim F$, where F is some continuous distribution.[22] In this context, the uncertainty about c could represent uncertainty about future opportunity cost of time, or how tired or motivated the individual might be.

For notational simplicity, we suppress the individual subscript for the rest of the paper. Assume individuals have quasi-hyperbolic preferences given by

$$U^t(u_t, u_{t+1}, ..., u_2) = u_t + \beta \sum_{\tau=t+1}^{2} \delta^\tau u_\tau$$

where u_t is the period t utility flow. So here we have:

$$u_1 = a \cdot (-c)$$

$$u_2 = a \cdot b + (1-a) \cdot m \cdot (-p)$$

Following O'Donoghue and Rabin (2001), we allow individuals to mispredict preferences. In period 0, individuals believe that their period 1 self will have a present-focus factor $\hat{\beta} \in [\beta, 1]$. To simplify things, we set $\delta = 1$ for all individuals.

To determine the conditions under which individuals will demand commitment in period 0, we work backwards from their decisions in period 1. In period 1, individuals perceive that they will choose $a = 1$ if $c \leq \hat{\beta}(b + mp)$. Define $V(m) = E(U^0|m)$, i.e., the individual's subjective expected utility from the perspective of period 0 when m is their commitment choice. An individual's perceived expected utility in period 0 is:

$$V(m) = \beta[\int_{c>\hat{\beta}(b+mp)} (-mp)dF + \int_{c\leq\hat{\beta}(b+mp)} (b-c)dF]$$

Our object of interest is $\Delta V = V(1) - V(0)$, i.e. the difference in perceived expected utility with and without taking up commitment in time 0. If $\Delta V > 0$, the individual demands commitment, i.e., they choose $m = 1$; they choose $m = 0$ otherwise. Note that:

$$\frac{\Delta V}{\beta} = \int_{c>\hat{\beta}(b+p)} (-p)dF + \int_{c=\hat{\beta}b}^{c=\hat{\beta}(b+p)} (b-c)dF$$

[22] We do not put restrictions on the distribution F except to require that $c > 0$ with positive probability.

5.2 Stylized Fact 1: People Without Perceived Self-Control Problems Have Low Demand for Commitment

Individuals with no perceived self-control problems fall into one of two types: either they have no time inconsistencies at all, in which case $\hat{\beta} = \beta = 1$, or they are totally naive about their self-control problems, i.e. $\hat{\beta} = 1, \beta < 1$. In both cases, $\hat{\beta} = 1$. We show that as $\hat{\beta} \to 1$, $\frac{\Delta V}{\beta} \to 0$, so individuals have no reason to demand commitment.

Let $\hat{\beta} \to 1$. We have that

$$\frac{\Delta V}{\beta} \to \int_{c>b+p} (-p)dF + \int_{c=b}^{c=b+p} (b-c)dF$$

First consider the case where $c > b$. Both terms are negative so taking up commitment is never worth it.

In the other case where $c \leq b$, neither of the terms are incurred. So we also have $\frac{\Delta V}{\beta} \to 0$, as needed.

5.3 Stylized Fact 2: Lower Perceived Effectiveness of the Commitment Drives Low Take-Up

A lower effectiveness of the alarm implies a lower penalty p. When an alarm pops up at users' self-set optimal limit time, users may still go on to exceed it because the psychological costs are small.

To see this, note that as $p \to 0$, we have $\frac{\Delta V}{\beta} \to 0$. The first term approaches 0 since the cost of exceeding the alarm is low even in the event of overuse. Moreover, the likelihood of incurring the second term becomes negligible—intuitively, the low effectiveness doesn't give a high enough penalty in future periods to be effective in limiting usage.

5.4 Stylized Fact 3: Timing of Invitations Influences Adoption

The timing of the invitation matters towards take-up: the more into their daily use they are, the less likely users are to take-up commitment. In context of our model, being in the heat of watching videos for hours is more akin to being in time period 1, as opposed to being in

0. We show that individuals will not want to take up commitment if they make the decision in period 1.

Here, the commitment decision is made at the same time as the usage decision. If they choose not to exceed their ideal use, i.e., $a = 1$, then the payoffs are the same regardless of their commitment choice and $-c + \hat{\beta}b$ is incurred. On the other hand, if they choose to exceed their ideal use, i.e., $a = 0$, then it is always better to avoid commitment and to receive a payoff of 0 as opposed to incurring $-\hat{\beta}p$. Thus individuals are better off choosing $m = 0$.

5.5 Stylized Fact 4: Sophisticated Users Can Have Lower Take-Up Than Naifs

This stylized fact means that demand for commitment can be non-monotonic in $\hat{\beta}$. In other words, when $\hat{\beta}$ is sufficiently high, then demand for commitment is decreasing in $\hat{\beta}$.[23] To see this, we first notice that as $\hat{\beta} \to 0, \Delta V/\beta \to -p$, which establishes that there exists $\underline{\beta}$ such that $\Delta V < 0$ for each p. We next note that because ΔV is continuous in $\hat{\beta}$, and because $\Delta V < 0$ for $\hat{\beta} = 1$ as we established above, we also have that there exists a $\bar{\beta}$ such that $\Delta V < 0$ if $\hat{\beta} > \bar{\beta}$.

What remains is to show that under certain conditions there exists $\hat{\beta}$ such that $\Delta V > 0$. For this, we assume for simplicity that the period 1 stochastic cost c follows a simple distribution: taking either a high value $\bar{c}$ with probability μ or a low value $\underline{c}$ with probability $(1 - \mu)$. Further assume that $\underline{c} < b$ and $\bar{c} > b$, for all individuals. We will show that there exists $\hat{\beta}$ such that $\Delta V > 0$ for sufficiently low μ, demonstrating the existence demand for commitment.

Consider the low realization $\underline{c}$. If $\underline{c} > \hat{\beta}(b+p)$, then the commitment contract is perceived to not be worthwhile, since it only increases the penalties incurred. If $\underline{c} < \hat{\beta}b$ then the commitment is also perceived as not worthwhile, because the individual already believes that they will choose action $a = 1$. However, for $\frac{\underline{c}}{b+p} < \hat{\beta} < \frac{\underline{c}}{b}$, the gains from commitment exceed the costs and the individual would perceive themselves to benefit from taking up the

[23] This section builds on the proofs in Heidhues and Kőszegi (2009), and Carrera et al. (2022) to show the conditions for non-monotonic demand for commitment.

contract.[24] Since the decision of whether to commit is dominated by the effect of commitment on outcomes for the *low* realization $\underline{c}$ when μ is sufficiently close to zero, then there would be demand for commitment for sufficiently low μ at the intermediate values of $\hat{\beta}$, such that $\frac{c}{b+p} < \hat{\beta} < \frac{c}{b}$, despite no demand for commitment for sufficiently high or low values of $\hat{\beta}$.

5.6 Stylized Fact 5: Those Who Overestimate or Underuse Have Higher Take-Up

The next result follows from the proof above. Recall that overestimation is defined by *actual-predicted* < 0 *for next 7 days* and underuse is defined by *actual-ideal* < 0 *for next 7 days*. Both can be driven by low c realizations: when c is low, users find it very easy not to overuse the platform, and so the actual time they spend on it ends up being very low in that period.

Those who overestimate or underuse mechanically have lower actual time realizations. As such, we could be selecting individuals that are more likely to have low realizations of c; in other words, we are selecting a population of individuals with low μ. Recall from above that $\Delta V > 0$ for small enough μ—that is, the lower μ, the higher the demand for commitment. Thus, it is possible for individuals who over-estimate or underuse to have higher demand for commitment.

6 Conclusion

This study explores the demand for, and effectiveness of, a Digital Wellness alarm designed to support users in fostering healthier digital habits. By partnering directly with a social media platform, we conducted a large-scale field experiment engaging over 13 million users in-situ. Our findings reveal that while demand for the alarm is modest, it highlights the complexities of scaling wellness tools in a real-world setting and the nuanced role of individual perceptions, tool effectiveness, and contextual design.

We find that adoption of the Digital Wellness alarm reached 4.8% among users after exposing them to up to several invitations. This is in contrast to some previous studies,

[24] Since we assumed that $\underline{c} < b$, it will be the case that $\frac{c}{b} < 1$, thus $\hat{\beta} < \frac{c}{b} \implies \hat{\beta} < 1$, so $\hat{\beta}$ in this range is a feasible value.

suggesting that it is important to consider how the selection of participants into voluntary experimental studies can introduce nonresponse bias. Even selection into a fairly quick, two-click survey drove take-up up to around 8%, over 70% higher than take-up in the wider experimental population, lending credence to the selection story.

In terms of impact, the tool reduced the average daily usage of adopters by about 9%, affirming the potential of such tools to positively impact individuals who actively seek greater control over their digital habits. The implications extend beyond time spent on the platform, however. Platforms implementing wellness tools can balance user autonomy with fostering healthier habits, potentially generating goodwill, enhancing user satisfaction, and preempting regulatory scrutiny. Governments and policymakers globally are increasingly concerned about the societal impacts of overuse, making proactive measures by platforms both timely and strategic.

However, effective digital wellness cannot rest solely on the availability of self-limiting tools on platforms alone. A key takeaway from this study is the need for a holistic, collaborative approach to digital wellness. Platforms alone cannot shoulder the responsibility for fostering healthier habits, given that individual consumers might not fully adopt these tools even when offered. Future initiatives could benefit from partnerships with mental health professionals, educators, and advocacy groups to create comprehensive solutions that address both behavioral and psychological dimensions of overuse. Tools like alarms can be powerful, but their effectiveness depends on complementing them with awareness campaigns, educational efforts, and motivational support for users.

Ultimately, understanding and addressing digital well-being requires a multi-faceted effort. Future research should not only explore the interplay between individual preferences and platform incentives but also examine how a collaborative ecosystem——including mental health experts, policymakers, and users—can collectively foster healthier digital habits. Such efforts can ensure that digital tools are not just accessible but truly transformative, aligning user experience with well-being.

7 Appendix

7.1 Balance Tables

Table 5: Balance Table, Split by Treatment Group

	(1) Control Mean	(2) $T = 20$ Mean	(3) $T = 60$ Mean	(4) $T = 120$ Mean	(5) $T = 20$ p-val	(6) $T = 60$ p-val	(7) $T = 120$ p-val
Percent Male	54.7	54.7	54.7	54.7	0.751	0.171	0.511
Age	32.5	32.5	32.5	32.5	0.656	0.801	0.137
Percent Urban	66.4	66.4	66.5	66.4	0.296	0.108	0.737
Percent Tier 1 City	26.4	26.3	26.4	26.3	0.319	0.298	0.213
Price of Phone (USD)	519.5	519.2	519.7	519.1	0.281	0.580	0.239
Percent of Days Active (Prior 28 Days)	91.4	91.4	91.4	91.4	0.764	0.600	0.590
Average Daily Minutes (Prior 28 Days)	152.1	152.1	152.2	152.2	0.924	0.203	0.222
Days Active Since Signup	417.0	417.1	417.2	417.1	0.628	0.394	0.605
Percent Ever Uninstalled App	54.1	54.1	54.0	54.1	0.773	0.136	0.666
Mean Uninstalls	1.3	1.3	1.3	1.3	0.781	0.976	0.951

Notes: Columns (1)-(3) report means for control and treatment groups. Column (4)-(7) report the t-test p-values.

Table 6: Balance Table, Split by Default Treatment Group

	(1) Control Mean	(2) Low default Mean	(3) Middle default Mean	(4) Low default p-val	(5) Middle default p-val
Percent Male	54.6	54.7	54.7	0.241	0.207
Age	32.5	32.5	32.5	0.154	0.088
Percent Urban	66.4	66.4	66.4	0.534	0.743
Percent Tier 1 City	26.3	26.4	26.3	0.107	0.256
Price of Phone (USD)	519.3	519.5	519.2	0.622	0.702
Percent of Days Sctive (Prior 28 Days)	91.4	91.4	91.4	0.757	0.474
Average Daily Minutes (Prior 28 Days)	152.1	152.2	152.2	0.129	0.031
Days Active Since Signup	417.2	416.9	417.2	0.262	0.997
Percent Ever Uninstalled App	54.1	54.1	54.1	0.937	0.767
Mean Uninstalls	1.3	1.3	1.3	0.674	0.892

Notes: Columns (1)-(3) report means for control and treatment groups. Column (4)-(5) report the t-test p-values.

Table 7: Balance Table, For Survey Participants Only ($n = 63,459$)

	(1) Control	(2) Treatment	(3) p-val	(4) sd
Percent Male	57.8	58.4	.158	-
Age	27.6	27.7	.306	7.3
Percent Urban	68.1	67.8	.524	-
Percent Tier 1 City	28.2	28.4	.61	-
Price of Phone (USD)	571.4	564.7	.039	324.9
Percent of Days Active (Prior 28 Days)	9191.4	9206.4	.467	2115.8
Avearage Daily Minutes (Prior 28 Days)	168.4	168.5	.914	93.4
Days Active Since Signup	517.8	513.8	.267	365.0
Percent Ever Uninstalled App	56.7	57.1	.37	-
Mean Uninstalls	1.4	1.4	.965	3.1

Notes: Columns (1) and (2) report means for control and treatment groups. Column (3) reports the t-test p-value, and Column (4) reports the standard deviation.

7.2 Take-Up

Table 8: Baseline Demographic Characteristics, by Take-Up Invitation Number

	(1) No Takeup	(2) Takeup Invite 1	(3) Takeup Invite 2-5	(4) Takeup Invite 6-14	(5) p-val (2) vs (3)	(6) p-val (3) vs (4)	(7) p-val (4) vs (5)
Percent Male	55.0	49.6	48.8	49.0	0.000	0.001	0.231
Age	32.5	31.5	32.6	32.9	0.000	0.000	0.000
Percent Urban	66.5	66.2	65.5	64.9	0.153	0.002	0.000
Percent Tier 1 City	26.5	26.1	25.5	25.1	0.019	0.003	0.003
Percent of Days Active (Prior 28 Days)	91.8	89.1	89.7	93.6	0.000	0.000	0.000
Avearage Daily Minutes (Prior 28 Days)	154.3	122.2	121.7	150.4	0.000	0.191	0.000
Days Active Since Signup	419.7	386.1	398.4	425.7	0.000	0.000	0.000
Percent Ever Uninstalled App	53.6	60.9	58.5	53.4	0.000	0.000	0.000
Mean Uninstalls	1.3	1.5	1.4	1.3	0.000	0.000	0.000

Notes: Columns (1) through (4) report means. Columns (5) through (7) report the t-test p-value.

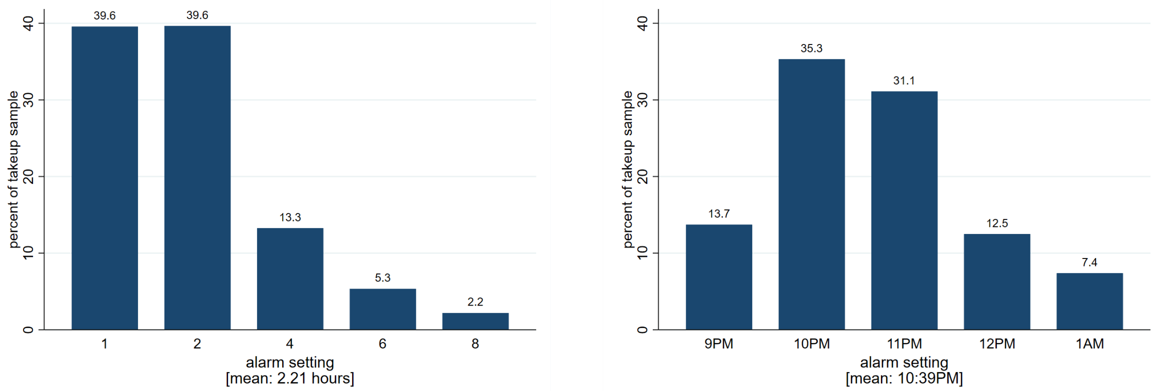
Table 9: Take-Up Details, by Treatment Sub-Group

	(1) $T = 20$	(2) $T = 60$	(3) $T = 120$
Share Saw Any Invitation	99.63	97.84	91.28
Mean Number of Invitations seen	5.07	4.32	2.94
Median Number of Invitations seen	5	4	3
Percent Set Alarm on First Invitation	2.39	2.13	1.65
Percent Set Alarm on First Invitation (Contingent on Seeing Invitation)	2.39	2.08	1.81
Percent Ever Set Alarm	6.02	5.06	3.68

Notes: Columns (2)-(4) report means for the different treatment groups.

Figure 8: Distribution of Endogenously Chosen Alarm Timings: All Treatments

(a) Daytime **(b)** Nighttime



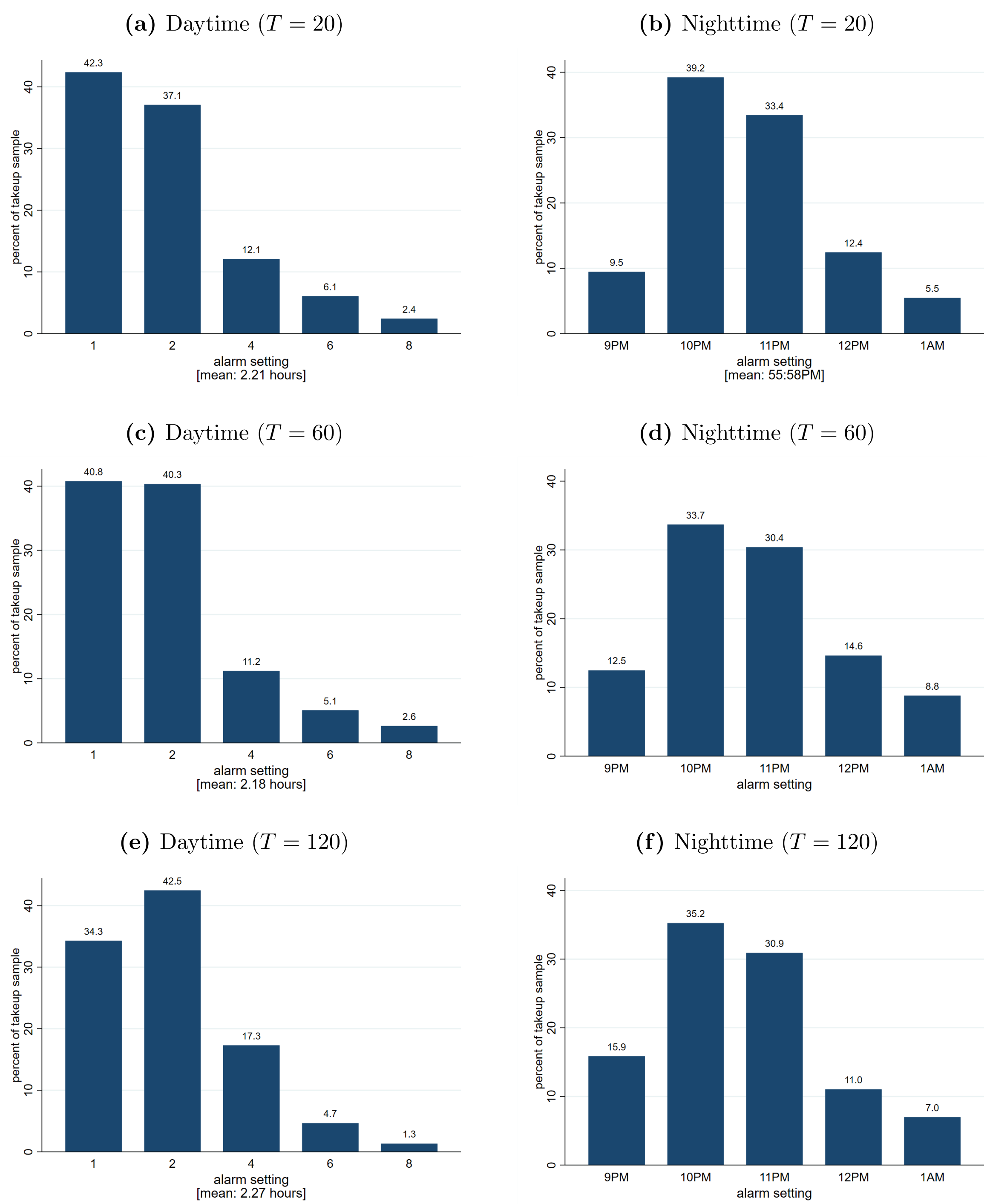
Notes: Bar heights plot the share of the subsample that set each alarm type. The left panel is for the subsample who were invited to set a daytime alarm, which could be set at 1 hour, 2 hours, 4 hours, 6 hours, 8 hours. The right panel is for the subsample who were invited to set a nighttiem alarm, which could be set at 9pm, 10pm, 11pm, 12am, or 1am.

Table 10: Limit Tightness

	(1) Pre-Period Mean	(2) Pre-Period Sd	(3) Post-Period Mean	(4) Post-Period Sd
Percent of Days Limit Binds	46.1	-	46.8	-
Percent Limit Ever Binds	97.8	-	93.0	-
Absolute Tightness (mins)	46.4	74.4	52.4	78.4
Absolute Tightness when Binding (Mins)	94.7	90.8	100.0	91.3
Relative Tightness (% of Daily Usage)	19.9	11.5	21.4	15.1
Relative Tightness When Binding (% of Daily Usage)	42.6	15.0	43.8	15.7

Notes: This table presents different measures of limit tightness of the reminders set by those who opt to take-up treatment at any point over the two week experimental period. Columns (2) and (3) compare the endogenously-chosen limit to users' pre-period daily usage (7 days prior to experimental period). Columns (4) and (5) compare the endogenously-chosen limit to users' post-period daily usage (7 days following the experimental period).

Figure 9: Distribution of Endogenously Chosen Alarm Timings: By Treatment Groups



Notes: Bar heights plot the share of the subsample that set each alarm type. Left panels are for the subsample who were invited to set a daytime alarm, which could be set at 1 hour, 2 hours, 4 hours, 6 hours, 8 hours. Right panels are for the subsample who were invited to set a nighttiem alarm, which could be set at 9pm, 10pm, 11pm, 12am, or 1am.

Table 11: Predicting Take-Up (At All Over 2-Week Period)

	(1) Takeup	(2) Takeup	(3) Takeup	(4) Takeup
60 minute treatment	-0.99 (0.02)***	-0.99 (0.02)***	-0.99 (0.02)***	-1.00 (0.02)***
120 minute treatment	-2.41 (0.02)***	-2.41 (0.02)***	-2.41 (0.02)***	-2.42 (0.02)***
Default middle	0.16 (0.01)***	0.16 (0.01)***	0.16 (0.01)***	0.16 (0.01)***
Percent of days active 28 days before experiment		2.73 (0.04)***	2.81 (0.04)***	2.88 (0.04)***
Days active since signup		-0.00 (0.00)***	-0.00 (0.00)***	-0.00 (0.00)***
Daily hours 28 days before experiment		-1.00 (0.02)***	-1.00 (0.02)***	-0.99 (0.02)***
Daily hours 14 days before experiment		0.16 (0.02)***	0.15 (0.02)***	0.14 (0.02)***
Daily hours 7 days before experiment		0.23 (0.01)***	0.23 (0.01)***	0.25 (0.01)***
Ever uninstalled			0.21 (0.02)***	0.41 (0.02)***
Recent uninstalled			0.21 (0.02)***	0.17 (0.02)***
Mean uninstalls			0.00 (0.00)*	0.01 (0.00)***
Male				-1.05 (0.01)***
Age 18-23				-0.10 (0.02)***
Age 24-30				1.04 (0.02)***
Age 31-40				1.69 (0.02)***
Observations	11182747	11182747	11182747	10445786
R^2	0.0022	0.0039	0.0040	0.0059

Notes: This table presents OLS regressions of limit adoption (=1 as long as participants adopts an alarm at any point over the two weeks) on different demographic variables, for all the participants assigned to any treatment group.

Table 12: Predicting Take-Up (On First Invitation)

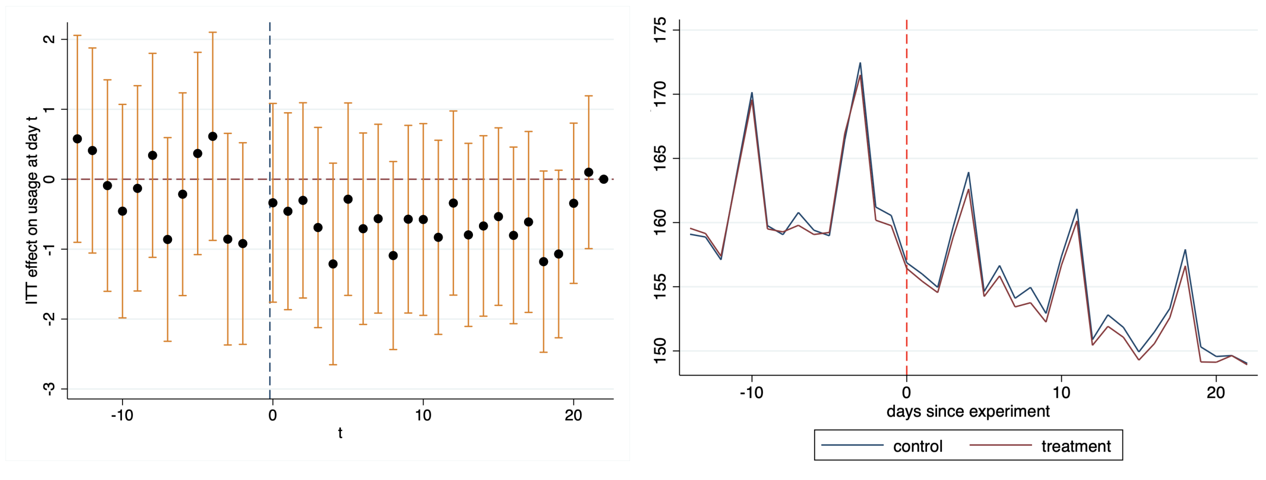
	(1) Takeup	(2) Takeup	(3) Takeup	(4) Takeup
60 minute treatment	0.01	0.01	0.01	0.01
	(0.01)	(0.01)	(0.01)	(0.01)
120 minute treatment	0.02	0.02	0.02	0.02
	(0.01)**	(0.01)**	(0.01)**	(0.01)**
Default middle	0.18	0.18	0.18	0.17
	(0.01)***	(0.01)***	(0.01)***	(0.01)***
Percent of days active 28 days before experiment		0.36	0.41	0.43
		(0.02)***	(0.02)***	(0.02)***
Days active since signup		-0.00	-0.00	-0.00
		(0.00)***	(0.00)***	(0.00)***
Daily hours 28 days before experiment		-0.32	-0.31	-0.31
		(0.01)***	(0.01)***	(0.01)***
Daily hours 14 days before experiment		0.06	0.06	0.05
		(0.01)***	(0.01)***	(0.01)***
Daily hours 7 days before experiment		0.00	0.00	0.01
		(0.00)	(0.00)	(0.00)
Ever uninstalled			0.19	0.21
			(0.01)***	(0.01)***
Recent uninstalled			0.05	0.04
			(0.01)***	(0.01)***
Mean uninstalls			0.00	0.00
			(0.00)**	(0.00)***
Male				-0.20
				(0.01)***
Age 18-23				0.28
				(0.01)***
Age 24-30				0.32
				(0.01)***
Age 31-40				0.39
				(0.01)***
Observations	11182747	11182747	11182747	10445786
R^2	0.0001	0.0014	0.0015	0.0018

Notes: This table presents OLS regressions of limit adoption on first invitation on different demographic variables, for all the participants assigned to any treatment group.

7.3 Treatment Effects

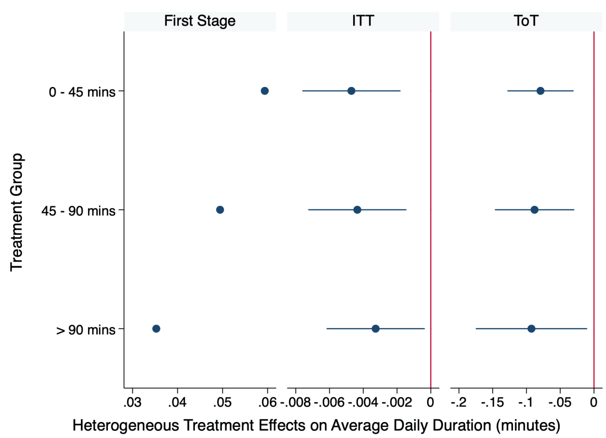
Figure 10: Daily Usage by Day

(a) Treatment effect **(b)** Daily Means



Notes: Panel (a) plots points estimates and 95% confidence intervals from estimating a daily version of the model in Equation (2). The outcome variable is minutes using the platform. Panel (b) plots the mean of minutes using the platform separately for the control group (blue) and treatment group (red).

Figure 11: Heterogeneous Treatment Effects on Log Average Daily Duration, By Treatment Group



Notes: This figure plots point estimates and 95% confidence intervals from first stage, ITT, and TOT regressions, estimating separate effects for each treatment timing group. The outcome variable is *log* of daily minutes. The left panel plots the first stage, the middle panel plots the ITT, and the right panel plots the TOT.

7.4 Survey Data

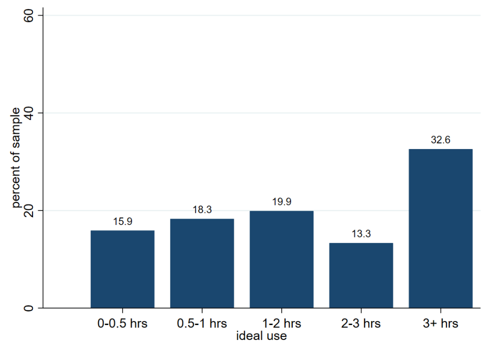
Table 13: Classification of Survey Users Into Overuse and Overestimation

	Overestimated (Actual $<$ Predicted)	No misestimation (Actual $\sim$ Predicted)	Underestimation (Actual $>$ Predicted)	Total
Underuse				12,521
(Actual $<$ Ideal)				(20%)
No misuse				15,198
(Actual $\sim$ Ideal)				(24%)
Overuse	5,818	10,972	18,950	35,740
(Actual $>$ Ideal)	(*16%*)	(*31%*)	(*53%*)	(56%)
Total				63,459

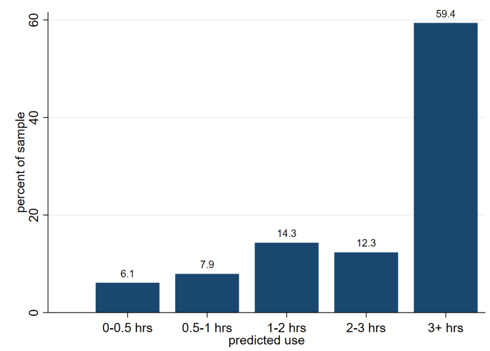
Notes: This table reports the number of survey participants according to the relationship between their predicted, ideal, and actual usage of the platform. The right column reports the number according to whether they underuse, do not misuse, or overuse. The "Overuse" row additionally reports the number of those who overuse who additionally overestimate, do not misestimate, or underestimate their future usage.

Figure 12: Ideal and predicted use

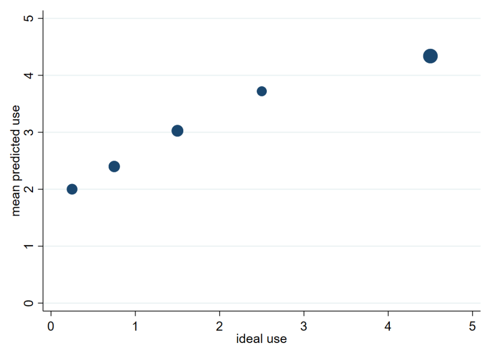
(a) Ideal Use



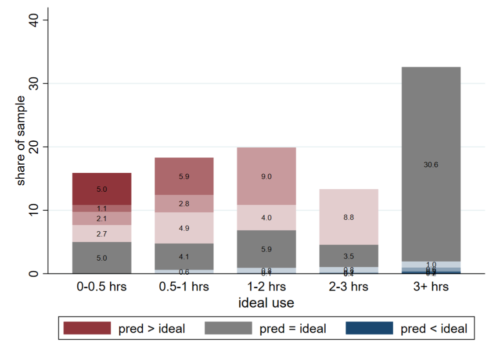
(b) Predicted Use



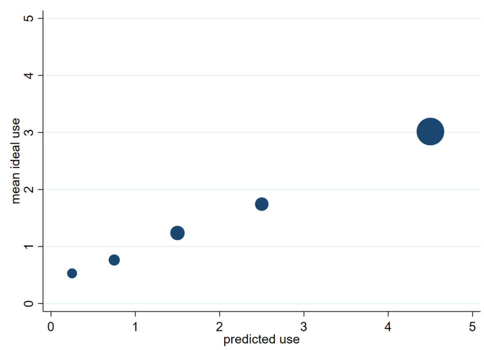
(c) Predicted vs Ideal



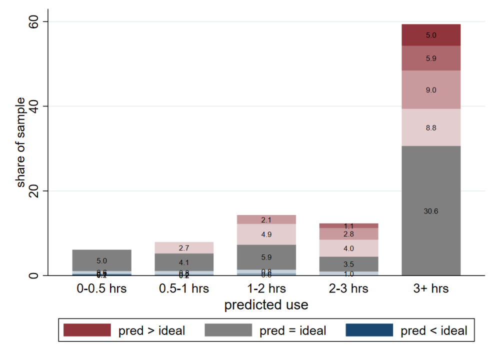
(d) Predicted vs Ideal



(e) Ideal vs Predicted



(f) Ideal vs Predicted



Notes: These figures plot the distribution of ideal and predicted usage among the 63k users who completed the survey. Panels (a), (b), (d), and (f) plot categorical answers. Panels (c) and (e) take the midpoint of each range, and assigns 4.5 hours for the > 3 category. Panels (a) and (b) plot distributions, (c) and (d) plot predicted use as a function of ideal use, and (e) and (f) vice-versa. Circle size is proportional to the number of observations.

7.5 Robustness Check

Table 14: Effects on Usage Behavior (All Treatments)

	(1) Daily usage	(2) Log daily usage	(3) Share days active	(4) Daily usage if active
First stage	0.048	0.048	0.048	0.048
	(0.000)***	(0.000)***	(0.000)***	(0.000)***
ITT	-0.239	-0.003	-0.000	-0.242
	(0.079)***	(0.001)**	(0.000)	(0.080)***
TOT	-4.960	-0.069	-0.008	-5.003
	(1.633)***	(0.027)**	(0.005)	(1.647)***
User & Time FE	Y	Y	Y	Y
Observations	25965254	25965254	25965254	25965254
Users	12982627	12982627	12982627	12982627
Control Mean	145.7	4.4	0.9	144.4
ITT R^2	0.002	0.010	0.007	0.002

Notes: This table presents the ITT and TOT effect. The pre-period is defined as the 7 days before the experiment, i.e. June 29th to July 5th. The post-period is defined as the 7 day period between July 21st to July 27th. Column (1) details results on average daily usage over the 7 days, Column (2) shows this in log time. Column (3) shows results on the share of the 7 days the user was active (the extensive margin), and Column (4) on the average daily usage over the active days. Standard errors in parentheses, $^{***}p < 0.01, ^{**}p < 0.05, ^{*}p < 0.1$.

References

Acland, Dan, and Vinci Chow. 2018. "Self-Control and Demand for Commitment in Online Game Playing: Evidence from a Field Experiment." *Journal of the Economic Science Association* 4 (1): 46–62.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. 2020. "The Welfare Effects of Social Media." *American Economic Review* 110 (3): 629–76.

Allcott, Hunt, Matthew Gentzkow, and Lena Song. 2022. "Digital Addiction." *American Economic Review* 112 (7): 2424–63.

Andreassen, Cecilie Schou, Mark D. Griffiths, Siri Renate Gjertsen, Elfrid Krossbakken, Siri Kvam, and Ståle Pallesen. 2013. "The Relationships Between Behavioral Addictions and the Five-Factor Model of Personality." *Journal of Behavioral Addictions* 2 (2): 90–99.

Andreassen, Cecilie Schou, Torbjørn Torsheim, Geir Scott Brunborg, and Ståle Pallesen. 2012. "Development of a Facebook Addiction Scale." *Psychological Reports* 110 (2): 501–517.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

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

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. "Working Over Time: Dynamic Inconsistency in Real Effort Tasks." *The Quarterly Journal of Economics* 130 (3): 1067–1115.

Becker, Gary S., and Kevin M. Murphy. 1988. "A Theory of Rational Addiction." *Journal of Political Economy* 96 (4): 675–700.

Beshears, John, James J. Choi, Christopher Harris, David Laibson, Brigitte C. Madrian, and Jung Sakong. 2020. "Which Early Withdrawal Penalty Attracts the Most Deposits to a Commitment Savings Account?" *Journal of Public Economics* 183 104144.

Bianchi, Adriana, and James G. Phillips. 2005. "Psychological Predictors of Problem Mobile Phone Use." *Cyberpsychology & Behavior* 8 (1): 39–51.

Bryan, Gharad, Dean Karlan, Scott Nelson et al. 2010. "Commitment Devices." *Annual Review of Economics* 2 (1): 671–698.

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

Choi, Seungwoo, Hayeon Jeong, Minsam Ko, and Uichin Lee. 2016. "Lockdoll: Providing Ambient Feedback of Smartphone Usage Within Social Interaction." In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 1165–1172.

Collis, Avinash, and Felix Eggers. 2022. "Effects of Restricting Social Media Usage on Wellbeing and Performance: A Randomized Control Trial Among Students." *Plos One* 17

(8): e0272416.

De Quidt, Jonathan, Johannes Haushofer, and Christopher Roth. 2018. "Measuring and Bounding Experimenter Demand." *American Economic Review* 108 (11): 3266–3302.

Elphinston, Rachel A., and Patricia Noller. 2011. "Time to Face It! Facebook Intrusion and the Implications for Romantic Jealousy and Relationship Satisfaction." *Cyberpsychology, Behavior, and Social Networking* 14 (11): 631–635.

Ericson, Keith Marzilli, and David Laibson. 2019. "Intertemporal Choice." In *Handbook of Behavioral Economics*, Volume 2. 1–67.

Genç, Hüseyin Ugur, and Aykut Coskun. 2020. "Designing for Social Interaction in the Age of Excessive Smartphone Use." *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* 1–13.

Giné, Xavier, Dean Karlan, and Jonathan Zinman. 2010. "Put Your Money Where Your Butt Is: A Commitment Contract for Smoking Cessation." *American Economic Journal: Applied Economics* 2 (4): 213–35.

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

Guo, Elaine. 2022. "Social Media and Teenage Mental Health: Quasi-Experimental Evidence." *Working Paper.*

Heidhues, Paul, and Botond Kőszegi. 2009. "Futile attempts at self-control." *Journal of the European Economic Association* 7 (2-3): 423–434.

Hong, Fu-Yuan, Der-Hsiang Huang, Hung-Yu Lin, and Su-Lin Chiu. 2014. "Analysis of the Psychological Traits, Facebook Usage, and Facebook Addiction Model of Taiwanese University Students." *Telematics and Informatics* 31 (4): 597–606.

Hoong, Ruru. 2021. "Self Control and Smartphone Use: An Experimental Study of Soft Commitment Devices." *European Economic Review* 140.

Hunt, Melissa G., Rachel Marx, Courtney Lipson, and Jordyn Young. 2018. "No More FOMO: Limiting Social Media Decreases Loneliness and Depression." *Journal of Social and Clinical Psychology* 37 (10): 751–768.

John, Anett. 2020. "When Commitment Fails: Evidence from a Field Experiment." *Management Science* 66 (2): 503–529.

Kan, Kamhon. 2007. "Cigarette Smoking and Self-Control." *Journal of Health Economics* 26 (1): 61–81.

Karaiskos, Dimitris, Elias Tzavellas, G. Balta, and Thomas Paparrigopoulos. 2010. "P02-232-Social Network Addiction: A New Clinical Disorder?" *European Psychiatry* 25 (S1): 1–1.

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

Kim, Inyeop, Gyuwon Jung, Hayoung Jung, Minsam Ko, and Uichin Lee. 2017. "Let's Focus: Location-Based Intervention Tool to Mitigate Phone Use in College Classrooms." In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 101–104.

Kim, Jaejeung, Joonyoung Park, Hyunsoo Lee, Minsam Ko, and Uichin Lee. 2019. "Lockntype: Lockout Task Intervention for Discouraging Smartphone App Use." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* 1–12.

Klasnja, Predrag, Sunny Consolvo, and Wanda Pratt. 2011. "How to Evaluate Technologies for Health Behavior Change in HCI Research." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 3063–3072.

Ko, Chih-Hung, Ju-Yu Yen, Cheng-Chung Chen, Sue-Huei Chen, Kuanyi Wu, and Cheng-Fang Yen. 2006. "Tridimensional Personality of Adolescents with Internet Addiction and Substance Use Experience." *The Canadian Journal of Psychiatry* 51 (14): 887–894.

Ko, Minsam, Subin Yang, Joonwon Lee et al. 2015. "NUGU: A Group-Based Intervention App for Improving Self-Regulation of Limiting Smartphone Use." *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* 1235–1245.

Koc, Mustafa, and Seval Gulyagci. 2013. "Facebook Addiction Among Turkish College Students: The Role of Psychological Health, Demographic, and Usage Characteristics." *Cyberpsychology, Behavior, and Social Networking* 16 (4): 279–284.

Kuss, Daria J., and Mark D. Griffiths. 2011. "Online Social Networking and

Addiction—A Review of the Psychological Literature." *International Journal of Environmental Research and Public Health* 8 (9): 3528–3552.

Laibson, David. 2015. "Why Don't Present-Biased Agents Make Commitments?" *American Economic Review* 105 (5): 267–72.

Laibson, David I., Andrea Repetto, Jeremy Tobacman, Robert E. Hall, William G. Gale, and George A. Akerlof. 1998. "Self-Control and Saving for Retirement." *Brookings Papers on Economic Activity* 1998 (1): 91–196.

Lee, Heyoung, Heejune Ahn, Samwook Choi, and Wanbok Choi. 2014. "The SAMS: Smartphone Addiction Management System and Verification." *Journal of Medical Systems* 38 (1): 1–10.

Lee, Yu-Kang, Chun-Tuan Chang, You Lin, and Zhao-Hong Cheng. 2014. "The Dark Side of Smartphone Usage: Psychological Traits, Compulsive Behavior and Technostress." *Computers in Human Behavior* 31 373–383.

Lenhart, Amanda, and Mary Madden. 2007. "Social Networking Websites and Teens: An Overview."Technical report, Pew Research Center.

Marotta, Veronica, and Alessandro Acquisti. 2017. "Online Distractions, Website Blockers, and Economic Productivity: A Randomized Field Experiment." *Working Paper.*

Mosquera, Roberto, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie. 2020. "The Economic Effects of Facebook." *Experimental Economics* 23 (2): 575–602.

O'Donoghue, Ted, and Matthew Rabin. 2001. "Choice and Procrastination." *The Quarterly Journal of Economics* 116 (1): 121–160.

Okeke, Fabian, Michael Sobolev, Nicola Dell, and Deborah Estrin. 2018. "Good Vibrations: Can a Digital Nudge Reduce Digital Overload?" *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services* 1–12.

Pelling, Emma L., and Katherine M. White. 2009. "The Theory of Planned Behavior Applied to Young People's Use of Social Networking Web Sites." *Cyberpsychology & Behavior* 12 (6): 755–759.

Ryan, Tracii, Andrea Chester, John Reece, and Sophia Xenos. 2014. "The Uses and

Abuses of Facebook: A Review of Facebook Addiction." *Journal of Behavioral Addictions* 3 (3): 133–148.

Sadoff, Sally, Anya Samek, and Charles Sprenger. 2020. "Dynamic Inconsistency in Food Choice: Experimental Evidence from Two Food Deserts." *The Review of Economic Studies* 87 (4): 1954–1988.

Sagioglou, Christina, and Tobias Greitemeyer. 2014. "Facebook's Emotional Consequences: Why Facebook Causes a Decrease in Mood and Why People Still Use It." *Computers in Human Behavior* 35 359–363.

Schou Andreassen, Cecilie, and Stale Pallesen. 2014. "Social Network Site Addiction-An Overview." *Current Pharmaceutical Design* 20 (25): 4053–4061.

Thaler, Richard H., and Shlomo Benartzi. 2004. "Save More Tomorrow™: Using Behavioral Economics to Increase Employee Saving." *Journal of Political Economy* 112 (S1): S164–S187.

Thaler, Richard H., and Hersh M. Shefrin. 1981. "An Economic Theory of Self-Control." *Journal of Political Economy* 89 (2): 392–406.

Tromholt, Morten. 2016. "The Facebook Experiment: Quitting Facebook Leads to Higher Levels of Well-Being." *Cyberpsychology, Behavior, and Social Networking* 19 (11): 661–666.

Vanman, Eric J., Rosemary Baker, and Stephanie J. Tobin. 2018. "The Burden of Online Friends: The Effects of Giving Up Facebook on Stress and Well-Being." *The Journal of Social Psychology* 158 (4): 496–508.

Wilson, Kathryn, Stephanie Fornasier, and Katherine M. White. 2010. "Psychological Predictors of Young Adults' Use of Social Networking Sites." *Cyberpsychology, Behavior, and Social Networking* 13 (2): 173–177.