The Well-Being Effects of Digital Mental Health Care*

Manuela Angelucci† Raissa Fábregas‡ Antonia Vazquez§

February 22, 2026

PRELIMINARY AND INCOMPLETE

The global adoption of smartphones and the rise of AI-powered digital mental health tools have generated enthusiasm about their potential to expand access to mental health support at low cost. Yet, questions remain about their effectiveness, safety, and the risk of exacerbating isolation or crowding out traditional care. We evaluate an AI-powered mental health app in a randomized controlled trial among Mexican women experiencing mild to severe psychological distress. Over a 6-month period, access to the app led to significant improvements of 0.1-0.3 standard deviations across multiple domains: reductions in anxiety, depression, and stress; enhanced subjective well-being; lower feelings of isolation; improved sleep quality; and greater uptake of healthful behaviors, including reductions in missed work. Treated participants were also more likely to seek traditional psychotherapy, though this is not the main driver of results. Engagement with the app is high in the first month but declines thereafter, a pattern common in digital interventions. Despite this decline, treatment effects persist. We find that participants sustain behavioral and cognitive practices promoted by the app, suggesting that even short-term engagement can generate durable improvements through skill or knowledge acquisition.

*We thank seminar and conference participants at the Texas Development Economics Workshop 2025, the World Bank's AI and the Future of Human Capital in the Global South symposium, CEGA Mental Health Convening, and Aalto University. This research was financially supported by The Agency Fund. This study is registered as Trial AEACTR-0015877 in the AEA RCT Registry and received IRB approval from The University of Texas at Austin (STUDY00006401). We thank Carolina Corral for outstanding research management and support, as well as Monica Vargas, Vasco Andrade Baldini, Estefania Capriata, and Jiqing Zhao for excellent research assistance. We thank the Population Research Center at UT Austin for grant management support.

†Department of Economics, The University of Texas at Austin (email: mangeluc@utexas.edu)

‡Lyndon B. Johnson School of Public Affairs, The University of Texas at Austin (email: rfabregas@utexas.edu)

§Lyndon B. Johnson School of Public Affairs, The University of Texas at Austin (email: antonia.vazquez@utexas.edu)

1 Introduction

Mental health disorders are a leading cause of disability, affecting nearly one billion people worldwide. Despite the scale of the problem, access to effective treatment remains limited due to demand- and supply-side barriers: high costs and stigma limit demand for care, while shortages of trained professionals and chronic underinvestment hinder the supply of publicly funded mental health services.

Recent technological advances offer new opportunities to address these barriers and expand access to mental health support, particularly for the large share of cases involving mild to moderate symptoms of anxiety, stress, or depression (Steel et al., 2014). The widespread adoption of smartphones has enabled the delivery of mental health services through apps and tools in more flexible, private, and cost-effective ways, particularly allowing for on-demand support to both clinical and subclinical populations.[1] More recently, the emergence of artificial intelligence (AI) has made it possible to personalize content to users' needs, adapt to changing symptoms, and optimize the timing and format of interventions. These tools are flexible and can provide immediate relief to users in distress *when* they are experiencing it, which may not be the case for inflexible, scheduled appointments. This flexibility is important, given the high intra-person variability in symptom severity, and the lack of flexibility is an acknowledged limitation of traditional care (Mohr et al., 2014, Kazantzis et al., 2010, Kleiman et al., 2017). However, alongside this potential, significant concerns have emerged, including questions about safety (Vaidyam et al., 2019), evidence of rapid user disengagement (Torous et al., 2020a, Firth et al., 2022), and the risk that digital tools could crowd out traditional care or exacerbate social isolation.

This paper provides experimental evidence on the effects of an AI-powered mental health care app on mental health and subjective well-being. In partnership with the app's developer, we conducted one of the largest randomized evaluations of such a tool to date, with follow-up extending to six months, well beyond the short horizons typical in this literature. We recruited approximately 2,000 women in Mexico who reported at least mild psychological distress and expressed interest in digital mental health support, and randomly assigned free access to the premium version of the app for up to six months. We focused on women from a middle-income

[1] The term "clinical" refers to individuals meeting standard diagnostic criteria for mental health disorders, and "subclinical" to individuals experiencing symptoms that do not meet diagnostic thresholds but may still impair well-being.

country because the mental health burden is concentrated in low- and middle-income countries and disproportionately affecting women through higher rates of depression and anxiety (World Health Organization, 2023, Collaborators et al., 2022). We measured outcomes at one, two, and six months using survey data. Two unique features of our study is that we complement survey data with weekly mood polls, which allow us to assess mental health impacts on a high-frequency basis and minimize recall bias, and administrative data, which provide accurate information about user engagement.

The premium version of the app, *Mindsurf*, combines an AI-powered conversational agent trained in cognitive behavioral therapy (CBT) techniques with structured tools, including short instructional videos and guided exercises, designed to support short-term emotional regulation and longer-term coping strategies. The AI assistant provides personalized guidance that adapts to users' inputs over time. Core features include guided psychoeducation delivered through written interactions, self-paced exercises to teach how to manage anxiety, depression, stress, and low self-esteem, and a journaling tool that encourages regular documentation and reflection on thoughts and emotions.

We present five main findings. First, we estimate large and persistent average treatment effects on mental health and psychological well-being, our primary preregistered outcomes. The intervention increased a composite mental health index of well-being, anxiety, depression, and stress by 0.30 standard deviations (SD) at one and two months, with effects of 0.23 SD persisting at six months. For context, in-person psychotherapy reduces concurrent symptoms of depression by 0.20–0.60 SD (Cuijpers et al., 2023) and pharmacotherapy by 0.35 SD (Gartlehner et al., 2017).

Second, we find no evidence of adverse effects for any subgroup, including people with severe baseline symptoms. In fact, the prevalence of severe depression and anxiety symptoms is consistently *lower* in the treatment group, alleviating concerns that this tool may be harmful or ineffective for higher-need populations.

Third, we find improvements across a range of pre-registered outcomes. Consistent with the psychological improvements, the frequency of healthful behaviors, exercise, outdoor activities, self-care, and helping others, increases. Importantly, we find positive labor-market impacts, with a 17% reduction in absenteeism in the prior week. Sleep quality also improves (0.22 SD at one month, with effects persisting). We document initial reductions in social-media use, which fade by six months.

We do not find evidence of detrimental effects on isolation or feelings of personal agency; instead, both improve. Feelings of isolation decline substantially (0.22 SD at one month; 0.16 SD at six months), and measures of perceived agency, capturing self-efficacy and locus of control, increase by 0.20 SD at two months and remain positive at six months.

Fourth, access to detailed usage data from the app provider allows us to document user engagement and link it to mental health outcomes. We find high initial engagement followed by a gradual decline: 95% of eligible participants logged into the app at least once during the first two months, but only 37% remained active by week 8, and fewer than 10% by month six. This pattern mirrors engagement dynamics documented in the broader literature of digital mental health tools (Smith et al., 2025). Declining engagement is often interpreted as evidence of limited efficacy or user value; however, it may also arise if the app helps users adopt symptom-relieving skills. Once the skills are adopted, the need for continued app use is lower. In this setting, the presence of large and persistent mental health improvements despite declining engagement is consistent with the latter interpretation. We provide suggestive evidence supporting a behavioral mechanism: healthful practices and sleep outcomes mediate up to half of the average mental health effects during the study period. Moreover, people with bigger mental health impacts also experience bigger impacts in healthful practices and sleep.

Fifth, access to the app crowds in in-person psychotherapy with trained professionals: treated people are 5 percentage points more likely to report attending psychotherapy in the previous month, and this increase alone does not account for the app's overall mental health impacts. Evidence suggests that the app may facilitate engagement with in-person care by reducing symptoms of psychological distress to more manageable levels among people with moderate to severe baseline distress.

This project contributes to several literatures. First, it adds to a literature that studies scalable mental health interventions, such as layman-provided psychotherapy (see, e.g., Angelucci and Bennett (2024, 2026), Angelucci et al. (2026), Baranov et al. (2020), Bhat et al. (2022), Zadey (2023)). This is especially useful given the global shortage of mental health treatment and the high demand- and supply-side barriers to care. Thus, this study establishes both that AI-powered digital mental health care can be an additional scalable intervention to reduce the mental health care gap and that it does not necessarily crowd-out other forms of care. It highlights the promise of low-cost, scalable tools to alleviate psychological distress without reducing the demand for

psychotherapy or increasing the use of social media and the feeling of isolation.

Second, it contributes to the emerging literature on AI-powered mental health interventions by filling key evidence gaps. We present evidence from the largest-scale experiment to date on these tools, randomizing roughly 2,000 participants and providing results from up to six months of follow-up, addressing limitations of prior work that has typically relied on small, selected samples and short study periods (Li et al., 2023, Heinz et al., 2024).[2] We also provide evidence from a low- and middle-income country (LMIC), helping close the evidence gap for the potential of these tools in contexts where most of the global mental health burden lies. Finally, beyond symptom and well-being outcomes, we document behavioral outcomes not previously shown in this literature, including patterns of user engagement, effects on wellness practices and daily functioning behaviors, impacts on use of traditional care, and pathways for sustained impact.

Third, we document a clear decoupling between engagement and effectiveness, with sustained mental-health benefits persisting despite declining product use. This finding has broader implications for how digital and AI tools should be evaluated: low or falling engagement, a common pattern in digital interventions, should not be taken automatically as evidence of low effectiveness or user value. Our evidence is suggestive that adoption of recommended behaviors mediates the sustained effects, implying that brief or intermittent interaction with a product can produce durable benefits. We therefore caution researchers and practitioners against inferring low perceived value or lack of impact solely from discontinued use, and urge the use of evaluation frameworks that measure behavioral outcomes in addition to raw engagement metrics.

Fourth, we document positive spillovers on labor-market outcomes, such as reduced absenteeism. These effects provide causal evidence that improving mental health can increase labor-market attachment, consistent with a negative causal link from mental disorders to poverty (Haushofer and Fehr, 2014, Ridley et al., 2020). These results also speak directly to the puzzle around mixed effects of mental health on economic outcomes in low and middle income countries: although a meta-analysis reports positive aggregate impacts (Lund et al., 2024), several experimental studies from East Asia, mostly with female samples, report null effects (Angelucci and Bennett, 2024, Baranov et al., 2020, Bhat et al., 2022), prompting the hypothesis that the lack of effects might stem from samples being predominantly women. Our finding of positive

[2]The average sample size in the 16 clinical RCTs reviewed by Li et al. (2023) was 140, with most studies measuring outcomes after a couple of weeks.

labor-market impacts in an all-female Mexican sample shows that a predominantly female sample alone cannot account for those null results.

2 AI-Powered Digital Mental Health Care

2.1 The Mental Health App

The intervention consisted of offering free access to a premium version of *Mindsurf*, a commercially available well-being app. This product offers personalized support through a blend of two evidence-based approaches, cognitive behavioral therapy (CBT) and mindfulness techniques. Mindfulness techniques are practical exercises designed to improve emotional regulation and awareness and lower stress, such as breathing exercises, body scans, and mindfulness. CBT treats mental disorders through cognitive restructuring (helping people re-frame dysfunctional thoughts) and behavioral activation (fostering engagement in mood-lifting, meaningful activities).

The app is available on smartphones and at USD $5 per month. At the time of the study, the core features of this premium version consisted of (i) a mood diary, (ii) an AI-powered conversational agent, (iii) audio and guided exercises, and (iv) self-assessment tools. The mood diary was also available in a free subscription.

Mood diary. Each user session begins with a prompt asking, *"How do you feel today?"* (Figure A.1). This feature allows users to monitor their emotional states over time and may also serve as an entry point to more tailored app content, as reported emotional states inform subsequent interactions and recommendations within the app.

AI-powered agent. This app's central feature provides personalized, text-based interactions using natural language processing (Figure A.2). Users may ask questions, describe experiences, or express emotions, and the agent responds with empathetic and contextually relevant prompts. For example, when a user reports elevated stress or low mood, the agent may acknowledge the emotion and prompt reflection or suggest a relevant coping exercise. Responses are in real time and adapt to the user's inputs, allowing for back-and-forth personalized interaction.

Audio and guided exercises. A library of short psycho-educational and skills-based modules covers topics such as anxiety, depression, stress, self-esteem, and emotional regulation (Figure A.3). Individual sessions typically last less than ten minutes and combine short explanations with interactive exercises. For instance, the depression module consists of 15 structured sessions

that introduce coping strategies. The self-assessment and AI agent guide users to self-guided exercises targeting specific concerns, including stress management, sadness, and anxiety.

Self-assessment tools. These tools are based on standardized self-reported screening questionnaires, intended to help users reflect on their emotional well-being and track symptoms over time and not to provide clinical diagnoses (Figure A.4).

2.2 Safety Protocols and Risk Management

To promote user safety and support individuals who may be experiencing elevated distress, the app incorporates several risk management safeguards, particularly with respect to the AI-powered conversational agent, which is trained exclusively for mental health–related interactions and is not designed to respond to general-purpose or unrelated queries. First, the company conducts periodic internal testing of the conversational agent using predefined risk-related phrases to evaluate its responses to potentially sensitive or high-risk content. Second, trained mental health professionals review a subset of anonymized conversations to monitor system performance and identify potential high-risk cases. Finally, a second dedicated AI-based risk detection system continuously scans user-generated text in real time to flag content suggestive of acute distress, triggering review and follow-up by the monitoring team when appropriate. Appendix A.2 provides additional details on the training and safeguards of the AI system.

3 Experimental Design and Data

3.1 Experimental Design

The experiment ran from April to December 2025. Figure 1 depicts the experimental design. Participants were recruited between April and May on a rolling basis through Facebook and Instagram advertisements describing the study as focusing on emotional well-being. The advertisements were designed to resonate with individuals experiencing psychological distress, with the goal of reaching individuals who might benefit from a digital mental health intervention.[3]

Participation in the study required completion of three short surveys. After clicking the ad, potential participants first completed a brief screening survey to assess eligibility. To qualify,

[3]In total, 825,727 users viewed the advertisements, and 31,259 clicked to learn more and begin the screening process. Figure B.1 shows sample ads.

respondents needed to be women aged 18–50, reside in Mexico, have no graduate education, report income below 26,000 MXP (the 7th decile of the national data), and score at least 3 on the PHQ-4, a brief screener for depression and anxiety. Of the 6,690 people who completed screening, 2,923 met the eligibility criteria. These women then completed a baseline survey two days later, followed by a commitment pledge indicating motivation to participate. A third short survey was sent one day later. These steps were meant to reduce long-term attrition, a common challenge in longitudinal studies.

A total of 1,964 participants completed all steps and were enrolled in the study. Enrolled participants were then randomly assigned to a treatment or control group. Half of the treatment group randomly received access to the premium version of *Mindsurf* for six months, and the other half for three months (understand effect persistence and app engagement). The control group received access after six months. Since the second randomization had no effects, we pool the two treatment arms in the main analysis and we discuss the findings from this second randomization in Appendix N). All participants received approximately US$15 for their participation.[4]

3.2 Baseline Data

At baseline, we collected mental health (PHQ-4 for anxiety and depression; WHo-5 for subjective wellbeing, and PSS-4 for stress), socioeconomic information, personality traits, a list of healthful activities, locus of control, self-efficacy, social media use, and psychotherapy use in the previous 12 months.

3.3 Outcome Data

We draw on four sources of data: (i) participant surveys administered at baseline and approximately at one month, two, and six months after after randomization; (ii) weekly affect data; (iii) administrative data (iv) expert predictions collected through the Social Science Prediction Platform (DellaVigna et al., 2019).

[4]We submitted an initial pre-analysis plan prior to the first round of data collection, specifying primary and secondary outcomes, outcome families, and regression specifications. We amended the plan to include six month outcomes, with the amendment submitted prior to launching the final survey wave.

3.3.1 *Survey Data*

We provide a brief description of all our survey outcomes. Appendix **??** provide a more detailed explanation.

Mental health and well-being. Mental health outcomes constitute our primary pre-registered outcome family. We measure four dimensions of mental health using validated self-report instruments: the PHQ-8, measuring symptoms of depression (0–24; scores ≥ 10 suggest moderate to severe depression) (Kroenke et al., 2009); the GAD-7, for generalized anxiety (0–21; scores ≥ 10 indicate moderate to severe anxiety) (García-Campayo et al., 2010); the WHO-5 Well-Being Index, which assesses subjective psychological well-being on a 0 to 100 scale, with scores below 50 indicating poor well-being (Topp et al., 2015); and the PSS-4, for perceived stress (0–16; higher scores indicate greater stress).

Social isolation. Feelings of isolation have been identified as both a correlate and a risk factor for poor mental health (e.g., Holt-Lunstad, 2022). Digital mental health lacks in-person components and may thus amplify social disconnection (DHHS, 2023). We use the six-item De Jong Gierveld Loneliness Scale (De Jong-Gierveld and van Tilburg, 2006). We report outcomes separately for emotional and social isolation, as well as for the overall index denoting less isolation.

Self-efficacy and locus of control. Self-efficacy and locus of control are psychological traits closely linked to resilience, emotional regulation, and overall well-being (Botha and Dahmann, 2024). Individuals with high self-efficacy believe in their ability to manage challenges effectively, while an internal locus of control reflects the belief that life outcomes are shaped by one's own actions rather than external circumstances. We report standardized scores for each scale separately, and a combined index for both scales, which we denote *Personal Agency Index*. Higher values indicate higher agency.

Sleep Quality. Poor sleep quality is both a cause and a consequence of mental distress (Harvey, 2011, Baglioni et al., 2016). To capture patterns in sleep behavior, we use a combination of self-reported measures. In addition to the sleep-related item from the PHQ-8 (*"Trouble falling or staying asleep, or sleeping too much"*), we compute total sleep duration by collecting the people went to sleep the previous night and the time they woke up that morning. We also asked how many times respondents woke up the previous night. We use these outcomes to construct a sleep quality index. Higher values indicate higher quality sleep.

Healthful behaviors. We include three indicators of daily functioning that are likely affected by

better mental health: job absenteeism, helping others, and emotional regulation, as well as three wellness practices widely thought to also benefit mental health: exercising, going out for leisure activities, and self-care (Mahindru et al., 2023, Fancourt et al., 2021). Specifically, respondents reported how many days in the past week they: missed work, raised their voice in anger, helped someone with their homework; and exercised, went out for non-school or work reasons, and spent at least 15 minutes on themselves. We report effects on the number of days for each behavior and a standardized index of healthful behaviors, as specified in our pre-analysis plan. We also create two subindices for daily functioning and wellness practices, since these two subgroups are conceptually distinct. To complement these data, at two and six months we also asked respondents how often they took the following actions the previous week: being kinder to self, setting boundaries, doing breathing exercises, adopting a sleeping routine, and journaling. These are all actions recommended by the app. We report the effects on the number of days, as well as a standardized index. Higher values correspond to more frequent (and more) actions.

Social media use. Social media use has been shown to decrease subjective well-being and mental health (Allcott et al., 2020, Braghieri et al., 2022). In our context, access to the app could plausibly affect social media use in *a priori* ambiguous ways. App access may reduce social media engagement by improving mental health or displacing the time that would otherwise be spent scrolling. On the other hand, the app access may increase phone use and exposure to social media. We measured days of social media use (Facebook, X, Instagram, and TikTok) in the previous week. We also measure average daily time spent on social media, from which we constructed an indicator for more than 2 hours of use per day. We report extensive margin measures and the likelihood of high-intensive margin use, as well as a combined use index.

Effort and emotional executive function games. To measure cognitive effort and emotion-related attentional control, we implemented two incentivized tasks. First, we asked participants to count the number of zeros in matrices of zeros and ones under time constraints. This is a standard measure of willingness to exert cognitive effort in repetitive, attention-demanding tasks (Abeler et al., 2011). We incentivized participants to complete as many matrices as possible within one minute via a lottery (each correct response gained one lottery ticket). Second, we implemented an emotional Stroop task to assess attentional bias and cognitive control in the presence of emotionally salient stimuli (Williams et al., 1996). Participants identified the color of emotionally charged or neutral words presented in separate blocks. Longer reaction times to emotional

words are commonly interpreted as reflecting difficulty disengaging from emotionally relevant stimuli. We adapted the task to Spanish using lexicographically similar words, and incentivized participants by offering lottery tickets proportional to their total correct answers within a fixed time window. Appendix C provides further details.

Psychotherapy use. We asked participants whether they had attended any sessions with a psychologist, psychiatrist, or therapist in the last 30 days. In the last survey round, we also asked respondents how many sessions they had attended each month up to six months before. We also collected information on perceived stigma surrounding psychotherapy, as well as self-reported measures of knowledge about and perceptions of psychotherapy.

Stigmatizing beliefs about people with mental illness. At 2 and 6 months, we collected beliefs about negative traits associated with people with mental illness. We measured both first-order beliefs (respondents' beliefs about people with mental illness) second-order beliefs (respondents' beliefs about their neighbors' belifs about people with mental illness). In both cases, we measure beliefs on a scale from 1 ("totally disagree") to 5 ("totally agree"). The standardized index has higher values corresponding to *less negative* beliefs.

3.3.2 *Weekly Affect Data*

Another unique feature of our study is that we collected weekly mood data. Beginning one week after app access, we sent all participants weekly WhatsApp messages asking about current mood. People could reply by selecting a happy, neutral, or sad emoji. These responses provide a high-frequency proxy for emotional state, allowing us to track mood dynamics over time. From these data, we constructed mood indicators. Appendix D provides details on these data and shows that there are high and statistically significant correlations between affect and the survey mental health outcomes (Table D.1).

3.3.3 *Administrative Data*

A unique feature of our study is the access to detailed administrative data to measure participants' app use, the number of sessions completed, total time spent on the app, and engagement with specific app features. To protect participant privacy, the research team did not have access to the content of individual AI-agent interactions.

3.3.4 *Expert Predictions*

To assess to what extent our results change the views of the scientific community, we solicited the opinions of experts on economics and mental health through the Social Science Prediction Platform (DellaVigna et al., 2019). The prediction survey was completed by 126 individuals.

4 Identification and Estimation of Intent to Treat (ITT) Effects

We estimate Intent to Treat (ITT) effects, defined as the average causal impact of app access, by estimating the following equation:

$$Y_i = \alpha + \beta \mathrm{T}_i + \gamma X_i + \varepsilon_i, \tag{1}$$

where Y_i is the outcome for individual i, the indicator T_i equals one if the individual was assigned to receive immediate app access, and X_i represents a list of predetermined covariates selected through LASSO following Chernozhukov et al. (2018) and adding strata indicators and the baseline dependent variable (when available) among the set of covariates (Cilliers et al., 2024). Our parameter of interest, β, identifies the average treatment effect of app access under the assumption that the randomization solves the missing counterfactual problem and that the SUTVA holds. To control for multiple hypothesis testing, for each family of outcomes, we create indices standardized around the control group mean (Anderson, 2008). We estimate treatment effects separately at one, two, and six months after randomization, and use heteroskedasticity-robust standard errors. When we pool data rounds, we cluster standard errors by participant.

5 Sample and Descriptive Statistics

5.1 Sample

Of the 6,690 individuals who completed the first screening survey, 44% (N=2,923) met the eligibility criteria. Among the eligible, 67% (N=1,964) completed all three pre-treatment surveys and were randomized (Appendix Table E.1 summarizes sample sizes and characteristics at each stage of recruitment). Observed characteristics are largely similar across recruitment stages, with differences concentrated in dimensions explicitly used for screening, particularly psychological distress, education, and household income. In contrast, characteristics such as age, fertility,

marital and labor market status, and personality traits remain similar across the screened and the final sample.

The final sample (N=1,964) consists of adult women with at least moderate levels of psychological distress (PHQ-4$\geq$3), recruited nationwide across all of Mexico (Appendix Figure E.1). On average, participants are in their late thirties, have approximately one child, and roughly half are engaged in paid work. Approximately 69% report having visited a psychologist at least once, and 31% report having done so in the prior year. Taken together, the sample represents women experiencing mental health distress who are familiar with, but not continuously engaged in, formal mental health care.

5.2 Balance and Survey Attrition

Table F.1 shows that the pre-determined characteristics of the control and treatment groups are balanced at baseline: we cannot reject the hypothesis that the coefficients are jointly identical across the two arms. Table F.2 shows no evidence of systematic attrition by treatment status. Survey completion rates were high, with 95% and 94% of participants completing surveys at 1 and 2 months, respectively (columns 1 and 2), and 82% of participants completing surveys at 6 months (column 3).

6 First Stage and Average Treatment Effects

6.1 App Take-Up

In the treatment group, 85% downloaded the app on their phones. Of these, 95% used the app at least once during the study period. Figure 2 shows that 57, 36 and 10 percent of the treatment group used the app at 1, 2, and 6 months. Section 7 will describe patterns on engagement in details. While disengagement from digital mental health care tools is a common concern (Torous et al., 2020b), patterns of declining use are not unique to digital interventions. In-person psychotherapy exhibits similar dynamics, with “premature" dropout rates of up to 50% (McGovern et al., 2024). Importantly, declining engagement does not necessarily imply limited impact or value: some users may discontinue use after achieving desired improvements. We return to this issue in Section 8.

6.2 Effects At a Glance

Figure 4 presents average intent-to-treat estimates of the treatment effects on standardized indices for each family of outcomes, where positive coefficients indicate improvements. The mental health index, our primary pre-registered outcome, improved by 0.3 SD at one and two months and by 0.2 SD at six months. We fail to reject the hypothesis that these three coefficients are statistically identical (p-value>0.05). Pairwise comparisons show a statistically significant differences between impacts at 2 and 6 month estimates, indicating a sustained, yet declining improvement over the study period.

Overall, there are statistically significant and positive impacts in all the other families, with effect sizes generally ranging between 0.1 and 0.3 SD. While some estimated effects are largest at the one-month mark, below we will show that they are often statistically indistiguishable from impacts at 2 and 6 months, confirming that effects persist beyond the initial period. The results for all families of outcomes are qualitatively unchanged if we restrict the sample to participants who respond to all surveys, as shown in Appendix G.2.

6.3 Effects on Mental Health

Effects on Depression, Anxiety, Subjective Well-Being, and Stress. Table 1 shows that receiving access to the app significantly improves all measured dimensions of psychological well-being, our primary pre-registered set of outcomes, for up to 6 months. Across all follow-up periods, we find statistically significant improvements on each individual mental health scale. At 1 month, treated participants report higher subjective well-being (WHO-5) and lower levels of depressive symptoms (PHQ-8), anxiety (GAD-7), and perceived stress (PSS-4). These improvements persist at 2 months and remain statistically significant at the 6-month follow-up.

Benchmark. To benchmark these effect sizes with the impact of other evidence-based treatments for anxiety and depression, Figure 5 shows impacts in terms of standardized effect sizes. Psychological well-being (WHO-5) increased by 0.3 SD at two months; the severity of depression (PHQ-8), generalized anxiety (GAD-7), and perceived stress (PSS-4) declined by 0.3 SD, 0.2 SD, and 0.25 SD. We compare these impacts to the broader literature on digital self-help mental health tools. Studies of AI-based digital interventions also find comparable effects of 0.3 SD on psychological well-being at one month and larger improvements of 0.7 SD on psychological

distress at 1 month (Li et al., 2023). However, most studies use college students or other highly selected populations, have small samples, and follow participants only for a few weeks. For example, the studies reviewed in Li et al. (2023) have a median sample of 94, a median duration of 1 month, and primarily recruit college students from the USA, the UK, or Japan. Interestingly, the impact of *Therabot* (a generative AI agent) on depression is a reduction of 1.5 points on the PHQ-9 scale (Heinz et al., 2024), closely matching our PHQ-8 reductions of 1.5–1.6 points.[5]

These effect sizes are also comparable to improvements achieved through more traditional mental health care, providing a useful benchmark for interpreting magnitudes. For example, meta-analytic evidence from mainly high-income countries indicates that CBT reduces depressive symptoms by approximately 0.22 SD in a meta-analysis of meta-analyses (Gartlehner et al., 2017), with more recent estimates ranging from 0.47 SD after correcting for publication bias to 0.60 SD when restricting attention to low-bias studies (Cuijpers et al., 2023). The magnitude of the effects documented here is therefore notable, given the light-touch, short-duration, and non-clinical nature of the intervention evaluated. Standard in-person CBT typically lasts between 1.5 and 5 months and, thus, is more expensive and time-consuming than app-provided digital mental health tools. A more relevant comparison may be with scalable alternatives such as short courses of layman-delivered, trans-diagnostic group therapy. Angelucci and Bennett (2026) and Angelucci et al. (2026) evaluate such interventions among adult women in Mexico and India, finding improvements in psychological distress of 0.2–0.3 SD, comparable to the effects documented here.

Distributional Impacts. One recent concern regarding the use of large language models (LLMs) to provide psychological support is the possibility that AI-generated responses may offer unsuitable advice or induce severe distress in some users. Although we do not observe the content of interactions between users and the AI agent, we can assess this concern by comparing the distribution of mental health outcomes across treatment and control groups. If exposure to the app were to exacerbate distress for a substantial share of users, we would expect to observe a higher prevalence of extreme outcomes, severe depression, anxiety, stress, or very low well-being, among treated individuals.

Figure 3 shows that this is not the case. In the pooled data, treated participants are no more likely than control participants to experience severe depression, severe anxiety, high stress, or

[5]This is the only comparable outcome. Heinz et al. (2024) measure the PHQ-9, which has a maximum score of 27, while this study uses the PHQ-8, which omits one of the PHQ-9 questions and has a maximum score of 24.

very low subjective well-being. If anything, the prevalence of extreme psychological distress is higher in the control group. Moreover, the treatment substantially reduces clinically relevant symptoms: access to the app lowers the prevalence of at least moderate depression by 12.7 pp (50.0% vs. 62.7%; $p < 0.01$) and at least moderate anxiety by 10 pp (34.0% vs. 44.0%; $p < 0.01$). We observe similar patterns for perceived stress and well-being.

High-Frequency Affect Data. We complement the 1, 2, and 6 month mental health data with weekly data. These data show a similar pattern as the survey data: the effects on happiness are highest for the first two months and then they partially decrease, although impacts remain positive and significant throughout the study period. Conversely, there are reductions in sadness and neutral feeling. These decreases are largest in the first half of the study period and gradually dissipate. Beyond validating the survey mental health outcomes, these high-frequency data show that the mental health benefits occur from very early on: the rate of happiness is 10 percentage points higher in the treatment group in week 1 and peaks at 20 percentage points in weeks 4 to 6. These very short-term benefits are consistent with the literature on digital mental health interventions (Li et al., 2023).

6.4 Effects on the Components of the Other Families of Outcomes

Effects on Healthful Behaviors. Table 2 documents improvements in several self-reported healthful behaviors, both overall and for daily functioning and wellness practices separately. Pooling across waves, we find a reduction of 0.08 days of missed work per week ($p<0.05$), a 17 percent decline relative to the control group mean. This finding is consistent with the reduced absenteeism of severely depressed patients treated with a short course of psychotherapy (Patel et al., 2017), and, more broadly, with positive labor market impacts of mental health interventions in LMICs (Lund et al., 2024). App access also improves altruistic behavior and emotional regulation. Overall, there is a positive 0.17 SD impact on the daily functioning index. Similarly, we find impacts of similar magnitudes for wellness practices (0.21), driven by all its components: app access increases the frequency of exercising, going out, and self-care. Aggregating across all six outcomes, the pooled standardized healthful behavior index increases by 0.26 SD ($p<0.01$), with higher impacts at one month (0.34 SD) and a partial attenuation, but still positive impacts, at two and six months (0.19 and 0.20 SD).

Effects on Sleep. Table 3 reports treatment effects on three sleep measures: hours slept (calculated from self-reported bedtime and wake-up time rather than a direct subjective assessment); the number of sleep interruptions in the previous night; and whether the respondent reports trouble sleeping. At one month, treated participants sleep approximately 13 additional minutes per night and are significantly less likely to report trouble sleeping, while the reduction in sleep interruptions is negative but not statistically significant. The standardized sleep index improves by 0.22 SD at one month. These improvements persist at two and six months, and tests of equality indicate that we cannot reject the null of equal effects over time for both the composite index and the individual components.

For comparison, Bessone et al. (2021) find that providing sleep aids and encouragement to low-income adults in India, an intervention explicitly targeting sleep, increased objectively measured sleep duration by 27 minutes per night. The gains we document are roughly half as large.[6]

Effects on Social Media Use. Table 4 show lower use of Facebook, X, and Instagram, with the largest reductions observed for Instagram. These platform-specific effects attenuate over time and are no longer statistically significant at the six months for most platforms. The standardized social media index (where positive coefficients denote less overall use) shows an improvement of 0.09 SD ($p<0.05$) at one month and 0.08 SD ($p<0.10$) at two months, but the effect is no longer distinguishable from zero at six months. We can reject equality between early and late follow-up periods for the index, consistent with a fade-out of effects over time. Thus, app access led to temporary reductions in social media use, with no evidence of sustained changes at longer horizons.

Effects on Measures of Isolation and Personal Agency. Table 5 shows that app access provides sustained reductions in both emotional and social isolation, with persistent impacts of about 0.2 SD at one, two, and six months. Table 6 sustained impacts on both self-efficacy and locus of control, with aggregate effects of 0.15-0.20 SD at one and two months, which attenuate to 0.10 SD at six months.

Effects on Effort and Attention Games. Appendix Table G.1 shows no statistically significant

[6]Our sleep duration measure is self-reported, derived from bedtime and wake-up time. Bessone et al. (2021) find that self-reported time in bed closely tracks actigraphy measures, though self-reported sleep duration overestimates objectively measured sleep. While this overestimation may affect the level of reported sleep, it is unlikely to systematically differ across treatment and control groups. Moreover, treatment effects on sleep in that study operated primarily through increased time in bed, suggesting that this measure may at least capture a margin along which sleep improvements would occur.

effects of app access on performance in the matrix effort task (column 1). In contrast, we find improvements in performance on the emotional Stroop task at the one month (Panel A, column 3), with no corresponding effects on placebo neutral words. These effects do not persist at two or six months. One possible explanation is that early improvements in attentional control are most salient when baseline psychological distress is higher.

6.5 Experimenter Demand Effects

Our results are unlikely to be driven by experimenter demand effects, as we show with four tests. First, treatment effects extend to outcomes that are less transparently manipulable, such as sleep duration constructed from reported bedtimes and wake-up times, or the emotional stroop test. Second, we find no evidence that treatment effects are concentrated among individuals more susceptible to social desirability at baseline (Dhar et al., 2022). Third, we find no treatment effect on respondents' perceived value of participating in the study. Lastly, in an additional experiment, explicitly informing randomly selected respondents about the study's main hypothesis does not lead to larger effects (De Quidt et al., 2018). Appendix J provides further details of these tests.

7 Patterns of Product Engagement

A valuable feature of our study is that we have access to administrative data. Thus, we do not need to rely on user self-reports for product use, and we can measure their patterns of use in detail. We exploit these data to explore how use of the four main app features (assessment, AI conversational agent, exercises, and mood diary) varies over time. . Figure 7 computes active users' weekly minutes, total and by app functions. First, conditional on positive use, intensity declines from roughly 55–60 minutes in week 1 to approximately 15 minutes by week 10 and then stabilizes between 10 and 18 minutes per week through week 26. This pattern indicates that the subset of users who remain engaged continue to devote non-trivial time to the application. Second, the composition of use across app functions changes over time. In the first week after enrollment, the use of self-assessment tools (5 minutes), AI conversations (20 minutes), and audio/guided exercises (20 minutes) are highest. Self-assessment stops by week 3, while use of the other two features gradually declines through approximately week 10 and then stabilizes at 5–8 minutes per week for the AI agent and 4–7 minutes for guided exercises. By contrast,

the mood diary exhibits a flatter trajectory over time, remaining relatively stable at roughly 3–5 minutes per week. Taken together, these patterns are consistent with a shift from assessment and learning towards maintenance-oriented use.[7]

Figure H.2 reports different dimensions of patterns of engagement. Approximately 55–60% of active days begin between 9:00 pm and 8:00 am on weekdays (Panel A), about 30% of sessions start between 9:00 pm and midnight (Panel B), and 25% occur during the weekend (Panel C). That is, a substantial share of app use occurs during times when access to in-person psychotherapy would be severely limited. Panel D shows the SD of hours of use within the day and the coefficient of variation (CV) days of use within the week.[8] These outcomes are generally zero for traditional psychotherapy, as people often schedule meetings at the same time and weekday. Instead, this figure shows high variability along both dimensions: 58% of the sample has highly variable days of use (CV>1) and 90% of the sample has highly variable hours of use (SD>3). Thus, most people use the app at different times of the day and on different days of the week. These patterns are consistent with flexible, opportunistic use rather than structured appointment-style scheduling.

The possibility of this type of flexible use could be important, given that symptom severity exhibits substantial intra-individual variability (Ebner-Priemer and Trull, 2009). The app can provide immediate support during moments of distress without requiring a scheduled appointment. Indeed, the literature acknowledges that scheduled psychotherapy sessions alone may be insufficient and that patients need between-visit tools (Kazantzis et al., 2010), and that traditional care models create a timing mismatch between the availability of and need for a provider (Mohr et al., 2014).

8 Use–Response Dynamics in Digital Mental Health Care

Disengagement from digital tools is widely documented across domains such as education and health, and is often interpreted as evidence of limited product value. However, declining engagement can also reflect a learning-based mechanism: digital tools may deliver information,

[7]Figure H.1 show the share of people who use the app at least 20 minutes per week, also showing declines in user frequency over time.

[8]We compute the CV (=SD/mean) for days of the week because we label weekdays from 1 to 7 and we want to capture regularity of use. A user who connects on days 1 and 5 of the week has a higher variance than a user who connects on days 1 and 3, but both have regular use. The CV normalizes this scale difference, yielding a comparable measure of regularity regardless of how often the user connects.

skills, or behavioral strategies that users internalize over time, reducing the need for continued interaction once the relevant knowledge has been acquired.

This interpretation is particularly relevant in the context of mental health interventions grounded in skill acquisition. Cognitive behavioral therapy (CBT), on which the app evaluated here is based, aims to improve psychological well-being by helping individuals identify and modify maladaptive thought patterns and behaviors. Once these cognitive and behavioral changes are internalized, the marginal benefit of continued app engagement declines. Consistent with this model, a standard course of CBT typically lasts between 1.5 and 5 months and is associated with sustained effects for at least six months (Van Dis et al., 2020).

If continued app use were essential for sustained mental health improvements, we would expect little or no improvement at two and six months among individuals who stopped using the app after initial exposure. Conversely, if sustained benefits do not require continued use, psychological well-being should remain elevated even among individuals with minimal or no subsequent app engagement. Moreover, if improvements in mental health operate in part through behavioral changes, we would expect (i) sustained improvements in psychological well-being, sleep, and healthful behaviors despite declining app use, (ii) evidence that sleep and healthful behaviors mediate the effects of the intervention on mental health, (iii) evidence of sustained increases in the use of tools recommended by the app, consistent with users internalizing cognitive and behavioral strategies.

The study's experimental design does not allow us to identify app-driven adoption of the skills and behaviors cause improvements in mental health. Nevertheless, the persistence of mental health gains despite declining app engagement suggests that continued use may not be necessary for sustained benefits. Below, we present several complementary pieces of evidence that are consistent with this interpretation.

First, we have already established that impacts on psychological well-being, sleep, and healthful behaviors are positive, statistically significant, and persist through the six-month follow-up. This pattern rules out a simple interpretation in which disengagement leads to a concurrent fade-out of mental health benefits and instead indicates that key outcomes continue to move in the right direction even when app use is minimal, e.g., in months 5 and 6, when only 10-20% of the sample uses the app.

Second, Table K.1 reports effects on the reported use of tools recommended by the app

measured as the number of days in the past week respondents report using each tool, measured at 2 and 6 months. Treated participants report significantly greater use of a broad set of cognitive and behavioral tools commonly taught through cognitive behavioral therapy (CBT), including being kinder to oneself, setting boundaries, practicing breathing exercises, writing about thoughts and feelings, and following a sleep routine. This evidence suggests sustained adoption of app-recommended practices, consistent with users internalizing skills that may reduce the marginal value of continued app engagement.[9]

Third, we conduct a mediation analysis to quantify the extent to which improvements in sleep and healthful behaviors account for the overall treatment effect on mental health. We construct a standardized index combining relevant sleep and behavioral outcomes following Anderson (2008), and estimate the average Natural Indirect Effect (NIE), the Natural Direct Effect (NDE), and the ratio NIE/ITT, where $NIE + NDE = ITT$. Identification relies on standard assumptions of cross-world independence and common support (Imai et al., 2010a,b).[10]

Table K.2 reports estimates of these parameters by survey round. Both the NIE and NDE are statistically significant at one, two, and six months. Improvements in sleep and healthful behaviors account for 49%, 38%, and 45% of the average ITT effect at one, two, and six months.

Finally, we examine heterogeneity by estimating Individualized Average Treatment Effects (IATEs) for mental health, sleep, and healthful behaviors, and studying their relationship (Athey and Wager, 2019, Chernozhukov et al., 2025). Mental health IATEs range from 0 to 0.5 SD, behavioral IATEs from 0 to 0.5 SD, and sleep IATEs from –0.05 to 0.4 SD, with effects positive for nearly all individuals. The fact that mental health impacts are generally positive also for non-active users is consistent with the notion that app use and mental health benefits are eventually decoupled. Table K.3 shows that individuals experiencing larger improvements in sleep and healthful behaviors also experience larger mental health gains.Early app engagement is weakly but statistically significantly correlated with mental health improvements at one and two months, but not at six months.

Taken together, these findings are consistent with a skill or behavior adoption-based mech-

[9]Since we did not pre-specify these outcomes, we excluded them from our main behavior index analysis.

[10]The Natural Indirect Effect (NIE) captures the portion of the treatment effect on mental health that operates through improvements in sleep and healthful behaviors, while the Natural Direct Effect (NDE) captures the remaining effect operating through other channels. By construction, $NIE + NDE = ITT$, and the ratio NIE/ITT measures the share of the total effect explained by the mediators, under the identifying assumptions of the mediation framework. We condition on baseline mental health, personality traits, geographic location, age, education, household income, marital and employment status, parity, and absenteeism to support the cross-world independence assumption.

anism: the app improves mental health partly by promoting cognitive and behavioral changes in sleep and healthful behaviors, which persist beyond active engagement. This interpretation closely aligns with the CBT literature, which emphasizes cognitive restructuring, behavioral activation, and skill practice as key drivers of sustained symptom reduction (see, e.g., Cuijpers et al. (2013, 2019, 2023), Kazantzis et al. (2018)).

The idea that app-promoted behavioral changes lead to protracted improvements in mental wellbeing after the app use is discontinued is consistent with a findings from psychology and behavioral science: behaviors are more likely to become habitual when their benefits are immediate, tangible, and experienced at the time of action, because early rewards strengthen learning and repetition through both reinforcement and motivation channels (Lattal, 2010, Woolley and Fishbach, 2017, 2018). The benefits from using the app arise from very early on, as shown by the high-frequency mood data. This may motivate people to continue to adopt the tools presented in the app, enjoying protracted mental health benefits, without needing to use the app.

9 Who Benefits from the App?

Conceptually, it is unclear which type of participant would benefit the most from the app in general, and how baseline psychological distress correlates with impacts in particular. This is because it is difficult to predict both who will use the app the most (selection effect) and who will have the biggest mental health impacts conditional on use (causal effect). For example, people with the most severe mental needs at baseline mechanically have the biggest scope for benefiting from the app. However, they likely also face the highest behavioral barriers to use (ambiguous selection). At the causal level, the features of users who benefit the most from a mental health care app are not well-established (unclear impacts). We investigate these issues in our sample.

9.1 Baseline Distress and App Use

To begin with, we want to understand which predetermined mental health, personality, and socioeconomic variables predict product engagement. To do so, we compute total engagement in minutes over six months and rank the correlates of app use in order of importance (defined as relative usefulness for prediction) using a random forest algorithm for treatment group members. Each predictors' importance is scaled relative to the most important predictor, whose importance

is normalized to 1. Figure L.1 show that mental health and personality traits are the strongest predictors of app use, while socioeconomic characteristics such as income, education, fertility, and marital and employment status are less important (conversely, age and absenteeism are important predictors of app use).

Random forest algorithms predict importance, but are not informative about the type of relationship between predictor and outcome. To understand how much time people with different baseline psychological distress use the app, we estimate this relationship non-parametrically, using a kernel-weighted local regression. Figure L.2 shows that app use generally varies non-monotonically between 100 to 230 total minutes with respect to baseline distress: use is lowest for people in the top quartile of the mental health index, who have the least need for mental health care, followed by people in the lowest quartile, who have the most need for mental health care, and highest for two middle quartiles. However, the between-quartile differences are not very large. These findings are loosely consistent with evidence that worse mental health generally increases perceived need but can impede follow-through unless engagement is facilitated (Westra et al., 2009).

9.2 Baseline Distress and Treatment Effects

We perform a similar exercise using the estimated mental health IATEs instead of app use: first, we rank the correlates of mental health impacts, and then show the non-parametric relationship between impacts and baseline mental health. Figure L.3 shows that baseline mental health and personality traits are more important predictors than socioeconomic covariates, echoing the findings for product use. Conversely, Figure L.4 shows that the estimated mental health impacts are biggest for people with the highest baseline distress, decline (generally monotonically) for people with lower distress, and are smallest for people with the least baseline distress. A meta-analysis finds similar effects of in-person care (Bower et al., 2013).

Overall these patterns of engagement and impacts suggest that the app is effective at engaging users with the highest needs, who also experience the largest mental health improvements. These findings also re-iterate our conclusion that there is not a strong dose-response function between product use and mental health impacts: the people with the biggest impacts are not also the most active users.

10 Does Digital Therapy Crowd Out Psychotherapy?

A common concern in the expansion of digital mental health tools is whether such interventions crowd out in-person care. The relationship between access to digital tools and demand for person-led psychotherapy, however, is theoretically ambiguous and may vary across individuals and over time. This ambiguity arises because digital mental health tools can affect both the *need for* professional care and the *propensity to* seek it.

We highlight three main channels through which the app may affect demand for psychotherapy. First, the app improves mental health. This may reduce the demand for psychotherapy for some (e.g., the ones whose symptoms have become mild or negligible), and increase it for others (e.g., the ones who have still clinical levels of distress, but whose symptoms have improved enough to be more manageable). For example, this could happen if the app reduce barriers to help-seeking by improving psychological functioning such as cognitive load, emotional overwhelm, or executive dysfunction. These people still *need* care, and now they also have a higher *propensity* to seek it. Second, the app may reduce negative beliefs about mental health, which could increase willingness to seek in-person care. Third, the app may increase knowledge about evidence-based mental health care and how to access it, lowering informational frictions and raising demand. The first channel has ambiguous net effects on demand, while the latter two channels unambiguously predict increased demand for psychotherapy.

The first channel predicts that psychotherapy use is less likely to increase among people with mild baseline distress (who likely have no longer need for psychotherapy after the app improved their mental health). Conversely, the app may crowd in psychotherapy for people whose symptoms were moderate or severe at baseline, by making their conditions more manageable. The second and third channels predict that psychotherapy use increases the most for people whose beliefs and knowledge improve the most.

We begin by showing average impacts. Table 7 shows 3, 5, and 8 percentage point increases in the likelihood of seeing a psychologist at 1, 2, and 6 months ($p<0.1$, <0.01, and <0.01). These are non-negligible impacts over a control group mean of 0.15. The increases in number of psychotherapy sessions is proportionally even larger. Thus, access to the app acted as a

complement rather than a substitute for traditional psychotherapy in our sample.[11]

Next, we examine how impacts on demand, mental health, beliefs, and information about care vary by baseline psychological distress. To do that, we group individuals into mild, moderate, and severe categories using PHQ-4 cutoffs of 3–5, 6–8, and 9–12, respectively. Table 8 pools observations across the three survey waves to increase precision and reports both subgroup-specific estimates (and tests of their equality) and the overall ITT effect, clustering standard errors by participant. Column 1 shows that increase in psychotherapy is concentrated among people with moderate and severe baseline distress, for whom we estimate effects of 0.07 ($p<0.01$) and 0.12 ($p<0.01$) pp. The effect on people with mild baseline distress is smaller and statistically insignificant, at 0.02 pp, with a statistically significant difference between the effects on people with mild and severe distress ($p<0.05$). Columns 2 and 3 show that app access reduces symptoms of depression and anxiety for all groups, but especially for people with moderate baseline distress. These results are consistent with the first channel: the app likely increases the demand for psychotherapy by making symptoms of depression and anxiety more manageable for some people.[12] Column 4 shows that first-order beliefs about mental illness do not change, possibly because negative views of mental illness are uncommon in this sample. Conversely, second-order beliefs improve (Column 5), but by a small amount and not more among people with moderate and severe baseline distress.[13] Thus, the higher demand for psychotherapy is unlikely to be driven by less negative views about mental illness. Lastly, column 6 shows that knowledge and views about psychotherapy do not change for any subgroup. Thus, it is also unlikely that this channel led to higher use of psychotherapy. To conclude, we interpret the available evidence as supportive of the view that app access may crowd in psychotherapy by alleviating mental health symptoms for people with moderate and severe baseline distress.[14]

[11] Although the app included referral options for online therapy (which involved out-of-pocket costs) as well as general information about accessing care, we do not detect any increase in therapy bookings made through the app during the study period. Consistent with this, Appendix Figure **??** shows that treated participants are no more likely than control participants to report having found their current psychologist through the app.

[12] We find similar impacts on stress and wellbeing, but we prefer to focus on instruments with a clear clinical interpretation.

[13] Appendix Tables M.2 and M.3 reports average impacts on all sub-components.

[14] The results are qualitatively unchanged if we group people by their baseline mental health index tercile. For this exercise, we preferred to group people by their baseline PHQ-4 because this is a common screening tool to identify people with likely anxiety and depression, who, is clinical, are offered psychotherapy or pharmacotherapy. There is no analogous protocol for PSS-4 and WHO-5, the other two outcomes that we use to create the baseline mental health index.

Does Psychotherapy explain the mental health impacts? Since access to the app increases the use of traditional psychotherapy, we examine whether psychotherapy mediates the observed mental health impacts. To do so, we conduct a mediation analysis using attendance at psychotherapy in the previous month as the mediator and estimate the Natural Indirect Effect (NIE), the Natural Direct Effect (NDE), and the NIE/ITT ratio. Table M.4 shows that traditional psychotherapy accounts for at most 5% of the average ITT effect at the two-month follow-up ($p < 0.05$), with smaller and statistically insignificant mediation effects at one and six months. These findings suggest that psychotherapy use explains only a limited share of the overall mental health improvements and are inconsistent with the interpretation that the app improves mental health primarily by inducing attendance at traditional psychotherapy.

As an additional check, Table M.5 compares the mental health IATE estimates for the same two subgroups. It shows that treatment group members who went to a psychotherapist the previous month do not have systematically different mental health impacts. This analysis, though methodologically different, confirms the previous conclusions.

Lastly, Figure M.1 restricts the sample to individuals who did not attend psychotherapy during the study period and re-estimate the main treatment effects. Although this exercise conditions on a post-treatment outcome and should therefore be interpreted with caution, we find treatment effects of similar magnitude to those in the full sample. This pattern further suggests that the mental health benefits of the app are not driven primarily by increased engagement with traditional psychotherapy.

11 Expert Predictions.

Table O.1 shows that, while expert predictions are systematically more accurate than non experts', all forecasts understate the magnitude of the estimated impacts. Experts predict mental health increases of 0.13 SD at 3 weeks and 0.12 SD at 8 weeks (whereas the 1- and 2-month effects are about 0.30 SD). A similar pattern holds for healthful habits, with forecasts of a 0.12 SD improvement at 3 weeks, compared to an estimated effect of 0.33 SD. This is partly due to a lower predicted app use in the first three weeks (61% vs 95%), although the predicted use at week 8 is very close to the actual one (33% vs 37%).

Lastly, most people do *not* correctly predict the observed impacts on psychotherapy use and

social media use, and only 60% of experts correctly predict reductions in social media use. Thus, we conclude that our results are surprising relative to priors in the research community.

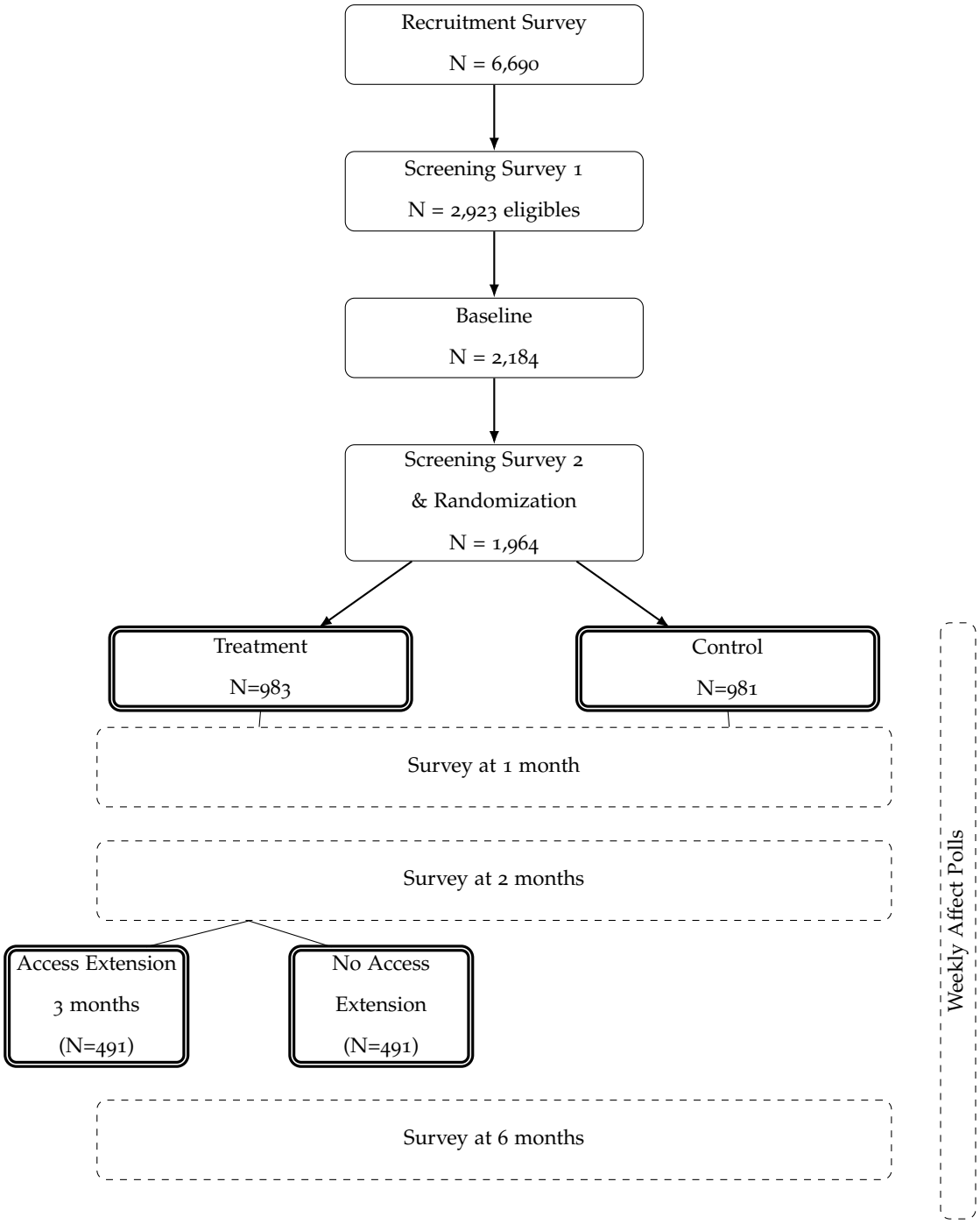
12 Conclusions

The prevalence of psychological distress is high and rising worldwide, yet the supply of mental health care lags behind its demand, causing an unmet need for mental health care. Digital care delivered via app and aided by artificial intelligence could be a way to alleviate human suffering, provided that this new technology is safe and effective and does not lead to isolation or crowd out traditional evidence-based therapy when available.

We present experimental evidence on the impact of an AI-powered mental health app among women with psychological distress and find that gaining access to the app led to rapid, sizable, and sustained well-being gains, without apparent harm to users and without displacing traditional mental health care. Existing evidence about the impacts of digital care delivered through apps is limited in size and does not address the questions of broader behavioral impacts, effect duration, and interactions with traditional therapy. This study addressed these gaps and showed that delivering mental health care via AI-powered apps can be a safe and effective tool to both decrease psychological distress and improve healthful behaviors after a few weeks of use, with effects that outlast product use and without crowding out psychotherapy. Given the low fixed costs and negligible marginal costs of this type of product, there may be scope for scaling the supply of AI-powered mental health care apps to digitally literate women with mild to severe psychological distress.

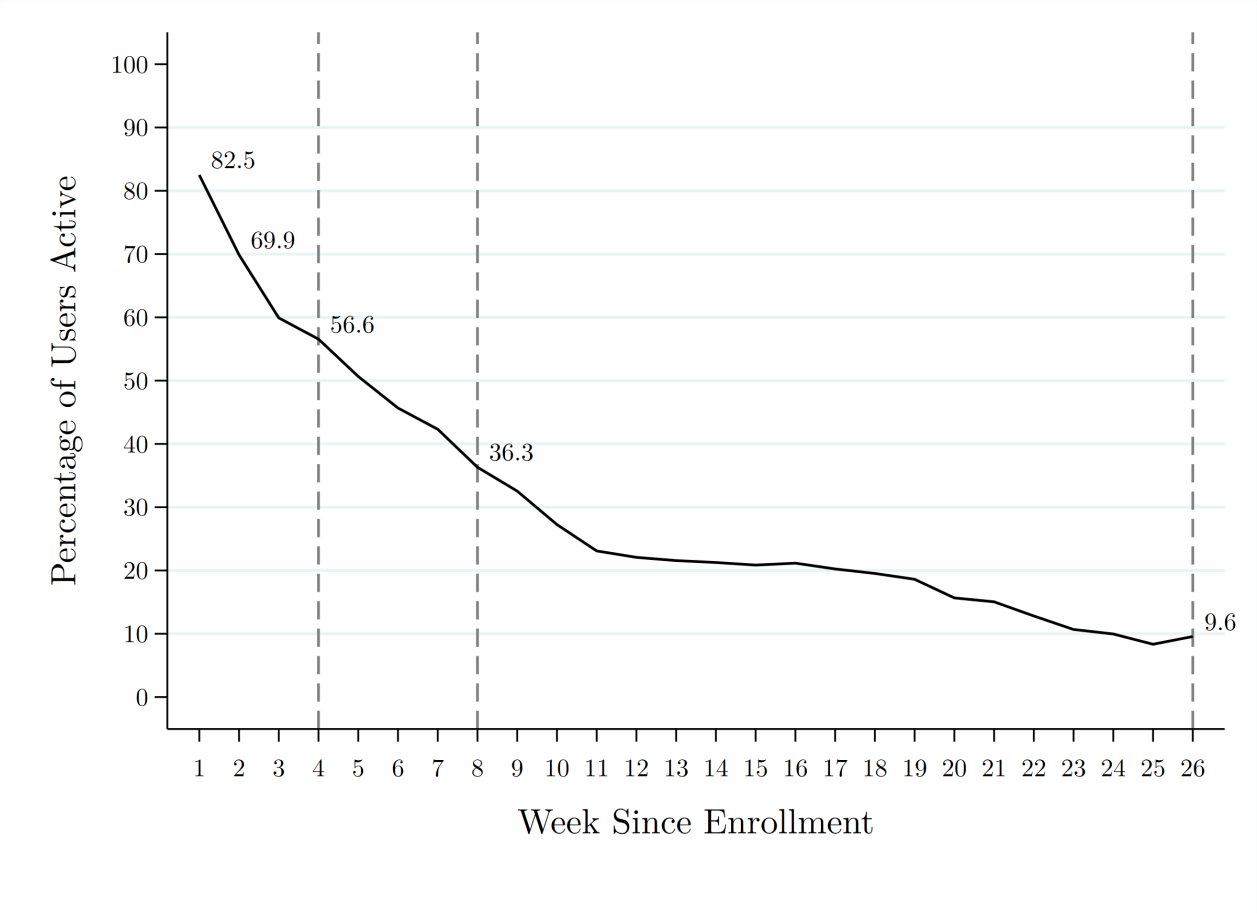
13 Tables and Figures

Figure 1: Experimental Design and Timeline



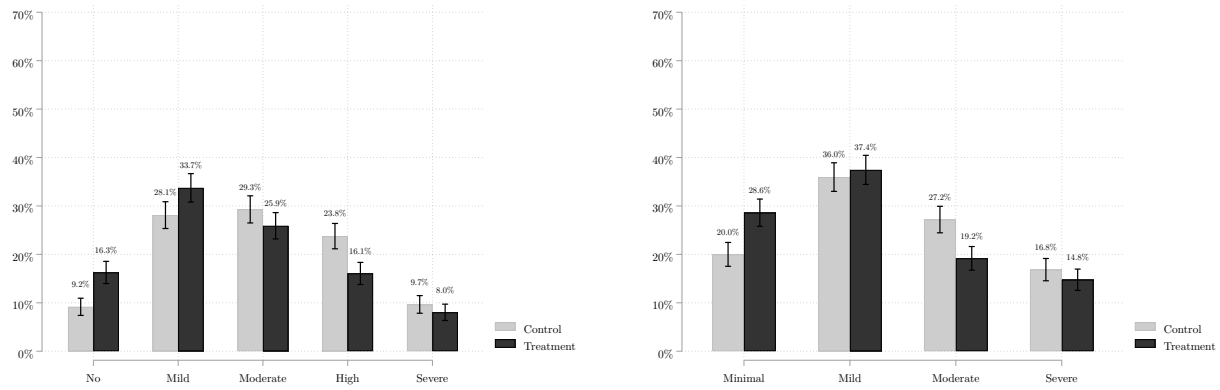
Notes: The figure shows a flowchart with the experimental design and timeline of the randomized controlled trial. The process begins with recruitment of N=6,690 participants, followed by Screening Survey 1 which yielded N=2,923 eligible participants (with 3,767 screened out). After baseline data collection (N=2,184), participants completed Screening Survey 2 and randomization (N=1964) and were randomly assigned to treatment (N=983) and control (N=981) groups. Both groups received Surveys at the 1 month, 2 months and 6 months mark. At three months half of people in the treatment group were randomly assigned to receive App access for three more months (N=491). Participants also received weekly affect polls throughout the study. The flowchart uses rectangular boxes connected by arrows to show the progression through each study phase, with sample sizes decreasing due to attrition at each stage.

Figure 2: Weekly App Engagement



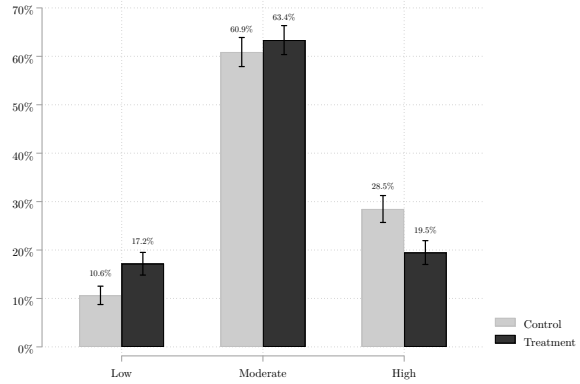
Notes: Engagement is defined as the percentage of individuals assigned to the treatment group who used the application at least once during a given week. Individuals who never downloaded the app are coded as zero usage. Vertical dashed lines indicate the timing of the follow-up surveys conducted at 1, 2, and 6 months after enrollment. Markers display the weekly engagement rate for selected weeks.

Figure 3: Mental health outcomes by treatment status at 2 months

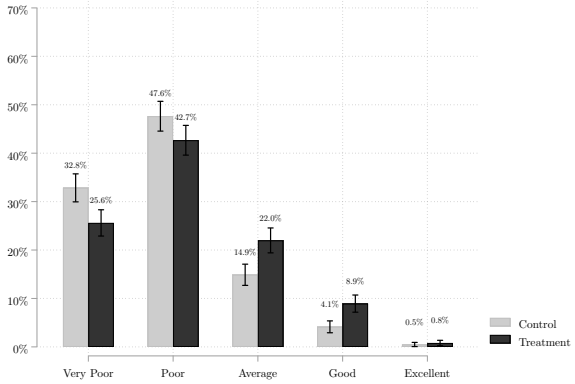


(a) Depression (PHQ-8)

(b) Anxiety (GAD-7)



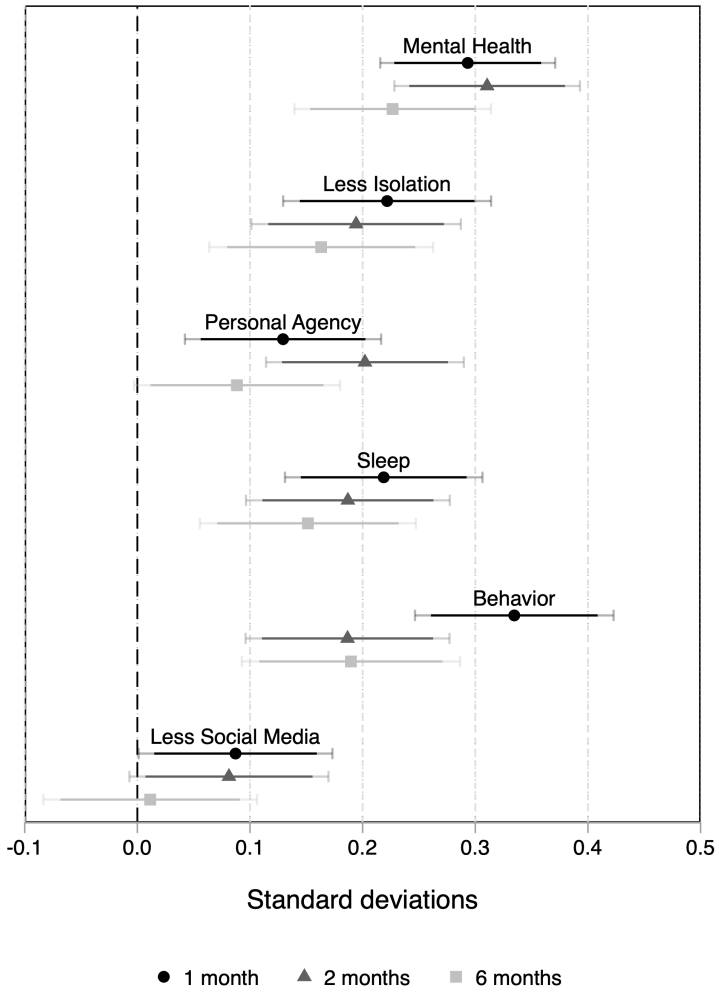
(c) Stress (PSS-4)



(d) Well-being (WHO-5)

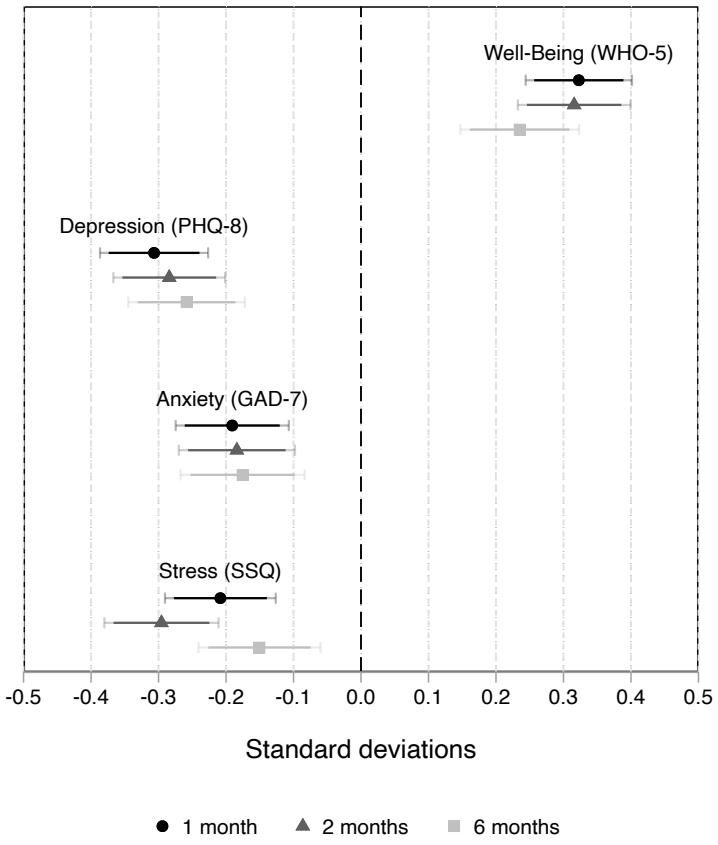
Notes: Note: Each panel shows the percent in each category within the treatment (vs control) group at 2 months, with 95% confidence intervals. The cutoffs used were: PHQ-8: 0–4 = No, 5–9 = Mild, 10–14 = Moderate, 15–19 = High, 20–24 = Severe; GAD-7: 0–4 = Minimal, 5–9 = Mild, 10–14 = Moderate, 15–21 = Severe; PSS-4: 0–5 = Low, 6–10 = Moderate, 11–16 = High; WHO-5: 0–25 = Very Poor, 26–51 = Poor, 52–71 = Average, 72–91 = Good, 92–100 = Excellent. Results are qualitatively similar at 1 and 6 months. Kolmogorov-Smirnov tests reject the hypothesis of equality of distributions by arm for all mental health outcomes at 1, 2, and 6 months ($p<0.01$).

Figure 4: Summary Impacts by Family Indices



Notes: Each point reports the estimated treatment effect on a standardized outcome index, constructed following Anderson (2008). *Higher values indicate better outcomes*, and all variables are oriented so that positive effects reflect improvements (i.e., outcomes are reverse-coded when necessary). The *Mental Health Index* combines the WHO-5 well-being score, PHQ-8, GAD-7, and PSS-4 (the latter three reversed). The *Isolation Index* is based on the six-item De Jong Gierveld Loneliness Scale (reversed), combining the social and emotional subscales. The *Efficacy and Control Index* combines scores from the General Self-Efficacy (GSE) Scale and locus of control measures. The *Sleep Index* aggregates hours of sleep, number of interruptions (reversed), and sleep difficulties (reversed). The *Behavioral Index* summarizes days in the past week participants reported engaging in positive daily activities (e.g., exercise, missed work, self-care, went out, yelled to someone in anger, helped someone with homework). The *Social Media Index* captures platform use (Facebook, X/Twitter, Instagram, TikTok) and frequent use (more than two hours per day, reversed). All regressions control for randomization strata and LASSO-selected baseline covariates.

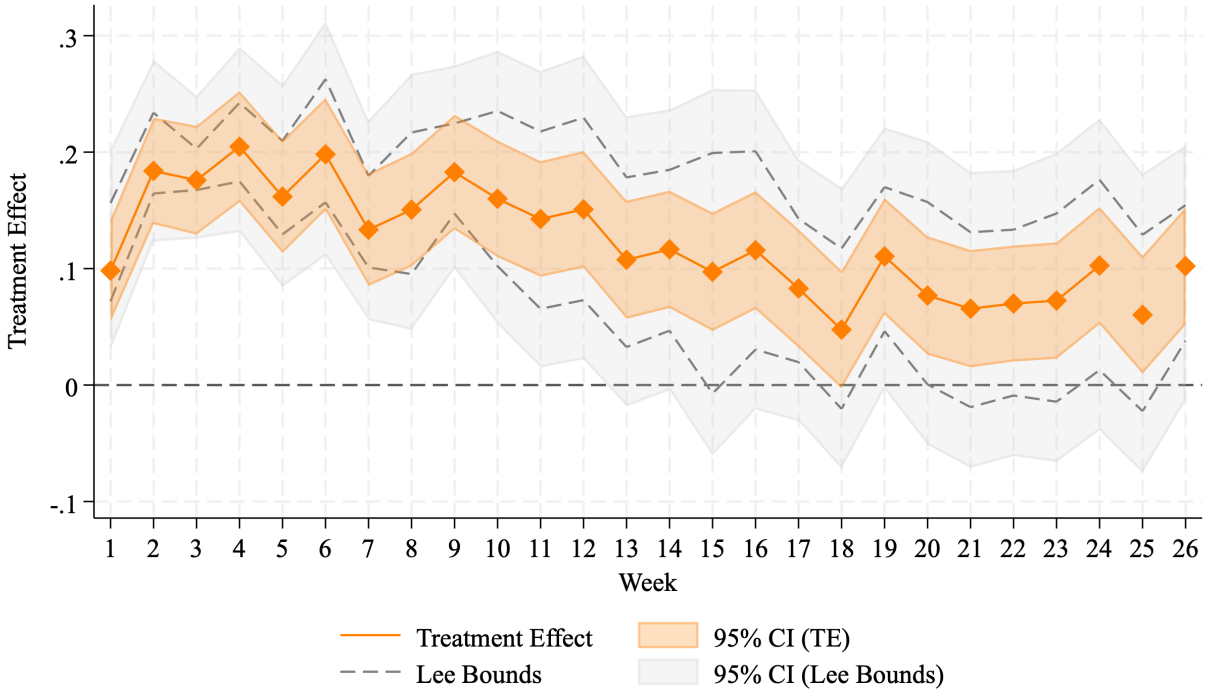
Figure 5: Summary Impacts by outcomes for mental health



Notes: Each point reports the estimated treatment effect on a standardized mental health outcome. Outcomes are expressed as z-scores, standardized using the control group mean and standard deviation within each wave. *Negative coefficients indicate improvement* for depression, anxiety, and stress, as lower scores on these scales reflect better mental health; *positive coefficients indicate improvement* for well-being. The *Well-Being (WHO-5)* score is based on the five-item World Health Organization Well-Being Index, measuring positive affect and functioning over the past two weeks. *Depression (PHQ-8)* is the eight-item Patient Health Questionnaire, assessing frequency of depressive symptoms. *Anxiety (GAD-7)* is the seven-item Generalized Anxiety Disorder scale, measuring worry, nervousness, and related symptoms. *Stress (SSQ)* is the four-item Stress Screener Questionnaire, capturing perceived stress over the past month. Circles, triangles, and squares denote 1-month, 2-month, and 6-month estimates, respectively. Thick and thin bars show 90% and 95% confidence intervals. All regressions control for randomization strata and LASSO-selected baseline covariates.

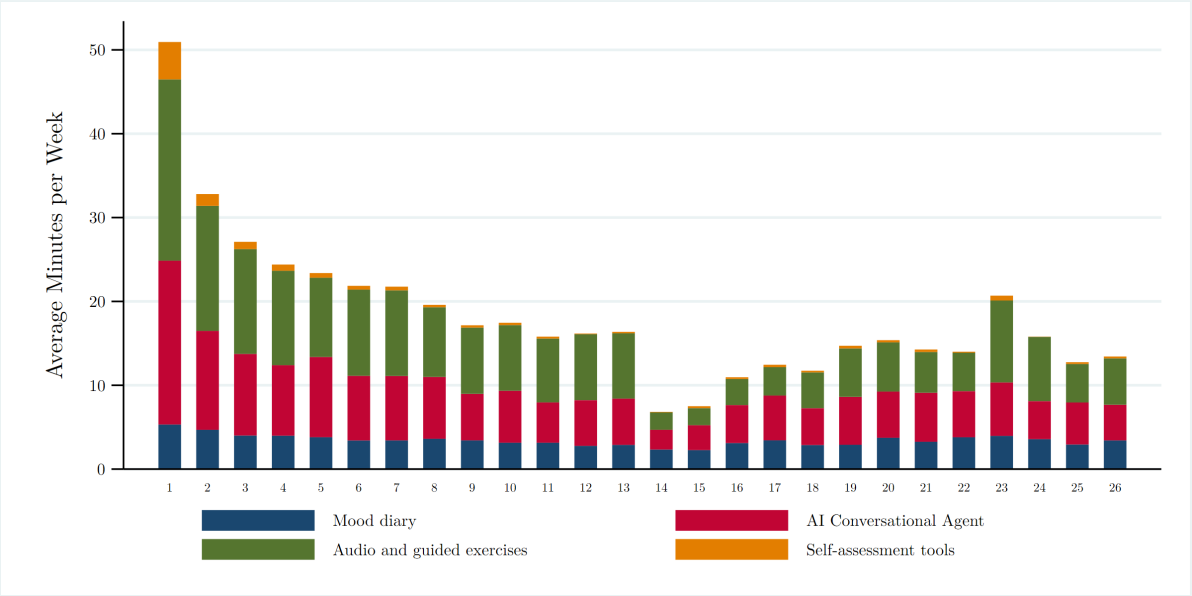
Figure 6: Impacts on Weekly Affect Data

(a) Happy



Notes: Each panel displays the estimated treatment effect on the probability of selecting a specific emoji response (happy, neutral, or sad) in weekly affect polls over 24 weeks. The orange/blue/gray line represents the point estimate with 95% confidence intervals (shaded area), calculated using regression with baseline controls and strata fixed effects, with robust standard errors. The gray dashed lines show Lee bounds accounting for differential attrition, with corresponding 95% confidence intervals (light gray shaded area). Lee bounds are calculated using age and baseline psychological distress (PHQ-4) strata as tightening covariates. The bounds indicate the range of plausible treatment effects under worst-case assumptions about selective attrition.

Figure 7: Weekly App Use by Activity Type



Notes: The figure plots average weekly minutes of app use among active users, decomposed into the four core app functions: AI conversational agent, audio/guided exercises, mood diary, and self-assessment tools. We define event duration from consecutive timestamps within sessions, accounting for background inactivity.

Table 1: Impacts on Mental Health

	Well-being (WHO-5) (1) [+]	Dep./Anx. (PHQ-8) (2) [−]	Anxiety (GAD-7) (3) [−]	Stress (PSS-4) (4) [−]	**Mental Health Index (std.)** (5) [+]
Panel A. Pooled Effects					
Treated	5.353***	-1.601***	-0.945***	-0.638***	0.285***
	(0.610)	(0.191)	(0.179)	(0.098)	(0.035)
Mean control	35.59	11.82	9.06	8.94	-0.00
Observations	5273	5272	5274	5273	5274
Panel B. Effects at 1 month					
Treated	5.592***	-1.687***	-0.971***	-0.580***	0.302***
	(0.683)	(0.219)	(0.211)	(0.109)	(0.039)
Mean control	33.79	12.30	9.49	9.16	-0.00
Observations	1850	1849	1850	1850	1850
Panel C. Effects at 2 months					
Treated	5.703***	-1.578***	-0.931***	-0.845***	0.313***
	(0.772)	(0.235)	(0.222)	(0.120)	(0.042)
Mean control	35.14	11.91	9.11	9.00	0.00
Observations	1824	1824	1825	1825	1825
Panel D. Effects at 6 months					
Treated	4.526***	-1.518***	-0.933***	-0.467***	0.235***
	(0.865)	(0.251)	(0.237)	(0.137)	(0.044)
Mean control	38.20	11.17	8.51	8.62	-0.00
Observations	1599	1599	1599	1598	1599
1 mo=2 mos (*p-val*)	0.81	0.67	0.87	0.02	0.70
2 mos=6 mos (*p-val*)	0.17	0.74	0.98	0.00	0.06
1 mo=6 mos (*p-val*)	0.29	0.48	0.89	0.42	0.14
1 mo=2 mos=6 mos (*p-val*)	0.42	0.78	0.98	0.01	0.16

Notes: Each cell reports the estimated treatment effect on the specified mental health outcome using each scale raw scores. [+] indicates that higher values represent better outcomes; [−] indicates that higher values represent worse outcomes. Regressions control for strata and lasso-selected baseline covariates. Outcomes include the WHO-5 well-being index (0–100; <50 indicates possible depression), the PHQ-8 for depression/anxiety (0–24; ≥ 10 indicates moderate to severe depression), the GAD-7 for anxiety (0–21; ≥ 10 indicates moderate to severe anxiety), and the PSS-4 for perceived stress (0–16; no formal cutoff, but scores ≥ 8 are typically considered high stress). The Mental Health Index is a standardized index of all outcomes following Anderson (2008), where higher values indicate better mental health. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 2: Impacts on Healthful Behaviors

	Daily Functioning				*Wellness Practices*				
	Missed Work (1) [−]	Helped (2) [+]	Angry (3) [−]	*Functioning Subindex (std)* (4) [+]	Exercise (5) [+]	Go Out (6) [+]	Self-care (7) [+]	*Wellness Subindex (std)* (8) [+]	**Behavior Index (std)** (9) [+]
Panel A. Pooled Effects									
Treated	-0.084**	0.187**	-0.217***	0.167***	0.239***	0.363***	0.175**	0.206***	0.246***
	(0.037)	(0.088)	(0.078)	(0.035)	(0.068)	(0.062)	(0.089)	(0.037)	(0.036)
Mean control	0.49	2.24	2.61	-0.00	1.56	1.69	3.46	0.00	-0.00
Observations	5274	5274	5274	5274	5274	5274	5274	5274	5274
Panel B. Effects at 1 month									
Treated	-0.096*	0.217**	-0.325***	0.223***	0.368***	0.528***	0.228**	0.300***	0.336***
	(0.053)	(0.103)	(0.090)	(0.045)	(0.084)	(0.082)	(0.108)	(0.046)	(0.045)
Mean control	0.48	2.26	2.68	-0.00	1.61	1.56	3.31	0.00	0.00
Observations	1850	1850	1850	1850	1850	1850	1850	1850	1850
Panel C. Effects at 2 months									
Treated	-0.068	0.162	-0.161*	0.141***	0.212**	0.243***	0.103	0.146***	0.190***
	(0.049)	(0.105)	(0.096)	(0.045)	(0.084)	(0.084)	(0.110)	(0.046)	(0.045)
Mean control	0.44	2.11	2.74	-0.00	1.43	1.70	3.58	0.00	-0.00
Observations	1825	1825	1825	1825	1825	1825	1825	1825	1825
Panel D. Effects at 6 months									
Treated	-0.087	0.138	-0.146	0.130***	0.123	0.308***	0.171	0.161***	0.199***
	(0.062)	(0.111)	(0.097)	(0.048)	(0.095)	(0.090)	(0.116)	(0.049)	(0.049)
Mean control	0.54	2.38	2.38	-0.00	1.66	1.82	3.49	0.00	-0.00
Observations	1599	1599	1599	1599	1599	1599	1599	1599	1599
1 mo=2 mos (*p-val*)	0.66	0.50	0.06	0.08	0.05	0.00	0.17	0.00	0.00
2 mos=6 mos (*p-val*)	0.81	0.87	0.97	0.94	0.46	0.53	0.50	0.79	0.99
1 mo=6 mos (*p-val*)	0.89	0.42	0.09	0.11	0.02	0.03	0.64	0.01	0.01
1 mo=2 mos=6 mos (*p-val*)	0.91	0.66	0.11	0.13	0.04	0.01	0.41	0.00	0.00

Notes: Each cell shows the estimated treatment effect on the number of days in the previous week in which the respondent had behavior-related outcomes. [+] indicates that higher values represent better outcomes; [−] indicates that higher values represent worse outcomes. (i) *Exercised*, (ii) *Go out*, whether the respondent went out for activities unrelated to school or work; (iii) *Missed work*, did not go to work (unconditional on whether they work); (iv) *Angry*, whether they raise their voice to someone else in anger, (v) *Helped*, helped their children or someone else with schoolwork, (vi) *Self Care*, spent at least 15 minutes on themselves, (vii) *Healthful Behavior Index (std)*, a standardized composite index combining measures using the inverse-covariance weighting method following Anderson (2008), oriented so that higher values indicate more healthful behaviors. Regressions control for strata and lasso-selected baseline covariates. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 3: Impacts on Sleep Outcomes

	Hours Slept (1) [+]	Interruptions (2) [−]	Sleep Issues (3) [−]	**Sleep Index (std)** (4) [+]
Panel A. Pooled Effects				
Treated	0.164***	-0.107**	-0.106***	0.188***
	(0.060)	(0.053)	(0.016)	(0.035)
Mean control	7.05	2.81	0.50	-0.00
Observations	5269	5274	5274	5274
Panel B. Effects at 1 month				
Treated	0.215***	-0.108	-0.126***	0.220***
	(0.081)	(0.071)	(0.022)	(0.045)
Mean control	6.93	2.88	0.54	-0.00
Observations	1847	1850	1850	1850
Panel C. Effects at 2 months				
Treated	0.167*	-0.119*	-0.097***	0.187***
	(0.086)	(0.068)	(0.022)	(0.046)
Mean control	7.08	2.82	0.50	-0.00
Observations	1825	1825	1825	1825
Panel D. Effects at 6 months				
Treated	0.098	-0.086	-0.094***	0.152***
	(0.091)	(0.075)	(0.023)	(0.049)
Mean control	7.17	2.71	0.45	-0.00
Observations	1597	1599	1599	1599
1 mo=2 mos (*p-val*)	0.64	0.90	0.26	0.52
2 mos=6 mos (*p-val*)	0.55	0.72	0.90	0.51
1 mo=6 mos (*p-val*)	0.28	0.83	0.24	0.21
1 mo=2 mos=6 mos (*p-val*)	0.55	0.93	0.41	0.45

Notes: Each cell shows the estimated treatment effect on one of four sleep-related outcomes. Regressions control for strata and lasso-selected baseline covariates. [+] indicates that higher values represent better outcomes; [−] indicates that higher values represent worse outcomes. Outcomes include: *Hours of sleep*, calculated as the self-reported difference between the time respondents went to bed and the time they woke up the previous night; *Interruptions*, the number of awakenings or disruptions during that sleep period; *Sleep issues*, based on the relevant item(s) from the PHQ-4 about trouble sleeping (e.g., "Not being able to sleep or sleeping too much"); and (iv) the *Sleep Index (std)*, a standardized composite index combining these three measures using the inverse-covariance weighting method following Anderson (2008), oriented so that higher values indicate better sleep. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 4: Impacts on Social Media Use

	Facebook (1) [-]	X (2) [-]	Instagram (3) [-]	TikTok (4) [-]	Use Freq. (5) [-]	**Less Social Media Index (std)** (6) [+]
Panel A. Pooled Effects						
Treated	-0.094	-0.121*	-0.163	0.132	-0.018	0.064
	(0.092)	(0.070)	(0.107)	(0.116)	(0.017)	(0.041)
Mean control	5.29	0.75	2.99	2.70	0.30	0.00
Observations	5274	5274	5274	5274	5273	5274
Panel B. Effects at 1 month						
Treated	-0.119	-0.167**	-0.293**	0.088	-0.010	0.088**
	(0.102)	(0.077)	(0.117)	(0.125)	(0.021)	(0.044)
Mean control	5.41	0.77	2.99	2.68	0.29	0.00
Observations	1850	1850	1850	1850	1850	1850
Panel C. Effects at 2 months						
Treated	-0.128	-0.084	-0.172	0.038	-0.034	0.085*
	(0.106)	(0.078)	(0.119)	(0.128)	(0.021)	(0.045)
Mean control	5.25	0.73	2.93	2.69	0.30	0.00
Observations	1825	1825	1825	1825	1825	1825
Panel D. Effects at 6 months						
Treated	-0.032	-0.119	-0.016	0.279**	-0.012	0.015
	(0.115)	(0.081)	(0.126)	(0.135)	(0.023)	(0.049)
Mean control	5.19	0.75	3.04	2.73	0.30	0.00
Observations	1599	1599	1599	1599	1598	1599
1 mo=2 mos (*p-val*)	0.91	0.14	0.15	0.62	0.23	0.93
2 mos=6 mos (*p-val*)	0.34	0.71	0.16	0.02	0.35	0.07
1 mo=6 mos (*p-val*)	0.41	0.39	0.01	0.06	0.89	0.07
1 mo=2 mos=6 mos (*p-val*)	0.62	0.34	0.04	0.05	0.42	0.14

Notes: Each cell shows the estimated treatment effect on one of seven social media–related outcomes. [+] indicates that higher values represent better outcomes (less use); [−] indicates that higher values represent worse outcomes. The first four outcomes capture the number of days in the past seven days that respondents used: (i) *Facebook*, (ii) *X* (formerly Twitter), (iii) *Instagram*, and (iv) *TikTok*. The fifth outcome, *Use frequency*, is an indicator for using social media for two hours or more per day. Column (6) reports *Social Media Index (std)*, is a standardized composite index combining these six measures using the inverse-covariance weighting method following Anderson (2008), oriented so that higher values indicate less time spent on social media (i.e., healthier digital habits). Regressions control for strata and lasso-selected baseline covariates. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 5: Impacts on Feelings of Isolation

	Emotional Score (1) [−]	Social Score (2) [−]	**Isolation Index (std.)** (3) [+]
Panel A. Pooled Effects			
Treated	-0.165***	-0.121***	0.200***
	(0.035)	(0.036)	(0.040)
Mean control	2.21	2.48	0.00
Observations	5274	5274	5274
Panel B. Effects at 1 month			
Treated	-0.173***	-0.144***	0.228***
	(0.040)	(0.042)	(0.047)
Mean control	2.30	2.51	-0.00
Observations	1850	1850	1850
Panel C. Effects at 2 months			
Treated	-0.189***	-0.094**	0.195***
	(0.043)	(0.042)	(0.047)
Mean control	2.23	2.48	0.00
Observations	1825	1825	1825
Panel D. Effects at 6 months			
Treated	-0.122**	-0.121**	0.167***
	(0.048)	(0.048)	(0.051)
Mean control	2.09	2.45	0.00
Observations	1599	1599	1599
1 mo=2 mos (*p-val*)	0.69	0.17	0.41
2 mos=6 mos (*p-val*)	0.19	0.53	0.51
1 mo=6 mos (*p-val*)	0.35	0.59	0.22
1 mo=2 mos=6 mos (*p-val*)	0.42	0.39	0.44

Notes: Each cell reports the estimated treatment effect on the specified social isolation outcome using raw scores from the six-item De Jong Gierveld Loneliness Scale. Regressions control for strata and lasso-selected baseline covariates. The scale includes three items capturing emotional loneliness and three items capturing social loneliness. Responses are scored following the standard protocol, with higher values indicating greater loneliness. We report effects separately for emotional and social loneliness subscales, as well as for an overall standardized isolation index constructed following Anderson (2008). * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 6: Impacts on Perceptions of Personal Agency

	Self-Efficacy (GSE) (1) [+]	Locus of control (2) [+]	**Personal Agency Index (std.)** (3) [+]
Panel A. Pooled Effects			
Treated	0.313***	0.502***	0.161***
	(0.121)	(0.108)	(0.037)
Mean control	15.86	12.10	-0.00
Observations	5273	5272	5274
Panel B. Effects at 1 month			
Treated	0.308**	0.420***	0.149***
	(0.137)	(0.124)	(0.043)
Mean control	15.66	12.15	-0.00
Observations	1850	1850	1850
Panel C. Effects at 2 months			
Treated	0.494***	0.640***	0.222***
	(0.143)	(0.128)	(0.043)
Mean control	15.81	12.02	-0.00
Observations	1824	1824	1825
Panel D. Effects at 6 months			
Treated	0.116	0.423***	0.105**
	(0.157)	(0.141)	(0.045)
Mean control	16.15	12.13	-0.00
Observations	1599	1598	1599
1 mo=2 mos (*p-val*)	0.17	0.06	0.07
2 mos=6 mos (*p-val*)	0.01	0.14	0.01
1 mo=6 mos (*p-val*)	0.18	0.85	0.36
1 mo=2 mos=6 mos (*p-val*)	0.03	0.13	0.02

Notes: Each cell reports the estimated treatment effect on locus of control or self-efficacy outcomes measured using the General Self-Efficacy (GSE) Scale. In those scales positive scores denote more self-efficacy and higher locus of control. [+] indicates that higher values represent better outcomes; [−] indicates that higher values represent worse outcome. We also report an overall standardized personal agency index constructed following Anderson (2008) using both scores, where higher scores correspond to more positive outcomes. Regressions control for strata and lasso-selected baseline covariates. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 7: Impacts on the Demand for Psychotherapy

	Visit Psychologist (1)	Num. Visits (2)
Panel A. Pooled Effects		
Treated	0.045***	
	(0.012)	
Mean control	0.15	
Observations	3579	
4 wks=8 wks (*p-val*)	0.26	
Panel B. Effects at 1 month		
Treated	0.034**	
	(0.017)	
Mean control	0.15	
Observations	1855	
Panel C. Effects at 2 months		
Treated	0.054***	0.152***
	(0.018)	(0.054)
Mean control	0.16	0.37
Observations	1829	1829
Panel D. Effects at 6 months		
Treated	0.077***	0.298***
	(0.019)	(0.075)
Mean control	0.15	0.51
Observations	1603	1603
1 mo=2 mo (*p-val*)	0.23	
2 mos=6 mos (*p-val*)	0.25	0.05
1 mo=6 mos (*p-val*)	0.04	
1 mo=2 mos=6 mos (*p-val*)	0.11	

Notes: psychotherapy utilization outcomes measured in the month prior to the survey. The first outcome, *Any psychotherapy use*, is an indicator equal to one if the respondent reports at least one visit with a psychologist or psychotherapist in the previous month. The second outcome, *Number of visits*, reports the unconditional number of psychotherapy visits in the previous month. Regressions control for strata and lasso-selected baseline covariates. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 8: Impacts on Psychotherapy, Depression, Anxiety, Beliefs, and Psychotherapy Awareness/Knowledge by Baseline PHQ-4

	(1) Psychotherapy use	(2) Mental Health Index	(3) Depression (PHQ-8)	(4) Anxiety (GAD-7)	(5) 1st-order beliefs (STD)	(6) 2nd-order beliefs (STD)	(7) Knowledge about therapy (STD)
Panel A. Mild baseline distress (PHQ-4 = 3–5)							
Treatment	0.019	0.224***	-1.174***	-0.644**	0.015	0.119*	-0.039
	(0.023)	(0.063)	(0.321)	(0.280)	(0.069)	(0.066)	(0.075)
Panel B. Moderate baseline distress (PHQ-4 = 6–8)							
Treatment	0.074***	0.352***	-2.401***	-1.585***	0.058	0.172**	0.056
	(0.028)	(0.069)	(0.372)	(0.338)	(0.081)	(0.079)	(0.090)
Panel C. Severe baseline distress (PHQ-4 = 9–12)							
Treatment	0.117***	0.255***	-1.365***	-0.565	-0.099	0.029	0.112
	(0.036)	(0.091)	(0.507)	(0.478)	(0.107)	(0.105)	(0.119)
Panel D. Overall ITT							
Treatment	0.056***	0.248***	-1.473***	-0.821***	0.002	0.120***	0.045
	(0.016)	(0.046)	(0.243)	(0.217)	(0.046)	(0.045)	(0.051)
Panel E. Tests of equality of treatment effects (p-values)							
Mild = Moderate	0.131	0.172	0.013	0.032	0.686	0.609	0.413
Mild = Severe	0.022	0.779	0.751	0.886	0.374	0.466	0.281
Moderate = Severe	0.345	0.397	0.099	0.081	0.243	0.278	0.707
Observations	4686	4686	4686	4686	2838	2837	1562
Clusters Individual Level	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Waves pooled	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Measured at 1 mo	Yes	Yes	Yes	Yes	No	No	No
Measured at 2 mos	Yes	Yes	Yes	Yes	No	Yes	No
Measured at 6 mos	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Panel E reports p-values from tests of equality of subgroup-specific ITT effects. "Mild = Moderate" and "Mild = Severe" correspond to tests that the relevant treatment interaction equals zero. "Moderate = Severe" corresponds to a test that the difference between the two interaction terms equals zero. Robust standard errors clustered at the individual level. STD = standardized. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

References

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, 101(2):470–492.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, 110(3):629–676.

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American statistical Association*, 103(484):1481–1495.

Angelucci, M. and Bennett, D. (2024). The economic impact of depression treatment in india: Evidence from community-based provision of pharmacotherapy. *American Economic Review*, 114(1):169–98.

Angelucci, M. and Bennett, D. (2026). Mental health and the willingness to invest: Evidence from group psychotherapy in india. Unpublished Manuscript.

Angelucci, M., Bennett, D., Fabregas, R., and Vazquez, A. (2026). Parental mental health and adolescent outcomes: Evidence from a low-cost scalable program in mexico. Working paper.

Athey, S. and Wager, S. (2019). Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51.

Baglioni, C., Battagliese, G., Feige, B., Spiegelhalder, K., Nissen, C., Voderholzer, U., Lombardo, C., and Riemann, D. (2016). Insomnia as a predictor of depression: A meta-analytic evaluation of longitudinal epidemiological studies. *Journal of Affective Disorders*, 186:10–19.

Baranov, V., Bhalotra, S., Biroli, P., and Maselko, J. (2020). Maternal depression, women's empowerment, and parental investment: Evidence from a randomized controlled trial. *American Economic Review*, 110(3):824–59.

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

Bhat, B., de Quidt, J., Haushofer, J., Patel, V. H., Rao, G., Schilbach, F., and Vautrey, P.-L. P. (2022). The long-run effects of psychotherapy on depression, beliefs, and economic outcomes. NBER Working Paper 30011.

Botha, F. and Dahmann, S. C. (2024). Locus of control, self-control, and health outcomes. *SSM-Population Health*, 25:101566.

Bower, P., Kontopantelis, E., Sutton, A., Kendrick, T., Richards, D. A., Gilbody, S., Knowles, S., Cuijpers, P., Andersson, G., Christensen, H., et al. (2013). Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. *Bmj*, 346.

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernández-Val, I. (2025). Fisher–schultz lecture: Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. *Econometrica*, 93(4):1121–1164.

Cilliers, J., Elashmawy, N., and McKenzie, D. (2024). *Using post-double selection Lasso in field experiments*. World Bank.

Collaborators, G. . M. D. et al. (2022). Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet Psychiatry*, 9(2):137–150.

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., and Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *The Canadian Journal of Psychiatry*, 58(7):376–385.

Cuijpers, P., Karyotaki, E., Reijnders, M., and Ebert, D. D. (2019). How effective are cognitive behavior therapies for major depression and anxiety disorders? a meta-analytic update of the evidence. *World Psychiatry*, 18(2):245–258.

Cuijpers, P., Miguel, C., Harrer, M., Plessen, C. Y., Ciharova, M., Ebert, D., and Karyotaki, E. (2023). Cognitive behavior therapy vs. control conditions, other psychotherapies, pharmacotherapies and combined treatment for depression: A comprehensive meta-analysis including 409 trials with 52,702 patients. *World Psychiatry*, 22(1):105–115.

De Jong-Gierveld, J. and van Tilburg, T. G. (2006). A 6-item scale for overall, emotional, and social loneliness: Confirmatory tests on survey data. *Research on aging*, 28(5):582–598.

De Quidt, J., Haushofer, J., and Roth, C. (2018). Measuring and bounding experimenter demand. *American Economic Review*, 108(11):3266–3302.

DellaVigna, S., Pope, D., and Vivalt, E. (2019). Predict science to improve science. *Science*, 366(6464):428–429.

Dhar, D., Jain, T., and Jayachandran, S. (2022). Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in india. *American Economic Review*, 112(3):899–927.

DHHS (2023). Our epidemic of loneliness and isolation: The u.s. surgeon general's advisory on the healing effects of social connection and community. Online report.

Ebner-Priemer, U. W. and Trull, T. J. (2009). Ecological momentary assessment of mood disorders and mood dysregulation. *Psychological assessment*, 21(4):463.

Fancourt, D., Aughterson, H., Finn, S., Walker, E., and Steptoe, A. (2021). How leisure activities affect health: a narrative review and multi-level theoretical framework of mechanisms of action. *The Lancet Psychiatry*, 8(4):329–339.

Firth, J., Torous, J., Sarris, J., Lichstein, K. L., Green, J., Roiser, J. P., Carney, R., Koyanagi, A., and Cosco, T. D. (2022). The "digital placebo effect" in mental health apps: a scoping review of the literature. *World Psychiatry*, 21(3):314–332.

García-Campayo, J., Zamorano, E., Ruiz, M. A., Pardo, A., Pibas, G., and Rejas, J. (2010). The assessment of generalized anxiety disorder: psychometric validation of the spanish version of the self-administered gad-7 scale in primary care. *Health and Quality of Life Outcomes*, 8(1):8.

Gartlehner, G., Wagner, G., Matyas, N., Titscher, V., Greimel, J., Lux, L., Gaynes, B. N., Viswanathan, M., Patel, S., and Lohr, K. N. (2017). Pharmacological and non-pharmacological treatments for major depressive disorder: review of systematic reviews. *BMJ Open*, 7(6):e014912.

Harvey, A. G. (2011). Sleep and circadian functioning: critical mechanisms in the mood disorders? *Annual Review of Clinical Psychology*, 7:297–319.

Haushofer, J. and Fehr, E. (2014). On the psychology of poverty. *science*, 344(6186):862–867.

Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., and Jacobson, N. C. (2024). Evaluating therabot: A randomized control trial investigating the feasibility and effectiveness of a generative ai therapy chatbot for depression, anxiety, and eating disorder symptom treatment. *NEJM AI*. Trial Registration Number: NCT06013137.

Holt-Lunstad, J. (2022). Social connection as a public health issue: The evidence and a systemic framework for prioritizing the "social" in social determinants of health. *Annual Review of Public Health*, 43(1):193–213.

Imai, K., Keele, L., and Tingley, D. (2010a). A general approach to causal mediation analysis. *Psychological Methods*, 15(4):309–334.

Imai, K., Keele, L., and Yamamoto, T. (2010b). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71.

Jaeger, S. R., Roigard, C. M., Jin, D., Vidal, L., and Ares, G. (2019). Valence, arousal and sentiment meanings of 33 facial emoji: Insights for the use of emoji in consumer research. *Food research international*, 119:895–907.

Kazantzis, N., Whittington, C., and Dattilio, F. (2010). Meta-analysis of homework effects in cognitive and behavioral therapy: A replication and extension. *Clinical Psychology: Science and Practice*, 17(2):144.

Kazantzis, N., Whittington, C., and Dattilio, F. (2018). Homework in cognitive behavioral therapy: A systematic review of adherence assessment in anxiety and depression. *Journal of Consulting and Clinical Psychology*, 86(9):703–716.

Kleiman, E. M., Turner, B. J., Fedor, S., Beale, E. E., Huffman, J. C., and Nock, M. K. (2017). Examination of real-time fluctuations in suicidal ideation and its risk factors: Results from two ecological momentary assessment studies. *Journal of abnormal psychology*, 126(6):726.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., and Mokdad, A. H. (2009). The phq-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1-3):163–173.

Lattal, K. A. (2010). Delayed reinforcement of operant behavior. *Journal of the Experimental Analysis of Behavior*, 93(1):129–139.

Li, H., Zhang, R., Lee, Y.-C., Kraut, R. E., and Mohr, D. C. (2023). Systematic review and meta-analysis of ai-based conversational agents for promoting mental health and well-being. *npj Digital Medicine*, 6(236).

Lund, C., Orkin, K., Witte, M., Walker, J. H., Davies, T., Haushofer, J., Murray, S., Bass, J., Murray, L., Tol, W., et al. (2024). The effects of mental health interventions on labor market outcomes in low-and middle-income countries. Technical report, National Bureau of Economic Research.

Mahindru, A., Patil, P., and Agrawal, V. (2023). Role of physical activity on mental health and well-being: a review. *Cureus*, 15(1):e33475.

McGovern, C., Athey, A., Beale, E. E., Overholser, J. C., Gomez, S. H., and Silva, C. (2024). Who will stay and who will go? identifying risk factors for psychotherapy dropout. *Counselling and Psychotherapy Research*, 24(4):1432–1441.

Mohr, D. C., Schueller, S. M., Montague, E., Burns, M. N., and Rashidi, P. (2014). The behavioral intervention technology model: an integrated conceptual and technological framework for ehealth and mhealth interventions. *Journal of medical Internet research*, 16(6):e146.

Patel, V., Weobong, B., Weiss, H. A., Anand, A., Bhat, B., Katti, B., Dimidjian, S., Araya, R., Hollon, S. D., King, M., et al. (2017). The healthy activity program (hap), a lay counsellor-delivered brief psychological treatment for severe depression, in primary care in india: a randomised controlled trial. *The Lancet*, 389(10065):176–185.

Ridley, M., Rao, G., Schilbach, F., and Patel, V. (2020). Poverty, depression, and anxiety: Causal evidence and mechanisms. *Science*, 370(6522):eaay0214.

Smith, K. A., Ward, T., Lambe, S., Ostinelli, E. G., Blease, C., Gant, T., Gold, S. M., Holmes, E. A., Paccoud, I., Vinnikova, A., et al. (2025). Engagement and attrition in digital mental health: current challenges and potential solutions. *npj Digital Medicine*, 8(1):398.

Steel, Z., Marnane, C., Iranpour, C., Chey, T., Jackson, J. W., Patel, V., and Silove, D. (2014). The global prevalence of common mental disorders: A systematic review and meta-analysis 1980–2013. *International Journal of Epidemiology*, 43(2):476–493.

Thompson, C. A., Novotny, P. J., Yost, K., Bartz, A. C., Rogak, L., and Dueck, A. C. (2025). Development and validation of emoji response scales for assessing patient-reported outcomes. *JCO Clinical Cancer Informatics*, 9:e2400148.

Topp, C. W., Østergaard, S. D., Søndergaard, S., and Bech, P. (2015). The who-5 well-being index: A systematic review of the literature. *Psychotherapy and Psychosomatics*, 84(3):167–176.

Torous, J., Jän Myrick, K., Rauseo-Ricupero, N., and Firth, J. (2020a). Digital mental health and covid-19: Using technology today to accelerate the curve on access and quality tomorrow. *JMIR Mental Health*, 7(3):e18848.

Torous, J., Lipschitz, J., Ng, M., and Firth, J. (2020b). Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. *Journal of affective disorders*, 263:413–419.

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., and Torous, J. B. (2019). Chatbots and conversational agents in mental health: A review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

Van Dis, E. A., Van Veen, S. C., Hagenaars, M. A., Batelaan, N. M., Bockting, C. L., Van Den Heuvel, R. M., Cuijpers, P., and Engelhard, I. M. (2020). Long-term outcomes of cognitive behavioral therapy for anxiety-related disorders: a systematic review and meta-analysis. *JAMA psychiatry*, 77(3):265–273.

Westra, H. A., Arkowitz, H., and Dozois, D. J. (2009). Adding a motivational interviewing pretreatment to cognitive behavioral therapy for generalized anxiety disorder: A preliminary randomized controlled trial. *Journal of anxiety disorders*, 23(8):1106–1117.

Williams, J. M. G., Mathews, A., and MacLeod, C. (1996). The emotional stroop task and psychopathology. *Psychological Bulletin*, 120(1):3–24.

Woolley, K. and Fishbach, A. (2017). Immediate rewards predict adherence to long-term goals. *Personality and Social Psychology Bulletin*, 43(2):151–162.

Woolley, K. and Fishbach, A. (2018). It's about time: Earlier rewards increase intrinsic motivation. *Journal of personality and social psychology*, 114(6):877.

World Health Organization (2023). World mental health report: Transforming mental health for all. https://www.who.int/publications/i/item/9789240049338. Accessed July 2025.

Zadey, S. (2023). Scale-up costs and societal benefits of psychological interventions for alcohol use and depressive disorders in india. *PLOS Global Public Health*, 3(9):e0002017.

Online Appendix

A The Mindsurf App **49**
A.1 The four main app features . . . 49
A.2 Training the AI Assistant . . . 52
A.3 Safety Protocols . . . 53

B Recruitment Ads **56**

C Effort and Emotional Stroop Games **57**
C.1 Effort Task (Matrix Counting) . . . 57
C.2 Emotional Stroop Task . . . 57

D Weekly affect data **59**

E Sample Characteristics **60**

F Balance and Attrition **62**
F.1 Attrition in Survey Data . . . 62

G Average Effects – Additional Results **64**
G.1 Impacts on Effort Task and Emotional Stroop Task . . . 64
G.2 Fixed Sample Across Survey Rounds . . . 65

H Patterns of App Use **71**

I Additional Specifications **73**
I.1 Heterogeneous effects by Baseline Psychotherapy Use . . . 73

J Experimenter Demand Effects **74**

K Use-Effect Dynamics **78**

L App Use and Mental Health Impacts **81**

M Psychotherapy Use **85**

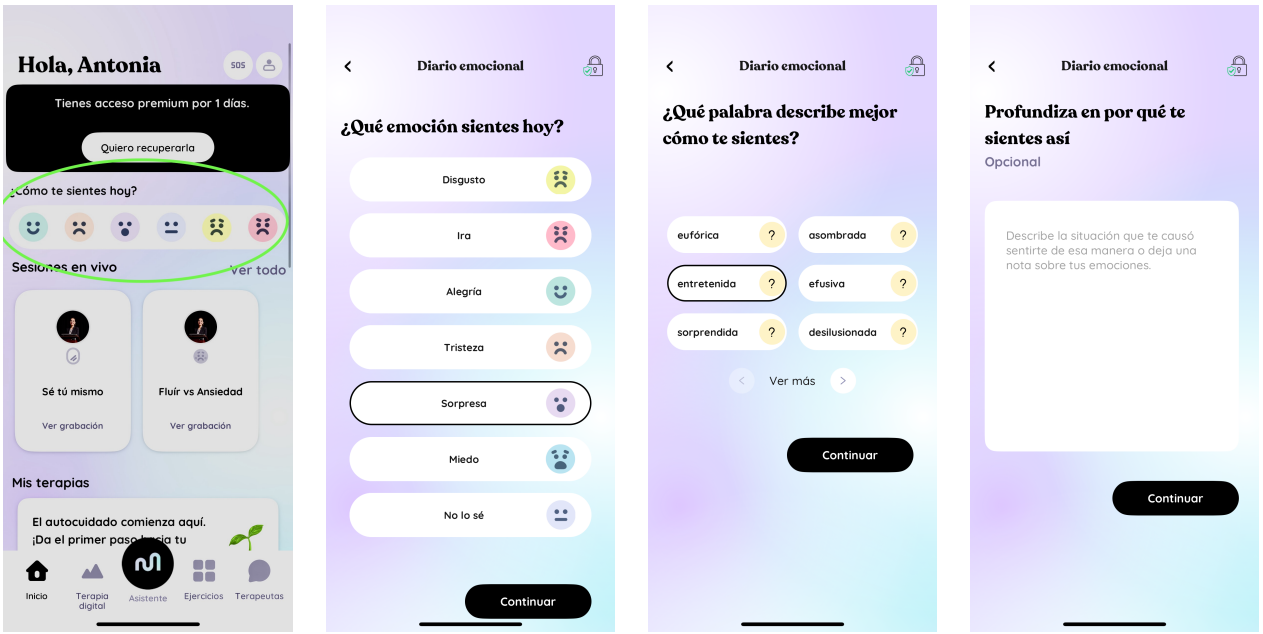
N Experimental Results from Extending App Access **91**

O Expert Predictions **93**

A The Mindsurf App

A.1 The four main app features

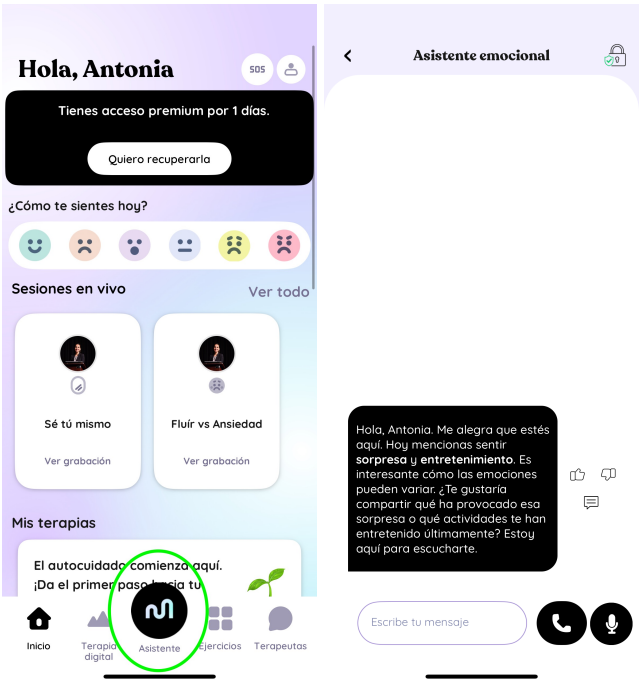
Figure A.1: App Interface: the Mood Diary



(a) Main dashboard (b) Selected emotion (c) Reflection prompt (d) Descriptive words

Notes: Screenshots showing the mood diary workflow in Spanish. Subfigure (a) displays the main dashboard with mood check-in question "¿Cómo te sientes hoy?" and six emoji options. Subfigure (b) shows the emotion selection interface where users choose from seven emotions with corresponding emoji faces. Subfigure (c) shows a reflection prompt that asks users to choose a word that describes how they are feeling. Subfigure (d) shows a second reflection prompt that invite user to describe in more detail why they are feeling that way.

Figure A.2: App Interface: the AI Conversational Agent



(a) Agent interface (b) Conversation flow

Notes: Screenshots showing the AI conversational agent interface and therapeutic interaction capabilities. Subfigure (a) displays the main screen where users can access the conversational agent interface Subfigure (b) shows the chat environment where users can engage in real-time dialogue with the AI assistant. In this example the AI agent starts a conversation saying *"Hello, Antonia. I'm glad you're here. Today you mentioned feeling surprised and entertained. It's interesting how emotions can vary. Would you like to share what has caused that surprise or what activities have entertained you recently? I'm here to listen."*

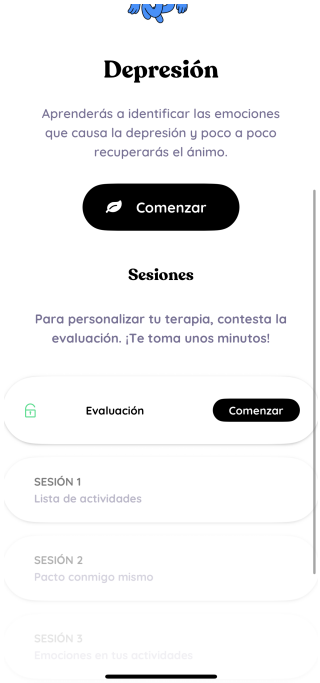
Figure A.3: App Interface: the Exercises



(a) Exercise menu



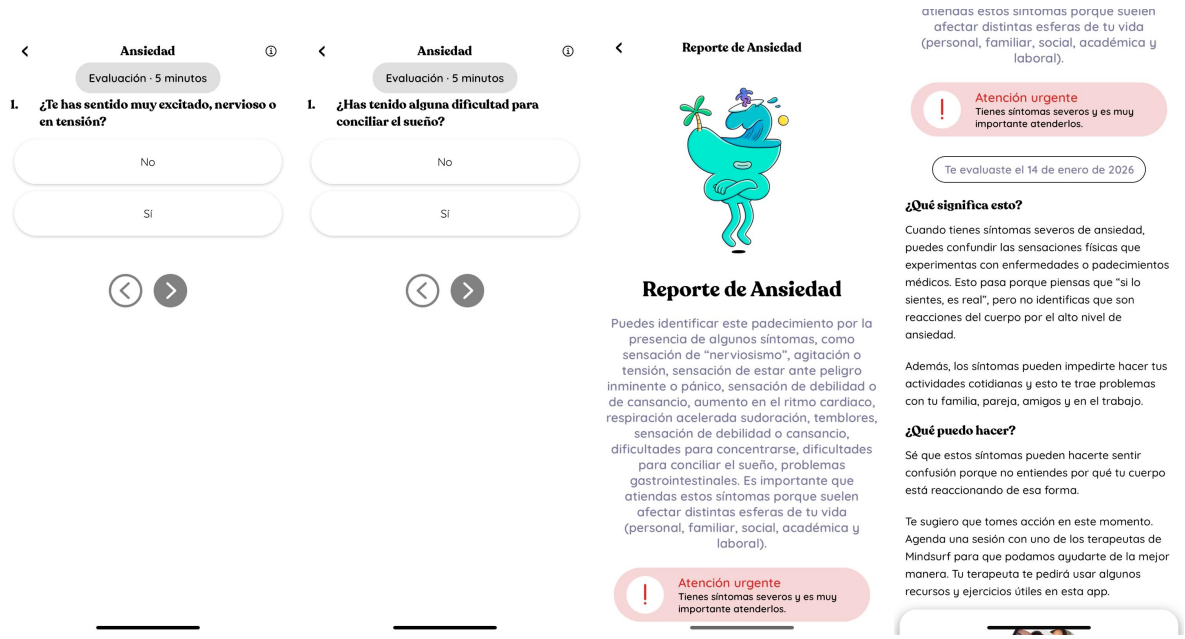
(b) Exercise interface



(c) Depression exercise: details

Notes: Subfigure (a) displays the exercise menu or library of psycho-educational modules covering key mental health areas such as anxiety, depression, stress, self-esteem, and emotional intelligence. Subfigure (b) and (c) show an example of the exercise interface.

Figure A.4: App Interface: the Self-Assessment Tools



(a) Self-Assessment Question (b) Self-Assessment Question (c) Results (d) Results

Notes: Screenshots showing examples of self-assessment questions for anxiety. Subfigures (a) and (b) display an examples of a question to assess anxiety levels ("*Have you felt very excited, nervous or tense?*" and "*Have you had trouble falling asleep?*") Subfigures (c) and (d) show results from the assessment test, telling the user that they have high levels of anxiety that require urgent attention and how could they get support (though the app or a psychotherapist).

A.2 Training the AI Assistant

The AI assistant is designed to provide dialogic support for psychosocial issues and is trained exclusively on curated mental health–related content. Training materials consist of a corpus of verified academic and clinical resources, including foundational texts, intervention manuals, diagnostic references, and structured educational materials in psychology, with a strong emphasis on evidence-based CBT approaches and meditation techniques.

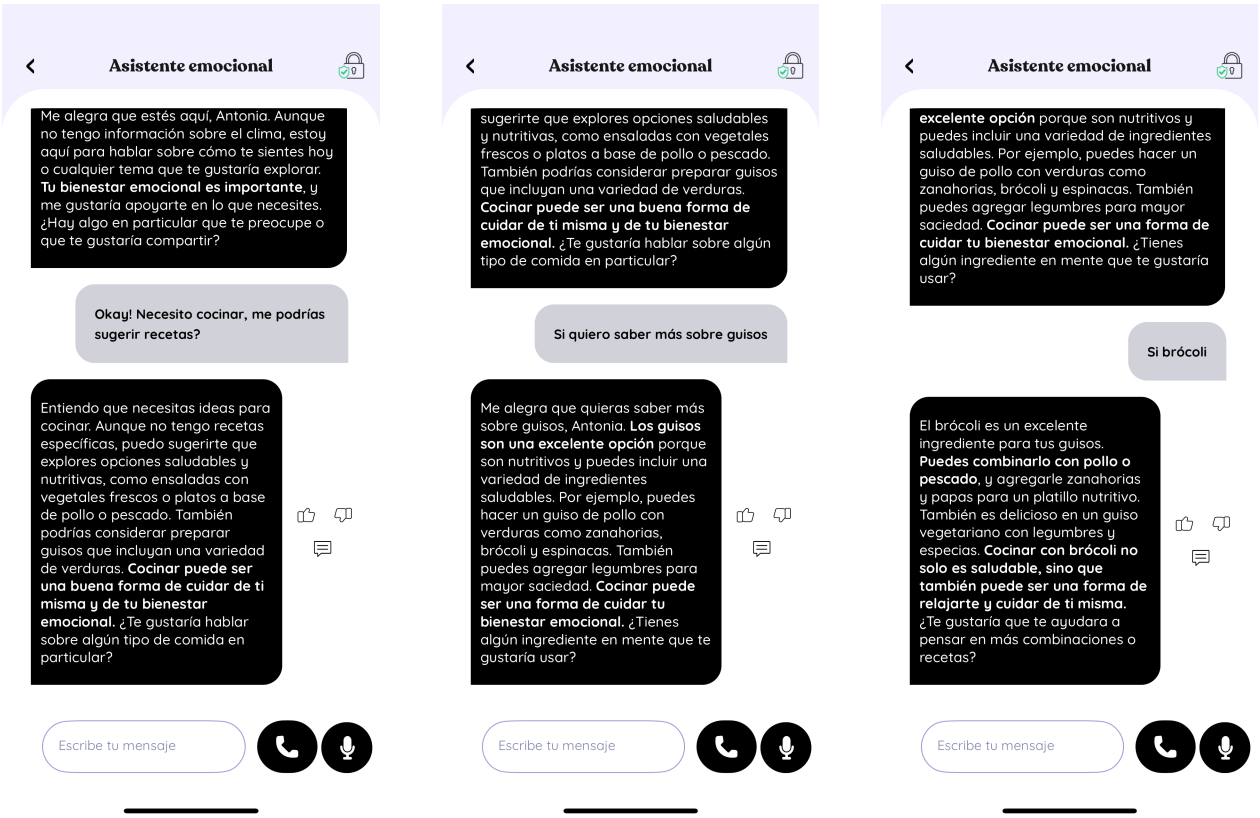
The training corpus covers a range of topics relevant to mental health support, including core CBT principles and applications; rational emotive behavior therapy (REBT); diagnostic classifications (DSM-5, ICD-10); CBT-based assessment tools; common cognitive distortions; and applied domains such as depression, anxiety and stress management, sleep and well-being, self-esteem, emotional regulation, interpersonal difficulties, and habit formation. The corpus draws on peer-reviewed academic and clinical sources, including foundational texts and established treatment manuals in cognitive and behavioral therapies.

The responses draw exclusively on verified source material. Retrieved content is then used to

condition the assistant's outputs, allowing it to provide contextually relevant guidance while remaining within its intended mental health support scope. This design limits responses to validated psychological content and supports consistency with established clinical knowledge.

A.3 Safety Protocols

Figure A.5: Example of Topical Boundaries: Cooking Recipe Request



(a) Recipe request (b) Redirection response 1 (c) Redirection response 2

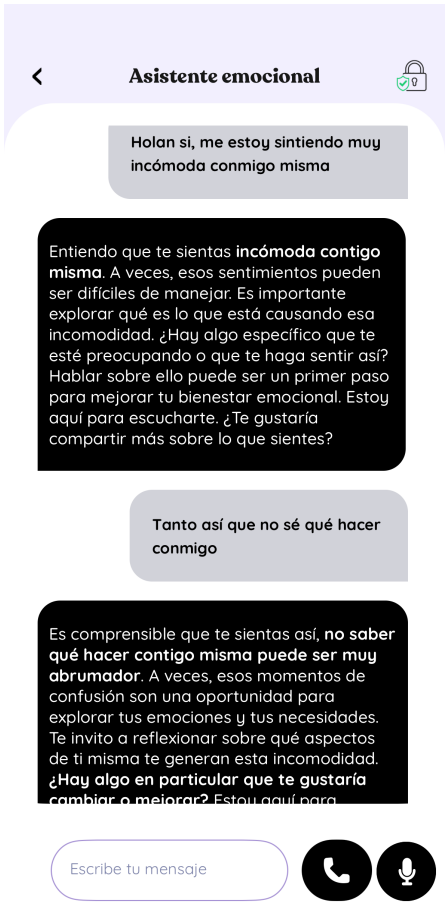
Notes: Screenshots demonstrating how the AI assistant handles off-topic requests. Subfigure (a) shows a conversation about cooking where the user says *"Okay! I need to cook, could you suggest recipes?"* and the assistant responds *"I understand that you need ideas for cooking. Although I don't have specific recipes, I can suggest that you explore healthy and nutritious options, such as salads with fresh vegetables or dishes based on chicken or fish. You could also consider preparing stews that include a variety of vegetables. Cooking can be a good way to take care of yourself and your emotional wellbeing. Would you like to talk about any particular type of food?"* Subfigure (b) shows that the user persists saying *"Yes, I want to know more about stews,"* and the Assistant answers *"I'm glad you want to know more about stews, Antonia. Stews are an excellent option because they're nutritious and you can include a variety of healthy ingredients. For example, you can make a chicken stew with vegetables like carrots, broccoli, and spinach. You can also add legumes for more satiety. Cooking can be a way to take care of your emotional wellbeing. Do you have any particular ingredient in mind that you'd like to use?"* Subfigure (c) shows the user answers *"Yes, broccoli"* about ingredient preferences. The assistant responds: *"Broccoli is an excellent ingredient for your dishes. You can combine it with chicken or fish, and add carrots and potatoes for a nutritious dish. It's also delicious in a vegetarian stew with legumes and spices. Cooking with broccoli is not only healthy, but can also be a way to relax and take care of yourself. Would you like me to help you think of more combinations or recipes?"*

The app has multiple ongoing protocols to ensure the safety of its users. First, the AI assistant is intentionally trained to discuss a limited range of topics. Off-topic queries are either directly acknowledged (e.g., by explaining that it is not able to provide the specific information), or re-directed towards related topics (e.g., when asked to provide a specific recipe, the assistant explained that it was not able to provide any specific recipe, but briefly described principles of healthful nutrition). This limited and specific training reduces the risk of hallucinations or off-topic conversations. See Figure A.5 for a specific example.

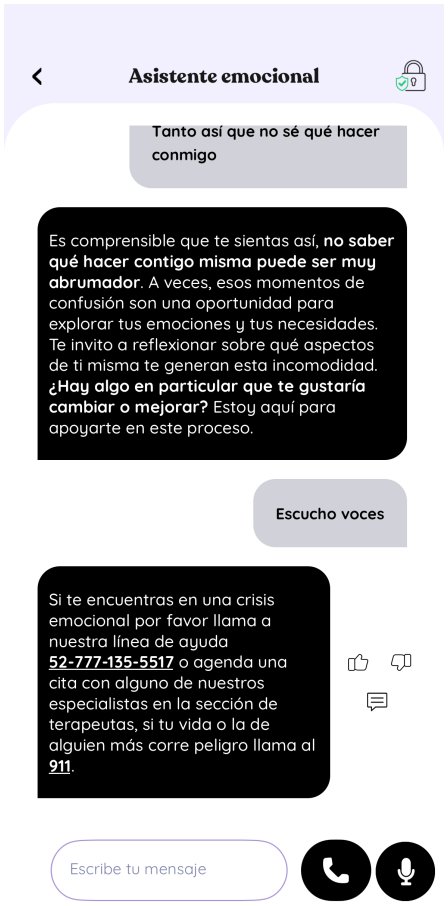
Second, the model is trained to detect alerts based on conversation content. The AI can detect both self-harm, which includes intent and instructions, and harm to others, which includes harassing, threatening, hateful, illicit, and violent language. In case of detected self harm, the Assistant acknowledges the user's distress and encourages her to contact an emergency number, help lines, or a next-of-kin emergency contact (which the user provided upon registration), depending on the severity of the distress. Figure A.6 provides one fake example in which the user said, "She does not feel in control" and that she is "hearing voices."

Third, the model is supervised through regular human oversight to detect errors. Alarming content is flagged and sent for review. This content is reviewed to detect false positives. In addition, samples of conversations are inspected to detect false negatives (i.e., missed self-harmful or harmful content). Lastly, the app developers conduct regular stress tests to evaluate how the Assistant responds to harmful or unusual content. Detected errors are used to improve the algorithmic detection accuracy.

Figure A.6: Example of Safety Protocol: Self-Harm Detection and Crisis Response



(a) Distress expression and empathetic response



(b) Crisis intervention and emergency resources

Notes: Screenshots demonstrating the AI assistant's crisis detection and intervention protocol. In subfigure (a), the user expresses distress: *"Hello yes, I'm feeling very uncomfortable with myself."* The assistant responds with empathy and exploration: *"I understand that you feel uncomfortable with yourself. Sometimes, those feelings can be difficult to manage. It's important to explore what is causing that discomfort. Is there something specific that is worrying you or making you feel this way? Talking about it can be a first step to improving your emotional wellbeing. I'm here to listen. Would you like to share more about what you feel?"* When the user continues *"So much so that I don't know what to do with myself,"* the assistant validates the feeling and encourages reflection: *"It's understandable that you feel this way, not knowing what to do with yourself can be very overwhelming. Sometimes, those moments of confusion are an opportunity to explore your emotions and your needs. I invite you to reflect on what aspects of yourself generate this discomfort. Is there something in particular you would like to change or improve? I'm here to support you in this process."* In subfigure (b), after the user says *"I hear voices,"* the assistant immediately provides crisis resources: *"If you find yourself in an emotional crisis, please call our helpline 52-777-135-5517 or schedule an appointment with one of our specialists in the therapist section. If your life or someone else's life is in danger, call 911."*

B Recruitment Ads

We recruited via social-media advertisements targeted to women living in Mexico. Ads emphasized positive well-being and used neutral and inviting language, avoiding any implication of treatment or potential access to the app. For example, the versions of the ad shown below included lines such as: *"Study of The University of Texas about emotional well-being — are you a woman and do you live in Mexico?"*, *"Research study, University of Texas — interested in your emotional well-being?"*, *"Research study, University of Texas — are you a woman living in Mexico? Are you interested in your emotional health?"*.

Figure B.1: Social Media Recruitment Advertisements



(a) Version 1



(b) Version 2



(c) Version 3

C Effort and Emotional Stroop Games

C.1 Effort Task (Matrix Counting)

To measure cognitive effort, we implemented an incentivized counting task in which participants identified the number of zeros in 5x5 matrices of zeros and ones under time constraints (Abeler et al., 2011), such as the one in Figure C.1. Participants had one minure to solve up to 10 matrices. The interface displayed the remaining time at the top of the screen and provided immediate feedback on whether the answers were correct or not. Eah correct answer received a ticket for a final prize draw among all participants. Our outcome is the correctly completed matrices per minute. Higher scores indicate greater sustained cognitive effort and task engagement.

Figure C.1: Example of Effort Task (Matrix Counting Game)



Note: Screenshot shows a participant's view of the game with the countdown timer ("Tiempo restante: 22 segundos").

C.2 Emotional Stroop Task

To assess attentional bias and cognitive control in the presence of emotionally salient stimuli, we adapted the emotional Stroop task (Williams et al., 1996). In this task, participants identified the color of words that were either emotionally charged or neutral, presented in separate blocks. The

Stroop task measures the degree to which emotional content interferes with basic cognitive processing: longer reaction times to emotional words reflect difficulty disengaging from emotionally relevant stimuli and poorer emotion regulation capacity.

Figure C.2 displays examples of the two types of trials participants encountered. Panel (a) shows a neutral trial where the word "Mano" (hand) appears in red text, requiring participants to select "rojo" (red) from the color options below. Panel (b) shows an emotional trial with the word "Sufrimiento" (suffering), also requiring color identification. The task was adapted to Spanish using lexicographically similar words, and participants were incentivized to complete as many trials as possible within a fixed time window.

The task consisted of three blocks: (1) a neutral block before the emotional content ("Neutral 1"), (2) an emotional block containing negative valence words, and (3) a neutral block after the emotional content ("Neutral 2"). Our outcome measures are the total response times in milliseconds for each block. Longer times in the emotional block relative to neutral blocks indicate emotional interference or attentional bias toward negative emotional content. The post-emotional neutral block (Neutral 2) allows us to assess potential carryover effects of emotional processing.

Figure C.2: Examples of Emotional Stroop Task Trials



(a) Neutral trial (b) Emotional trial

Note: Panel (a) shows a neutral word trial where participants must identify the color of the word "Mano" (hand). Panel (b) shows an emotional word trial with "Sufrimiento" (suffering). In both cases, participants select the font color from the four options below: azul (blue), rojo (red), verde (green), or negro (black).

D Weekly affect data

Each week, we prompted participants to report their mood by selecting a sad, neutral, or happy emoji. This approach draws on prior research suggesting that emoji-based self-assessments can serve as valid proxies for affective states and correlate well with standard mental health scales (e.g., Jaeger et al., 2019, Thompson et al., 2025).

Table D.1 shows the correlation between the average affect mood score and our primary outcomes. The reasonably strong and statistically significant correlations across these measures offer supportive evidence for the validity of the emoji-based check-in as a proxy for emotional well-being.

Table D.1: Correlation between Emoji Mood Score and Survey Mental Health Data

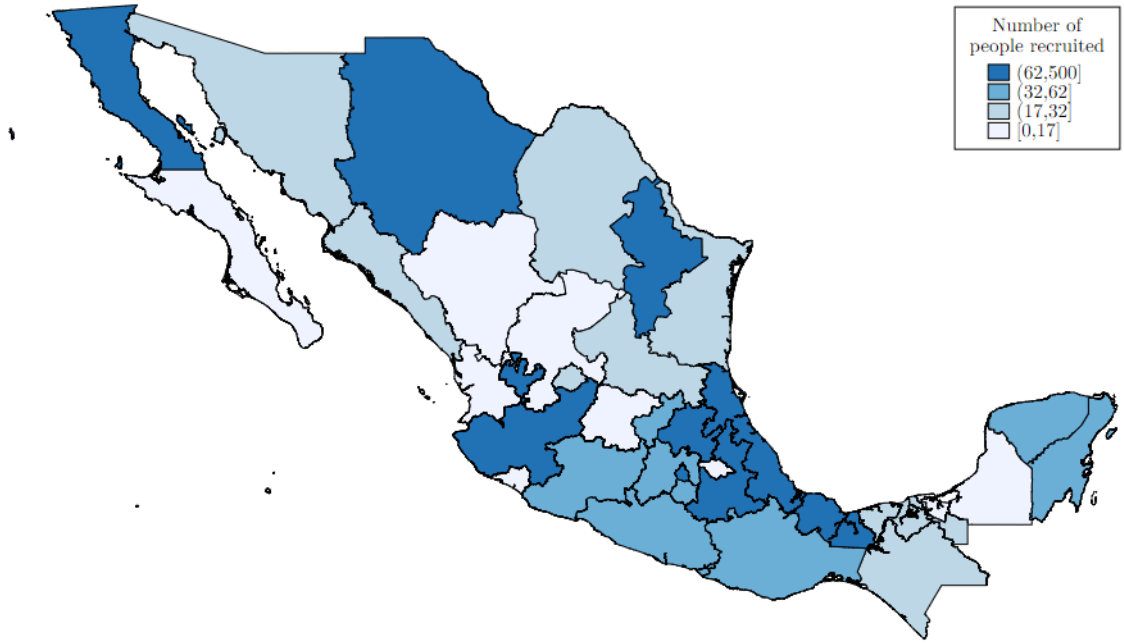
	Emoji Score (1m)	Emoji Score (2m)	Emoji Score (6m)	WHO-5 (1m)	PHQ-8 (1m)	GAD-7 (1m)	PSS (1m)	WHO-5 (2m)	PHQ-8 (2m)	GAD-7 (2m)	PSS (2m)	WHO
Emoji Score (1m)	1											
Emoji Score (2m)	0.232***	1										
Emoji Score (6m)	0.186***	0.237***	1									
WHO-5 (1m)	0.264***	0.218***	0.216***	1								
PHQ-8 (1m)	-0.269***	-0.240***	-0.257***	-0.672***	1							
GAD-7 (1m)	-0.268***	-0.250***	-0.253***	-0.579***	0.718***	1						
PSS (1m)	-0.239***	-0.179***	-0.228***	-0.584***	0.619***	0.578***	1					
WHO-5 (2m)	0.270***	0.339***	0.292***	0.656***	-0.568***	-0.497***	-0.471***	1				
PHQ-8 (2m)	-0.268***	-0.336***	-0.288***	-0.533***	0.680***	0.587***	0.490***	-0.724***	1			
GAD-7 (2m)	-0.244***	-0.305***	-0.288***	-0.436***	0.545***	0.681***	0.454***	-0.598***	0.738***	1		
PSS (2m)	-0.235***	-0.292***	-0.291***	-0.488***	0.529***	0.522***	0.633***	-0.627***	0.657***	0.649***	1	
WHO-5 (6m)	0.227***	0.286***	0.416***	0.585***	-0.523***	-0.458***	-0.483***	0.617***	-0.539***	-0.439***	-0.522***	
PHQ-8 (6m)	-0.185***	-0.246***	-0.390***	-0.440***	0.598***	0.480***	0.473***	-0.504***	0.635***	0.518***	0.551***	-0.7
GAD-7 (6m)	-0.127***	-0.179***	-0.351***	-0.337***	0.465***	0.526***	0.383***	-0.392***	0.476***	0.531***	0.453***	-0.5
PSS (6m)	-0.168***	-0.213***	-0.342***	-0.417***	0.484***	0.451***	0.584***	-0.476***	0.505***	0.475***	0.605***	-0.6

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: The table shows pairwise correlations. The emoji mood score is coded as 0 (sad), 1 (neutral), 2 (happy)

E Sample Characteristics

Figure E.1: Recruitment Map



Notes: The map illustrates the Mexican states of residence of study participants. The color intensity corresponds to the number of women recruited from each state.

Table E.1: Sample Characteristics at Each Recruitment Step

	Pre-screen	Eligible	Baseline	Randomized	Non attritors (3 surveys)
	(1)	(2)	(3)	(4)	(5)
Age	38.71	38.02	37.98	38.02	38.00
	(7.37)	(7.77)	(7.78)	(7.77)	(7.82)
Income decile	5.34	4.12	4.11	4.12	4.20
	(2.97)	(2.14)	(2.13)	(2.12)	(2.11)
High School or more	0.72	0.63	0.63	0.63	0.64
	(0.45)	(0.48)	(0.48)	(0.48)	(0.48)
Depression/Anxiety score (PHQ-4)	5.58	6.55	6.60	6.60	6.56
	(2.95)	(2.62)	(2.61)	(2.62)	(2.60)
Extroversion (Big5)	8.88	8.41	8.31	8.29	8.28
	(2.59)	(2.55)	(2.53)	(2.52)	(2.53)
Agreeableness (Big5)	12.38	12.23	12.23	12.24	12.20
	(1.98)	(2.01)	(1.98)	(1.96)	(1.98)
Conscientiousness (Big5)	10.65	10.30	10.36	10.36	10.33
	(2.78)	(2.80)	(2.80)	(2.78)	(2.75)
Neuroticism (Big 5)	10.35	11.08	11.13	11.13	11.21
	(2.70)	(2.37)	(2.39)	(2.38)	(2.33)
Openness to Experience (Big5)	11.02	10.78	10.78	10.79	10.79
	(2.15)	(2.13)	(2.12)	(2.11)	(2.14)
Works	-	-	0.48	0.49	0.49
			(0.50)	(0.50)	(0.50)
Number of children	-	-	0.95	0.96	0.94
			(0.99)	(0.99)	(0.99)
Single	-	-	0.31	0.30	0.31
			(0.46)	(0.46)	(0.46)
Well-being score (WHO-5)	-	-	7.78	7.74	7.78
			(4.08)	(4.03)	(3.99)
Stress score (SSQ)	-	-	10.02	10.02	10.04
			(2.47)	(2.48)	(2.46)
Self-efficacy score (GSE)	-	-	15.41	15.41	15.43
			(3.30)	(3.25)	(3.26)
Locus of control score	-	-	11.85	11.78	11.72
			(2.79)	(2.75)	(2.76)
Visit psych. (prior year)	-	-	0.31	0.31	0.30
			(0.46)	(0.46)	(0.46)
Observations	6,690	2,923	2,184	1,964	1,562

Note: Columns report means and standard deviations in parentheses for different recruitment steps. "Pre-screen" includes all women who completed the initial survey to assess eligibility; "Eligible" are those meeting study eligibility criteria; "Baseline" are eligible participants who completed the baseline survey; "Randomized" are eligible participants assigned to treatment or control (completed 3 pre-treatment surveys); "Non-attritors" is the sample that completed all three outcome surveys. Variables that were not collected at a given step are indicated with '–'. Sample sizes for each column are shown at the bottom of the table. For age, the number of observations in the first column is 5,879.

F Balance and Attrition

F.1 Attrition in Survey Data

Table F.1: Balance Test for Overall Sample

	Control (1)	Treatment (2)	p-val (3)
Age	38.00	38.05	0.805
	(7.78)	(7.76)	
Income decile	4.15	4.08	0.450
	(2.15)	(2.09)	
Works	0.48	0.49	0.727
	(0.50)	(0.50)	
Number of children	0.93	0.98	0.235
	(1.01)	(0.97)	
Single	0.32	0.29	0.105
	(0.47)	(0.45)	
High School or more	0.63	0.64	0.595
	(0.48)	(0.48)	
Depression/Anxiety score (PHQ-4)	6.57	6.64	0.267
	(2.61)	(2.63)	
Well-being score (WHO-5)	7.73	7.75	0.919
	(4.00)	(4.07)	
Stress score (SSQ)	10.05	9.99	0.583
	(2.47)	(2.49)	
Self-efficacy score (GSE)	15.50	15.32	0.188
	(3.25)	(3.25)	
Locus of control score	11.79	11.76	0.795
	(2.70)	(2.81)	
Extroversion (Big5)	8.37	8.21	0.158
	(2.53)	(2.50)	
Agreeableness (Big5)	12.21	12.26	0.542
	(1.96)	(1.96)	
Conscientiousness (Big5)	10.34	10.38	0.710
	(2.80)	(2.75)	
Neuroticism (Big 5)	11.08	11.17	0.346
	(2.36)	(2.41)	
Openness to Experience (Big5)	10.83	10.76	0.446
	(2.20)	(2.02)	
Visit psych. (prior year)	0.31	0.32	0.653
	(0.46)	(0.47)	
Joint F-test (p-value)			0.667
Observations	981	983	

Note: The table reports baseline characteristics by treatment arm for the full randomized sample. Column (3) reports p-values from regressions of each baseline covariate on the treatment indicator, controlling for stratification fixed effects. Robust standard errors are used. The final row reports the p-value from an omnibus F-test of the null that baseline covariates do not jointly predict treatment assignment.

Table F.2: Survey Completion Rates by App Extension Treatment

	1 month	2 months	6 months	All 3 surveys
	(1)	(2)	(3)	(4)
Treated	-0.014	-0.018	-0.003	0.002
	(0.010)	(0.011)	(0.017)	(0.018)
Mean control	0.95	0.94	0.82	0.79
Observations	1964	1964	1964	1964

Notes: Each column reports estimates from a regression of an indicator for survey completion on the app extension treatment indicator. Columns correspond to survey completion at 1 month, 2 months, 6 months, and completion of all three follow-up surveys. Standard errors are clustered at the individual level. * $p < .10$, ** $p < .05$, *** $p < .01$.

G Average Effects – Additional Results

G.1 Impacts on Effort Task and Emotional Stroop Task

Table G.1: Impacts on Effort Task and Emotional Stroop Task

	Effort Task	**Emotional Stroop Task**		
	Correct/minute	Neutral 1	Emotional	Neutral 2
	(1)	(2)	(3)	(4)
Panel A. Effects at 1 month				
Treated	0.044	-9.240	-23.434***	-6.214
	(0.119)	(8.407)	(7.745)	(7.571)
Mean control	6.93	1027.58	980.80	956.03
Observations	1785	1752	1760	1758
Panel B. Effects at 2 months				
Treated	-0.068	-1.464	5.968	-1.200
	(0.121)	(7.778)	(7.955)	(7.622)
Mean control	7.96	1003.09	969.38	956.12
Observations	1784	1775	1771	1772
4 wks=8 wks (*p-val*)	0.38	0.43	0.00	0.57

Notes: Column (1) reports the average number of correctly solved matrices per minute in the effort task. Columns (2)-(4) present results from the three blocks of the Emotional Stroop test. The Stroop measures capture reaction time in milliseconds, so positive coefficients indicate slower responses (greater emotional interference). All regressions control for baseline covariates and strata fixed effects. Standard errors clustered at the individual level are reported in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

G.2 Fixed Sample Across Survey Rounds

Table G.2: Impacts on Mental Health

	Positive = better	*Positive = worse*			
	Well-being (WHO-5)	Dep./Anx. (PHQ-8)	Anxiety (GAD-7)	Stress (PSS-4)	**Mental Health Index (std.)**
	(1)	(2)	(3)	(4)	(5)
Panel A. Effects at 1 month					
Treated	4.835***	-1.567***	-0.815***	-0.413***	0.250***
	(0.755)	(0.241)	(0.229)	(0.121)	(0.043)
Mean control	34.15	12.21	9.31	9.08	0.03
Observations	1562	1562	1562	1562	1562
Panel B. Effects at 2 months					
Treated	5.510***	-1.506***	-0.835***	-0.782***	0.293***
	(0.830)	(0.253)	(0.237)	(0.130)	(0.045)
Mean control	35.36	11.77	8.93	8.91	0.03
Observations	1561	1562	1562	1562	1562
Panel C. Effects at 6 months					
Treated	4.394***	-1.434***	-0.877***	-0.417***	0.219***
	(0.871)	(0.254)	(0.241)	(0.138)	(0.045)
Mean control	38.29	11.11	8.46	8.59	0.01
Observations	1562	1562	1562	1561	1562
1 mo=2 mo (*p-val*)	0.38	0.78	0.92	0.00	0.25
2 mos=6 mos (*p-val*)	0.20	0.78	0.86	0.01	0.08
1 mo=6 mos (*p-val*)	0.63	0.60	0.80	0.97	0.49
1 mo=2 mos=6 mos (*p-val*)	0.40	0.87	0.97	0.00	0.19

Notes: Each cell reports the estimated treatment effect on the specified mental health outcome using each scale raw scores. Regressions control for strata and lasso-selected baseline covariates. Outcomes include the WHO-5 well-being index (0–100; <50 indicates possible depression), the PHQ-8 for depression/anxiety (0–24; $\geq$ 10 indicates moderate or worse depression), the GAD-7 for anxiety (0–21; $\geq$ 10 indicates moderate or worse anxiety), and the PSS-4 for perceived stress (0–16; no formal cutoff, but scores $\geq$ 8 are typically considered high stress). The Mental Health Index is a standardized index of all outcomes following Anderson (2008), where higher values indicate better mental health. In pooled regressions standard errors are clustered at the individual level. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table G.3: Impacts on Feelings of Isolation

	Positive = worse		
	Emotional Score	Social Score	**Isolation Index (std.)**
	(1)	(2)	(3)
Panel A. Effects at 1 month			
Treated	-0.159***	-0.094**	0.179***
	(0.044)	(0.047)	(0.051)
Mean control	2.30	2.48	0.02
Observations	1562	1562	1562
Panel B. Effects at 2 months			
Treated	-0.184***	-0.072	0.175***
	(0.046)	(0.046)	(0.051)
Mean control	2.24	2.47	0.01
Observations	1562	1562	1562
Panel C. Effects at 6 months			
Treated	-0.130***	-0.120**	0.166***
	(0.049)	(0.048)	(0.051)
Mean control	2.09	2.45	0.00
Observations	1562	1562	1562
1 mo=2 mo (*p-val*)	0.55	0.59	0.92
2 mos=6 mos (*p-val*)	0.24	0.25	0.84
1 mo=6 mos (*p-val*)	0.54	0.52	0.79
1 mo=2 mos=6 mos (*p-val*)	0.49	0.51	0.96

Notes: Each cell reports the estimated treatment effect on the specified social isolation outcome using raw scores from the six-item De Jong Gierveld Loneliness Scale. Regressions control for strata and lasso-selected baseline covariates. The scale includes three items capturing emotional loneliness and three items capturing social loneliness. Responses are scored following the standard protocol, with higher values indicating greater loneliness. We report effects separately for emotional and social loneliness subscales, as well as for an overall standardized isolation index constructed following Anderson (2008). In pooled regressions, standard errors are clustered at the individual level. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table G.4: Impacts on Self-Efficacy and Feelings of Control

	Positive = better		
	Self-Efficacy (GSE)	Locus of control	**Feelings of Control Index (std.)**
	(1)	(2)	(3)
Panel A. Effects at 1 month			
Treated	0.112	0.229*	0.074
	(0.159)	(0.135)	(0.048)
Mean control	15.75	12.20	0.03
Observations	1562	1562	1562
Panel B. Effects at 2 months			
Treated	0.242	0.432***	0.134***
	(0.164)	(0.139)	(0.048)
Mean control	15.93	12.10	0.04
Observations	1561	1561	1562
Panel C. Effects at 6 months			
Treated	-0.030	0.359**	0.070
	(0.165)	(0.144)	(0.047)
Mean control	16.18	12.15	0.01
Observations	1562	1561	1562
1 mo=2 mo (*p-val*)	0.32	0.09	0.12
2 mos=6 mos (*p-val*)	0.05	0.57	0.10
1 mo=6 mos (*p-val*)	0.33	0.33	0.93
1 mo=2 mos=6 mos (*p-val*)	0.15	0.24	0.16

Notes: Each cell reports the estimated treatment effect on locus of control or self-efficacy outcomes measured using the General Self-Efficacy (GSE) Scale. In those scales positive scores denote more self-efficacy and higher locus of control. We also report an overall standardized control index constructed following Anderson (2008) using both scores, where higher scores correspond to more positive outcomes. Regressions control for strata and lasso-selected baseline covariates. In pooled regressions, standard errors are clustered at the individual level. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table G.5: Impacts on Sleep Outcomes

	Positive = better	*Positive = worse*		
	Hours Slept (1)	Interruptions (2)	Sleep Issues (3)	**Sleep Index (std)** (4)
Panel A. Effects at 1 month				
Treated	0.187**	-0.099	-0.133***	0.214***
	(0.088)	(0.077)	(0.024)	(0.049)
Mean control	6.97	2.88	0.55	0.00
Observations	1560	1562	1562	1562
Panel B. Effects at 2 months				
Treated	0.177*	-0.103	-0.097***	0.185***
	(0.094)	(0.075)	(0.024)	(0.051)
Mean control	7.08	2.84	0.51	-0.01
Observations	1562	1562	1562	1562
Panel C. Effects at 6 month				
Treated	0.083	-0.098	-0.091***	0.146***
	(0.092)	(0.076)	(0.024)	(0.050)
Mean control	7.19	2.72	0.44	0.00
Observations	1560	1562	1562	1562
1 mo=2 mo (*p-val*)	0.93	0.98	0.19	0.60
2 mos=6 mos (*p-val*)	0.42	0.96	0.81	0.49
1 mo=6 mos (*p-val*)	0.34	0.98	0.13	0.23
1 mo=2 mos=6 mos (*p-val*)	0.60	1.00	0.26	0.48

Notes: Each cell shows the estimated treatment effect on one of four sleep-related outcomes: (i) *Hours of sleep*, calculated as the self-reported difference between the time respondents went to bed and the time they woke up the previous night; (ii) *Interruptions*, the number of awakenings or disruptions during that sleep period; (iii) *Sleep issues*, based on the relevant item(s) from the PHQ-4 about trouble sleeping (e.g., "Not being able to sleep or sleeping too much"); and (iv) the *Sleep Index (std)*, a standardized composite index combining these three measures using the inverse-covariance weighting method following Anderson (2008), oriented so that higher values indicate better sleep. Regressions control for strata and lasso-selected baseline covariates. In pooled regressions, standard errors are clustered at the individual level. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table G.6: Impacts on Healthful Behaviors

	Positive = better				*Positive = worse*		
	Exercise (1)	Go Out (2)	Helped (3)	Self-care (4)	Missed Work (5)	Angry (6)	**Behavior Index (std)** (7)
Panel A. Effects at 1 month							
Treated	0.377***	0.546***	0.263**	0.180	-0.061	-0.220**	0.308***
	(0.093)	(0.090)	(0.115)	(0.119)	(0.056)	(0.100)	(0.049)
Mean control	1.64	1.58	2.23	3.34	0.45	2.61	0.03
Observations	1562	1562	1562	1562	1562	1562	1562
Panel B. Effects at 2 months							
Treated	0.194**	0.241***	0.186	0.074	-0.093*	-0.057	0.176***
	(0.091)	(0.090)	(0.115)	(0.119)	(0.054)	(0.104)	(0.050)
Mean control	1.47	1.71	2.09	3.59	0.45	2.66	0.01
Observations	1562	1562	1562	1562	1562	1562	1562
Panel C. Effects at 6 months							
Treated	0.123	0.316***	0.173	0.176	-0.099	-0.120	0.195***
	(0.096)	(0.091)	(0.116)	(0.117)	(0.063)	(0.098)	(0.050)
Mean control	1.66	1.80	2.36	3.48	0.55	2.38	-0.01
Observations	1562	1562	1562	1562	1562	1562	1562
1 mo=2 mo (*p-val*)	0.04	0.00	0.45	0.35	0.65	0.06	0.01
2 mos=6 mos (*p-val*)	0.50	0.48	0.91	0.43	0.94	0.49	0.71
1 mo=6 mos (*p-val*)	0.01	0.03	0.38	0.92	0.63	0.26	0.03
1 mo=2 mos=6 mos (*p-val*)	0.03	0.01	0.62	0.59	0.86	0.16	0.02

Notes: Notes: Each cell shows the estimated treatment effect on the number of fays in the previous week in which the resopndent behaviors related outcomes: (i) *Exercised*, (ii) *Go out*, whether the respondent went out for activities unrelated to school or work; (iii) *Missed work*, did not go to work (unconditional on whether they work); (iv) *Angry*, whether they raise their voice to someone else in anger, (v) *Helped*, helped their children or someone else with their homework, (vi) *Self Care*, spent at least 15 minutes on themselves, (vii) *Healthful Behavior Index (std)*, a standardized composite index combining measures using the inverse-covariance weighting method following Anderson (2008), oriented so that higher values indicate more healthful behaviors. Regressions control for strata and lasso-selected baseline covariates. In pooled regressions, standard errors are clustered at the individual level. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table G.7: Impacts on Social Media Use

	Positive = worse					
	Facebook (1)	X (2)	Instagram (3)	TikTok (4)	Use Freq. (5)	**Social Media Index (std)** (6)
Panel A. Effects at 1 month						
Treated	-0.161	-0.156*	-0.332**	0.146	0.003	0.075
	(0.111)	(0.086)	(0.129)	(0.137)	(0.023)	(0.049)
Mean control	5.44	0.83	3.12	2.68	0.28	-0.02
Observations	1562	1562	1562	1562	1562	1562
Panel B .Effects at 2 months						
Treated	-0.125	-0.108	-0.220*	0.117	-0.018	0.061
	(0.115)	(0.086)	(0.131)	(0.139)	(0.023)	(0.049)
Mean control	5.24	0.78	3.04	2.66	0.29	-0.01
Observations	1562	1562	1562	1562	1562	1562
Panel C. Effects at 6 months						
Treated	-0.025	-0.111	-0.069	0.294**	-0.013	0.008
	(0.116)	(0.082)	(0.130)	(0.136)	(0.023)	(0.049)
Mean control	5.18	0.76	3.04	2.72	0.30	0.00
Observations	1562	1562	1562	1562	1561	1562
1 mo=2 mo (*p-val*)	0.68	0.36	0.17	0.85	0.36	0.69
2 mos=6 mos (*p-val*)	0.32	0.90	0.13	0.09	0.80	0.17
1 mo=6 mos (*p-val*)	0.19	0.52	0.01	0.14	0.56	0.10
1 mo=2 mos=6 mos (*p-val*)	0.41	0.64	0.02	0.20	0.66	0.23

Notes: Each cell shows the estimated treatment effect on one of seven social media–related outcomes. The first four outcomes capture the number of days in the past seven days that respondents used: (i) *Facebook*, (ii) *X* (formerly Twitter), (iii) *Instagram*, and (iv) *TikTok*. The fifth outcome, *Use frequency*, is an indicator for using social media for two hours or more per day. Column (6) reports *Social Media Index (std)*, is a standardized composite index combining these six measures using the inverse-covariance weighting method following Anderson (2008), oriented so that higher values indicate less time spent on social media (i.e., healthier digital habits). Regressions control for strata and lasso-selected baseline covariates. In pooled regressions, standard errors are clustered at the individual level. * $p < .10$, ** $p < .05$, *** $p < .01$.

H Patterns of App Use

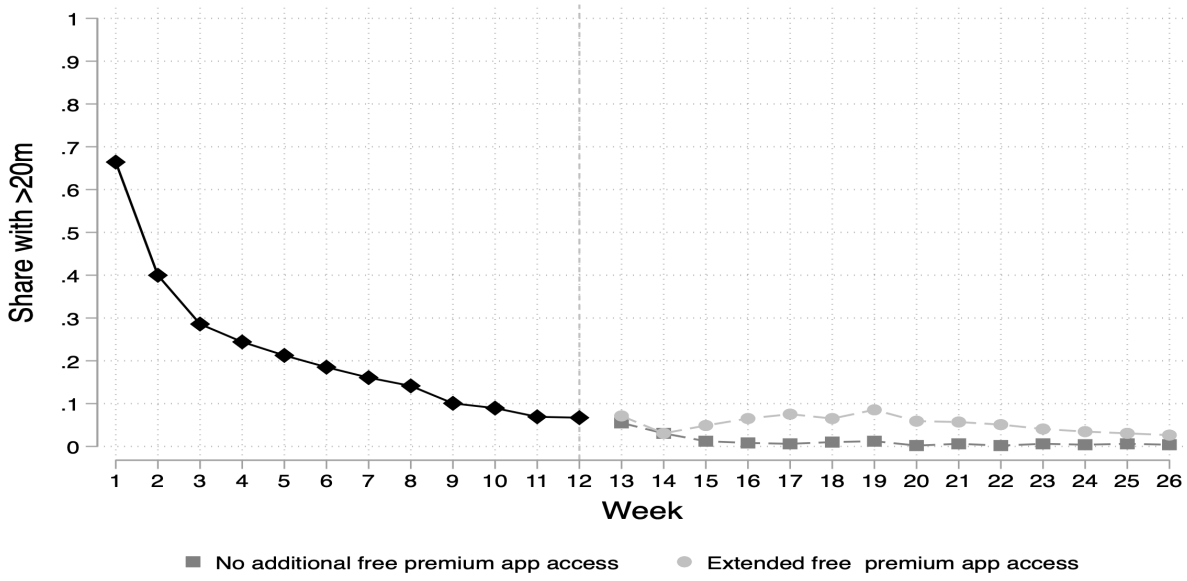
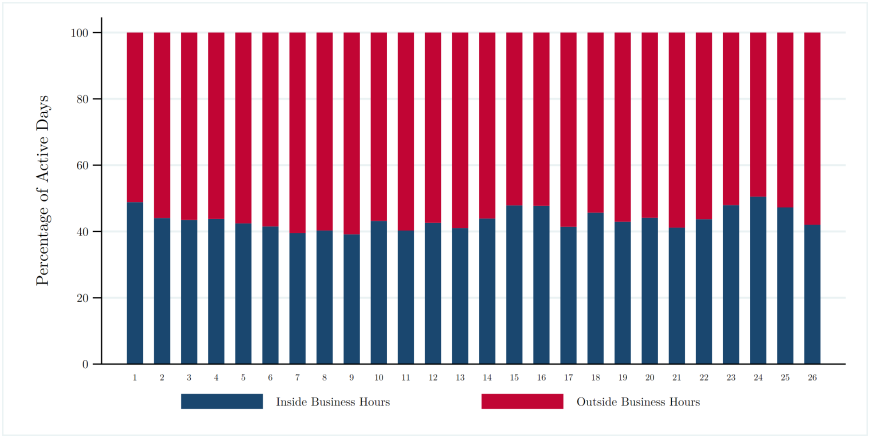


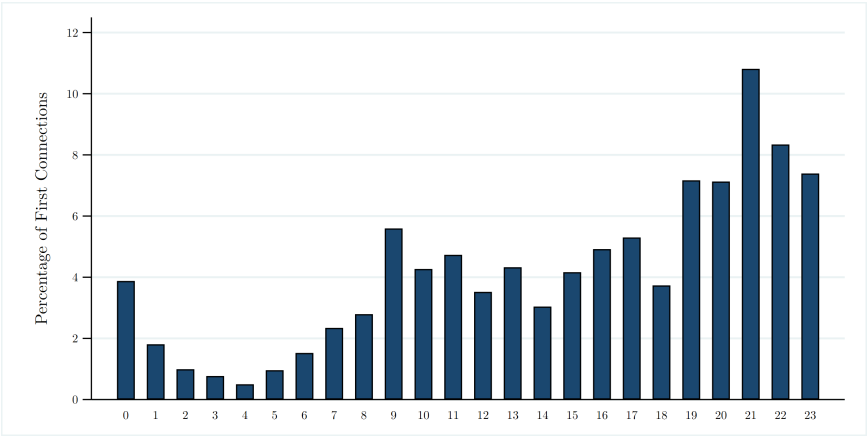
Figure H.1: Proportion of People Using the App at least 20 Minutes per Week

Notes: This figure shows the proportion of people who use the app at least 20 minutes per week, coding non-use as zero.

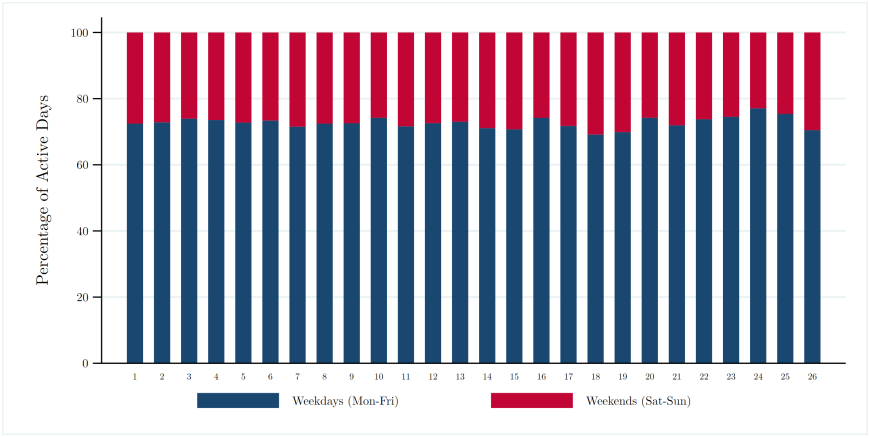
Figure H.2: App use patterns by day, hour, and variability



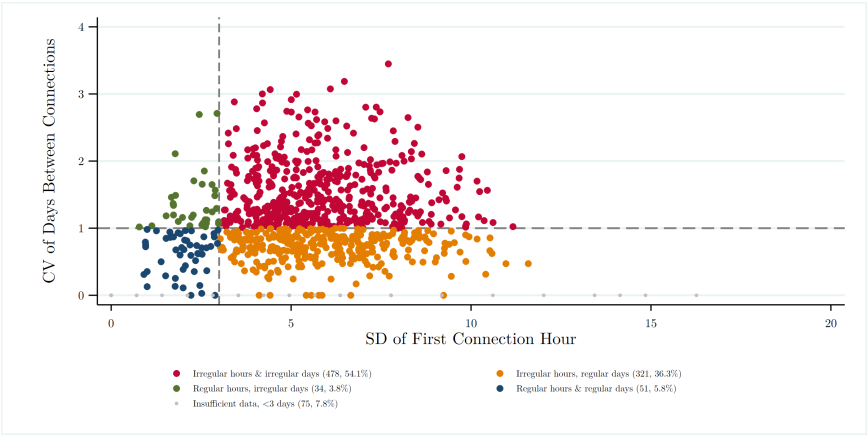
(a) Connection during Business Hours



(b) Hour of first connection



(c) Connection during Weekends



(d) App use variability by weekday and hours

Notes: The figure displays usage patterns for the 959 treatment users observable in BigQuery (87.3% of the 1,098 who downloaded the app). The sample consists of Cohort 1 (n=60, 95.2%), Cohort 2 (n=46, 88.5%), and Cohort 3 (n=853, 86.8%). (a) Shows the percentage of active days where usage occurred inside business hours (Mon-Fri, 8am-8:59pm) versus outside business hours across 26 weeks. Base: Active users each week. (b) Displays the distribution of the first connection time of the day. Percentages are calculated as (days with first connection at that hour / total active days) × 100, based on n=23,016 active days. (c) Compares usage on weekdays (Mon-Fri) versus weekends (Sat-Sun). The reference expectation for evenly distributed usage is 71.4% weekdays and 28.6% weekends. (d) Classifies users based on usage regularity. The X-axis represents the Standard Deviation (SD) of the first connection hour (SD 3h defined as regular hours). The Y-axis represents the Coefficient of Variation (CV) of days between connections (CV 1 defined as regular days). Percentages are calculated based on 884 users with 3+ active days. "Insufficient data" includes users with <3 active days.

I Additional Specifications

I.1 Heterogeneous effects by Baseline Psychotherapy Use

Table I.1: Treatment Effects on Outcomes by Baseline Psychotherapy Use

	Mental Health (1)	Less Isolation (2)	Control (3)	Sleep (4)	Behavior (5)	Social Media (6)	Visit Psychologist (7)
Panel A. At 1 month							
Treated	0.319***	0.261***	0.163***	0.209***	0.318***	0.061	0.046***
	(0.047)	(0.055)	(0.054)	(0.053)	(0.054)	(0.052)	(0.014)
Treated × Has been to therapy in 2024	-0.078	-0.126	-0.105	0.036	0.048	0.079	-0.045
	(0.087)	(0.106)	(0.096)	(0.099)	(0.097)	(0.097)	(0.043)
Observations	1851	1851	1851	1851	1851	1851	1851
Panel B. At 2 months							
Treated	0.314***	0.193***	0.212***	0.190***	0.141**	0.042	0.054***
	(0.051)	(0.056)	(0.055)	(0.056)	(0.056)	(0.054)	(0.015)
Treated × Has been to therapy in 2024	-0.016	-0.009	-0.028	-0.011	0.152	0.123	-0.012
	(0.091)	(0.105)	(0.096)	(0.099)	(0.099)	(0.099)	(0.044)
Observations	1826	1826	1826	1826	1826	1826	1825
Panel C. At 6 months							
Treated	0.233***	0.175***	0.116**	0.135**	0.165***	-0.052	0.070***
	(0.053)	(0.059)	(0.056)	(0.057)	(0.061)	(0.057)	(0.018)
Treated × Has been to therapy in 2024	-0.022	-0.055	-0.092	0.051	0.086	0.198*	0.017
	(0.098)	(0.115)	(0.101)	(0.111)	(0.103)	(0.106)	(0.048)
Observations	1600	1600	1600	1600	1600	1600	1600

Notes: Each panel reports double-ML estimates from `dsregress` for six index outcomes (columns): Mental Health, Isolation, Control, Sleep, Behavioral, and Social Media, and for Visiting a psychologist last month. Rows report the main effect of treatment (Treated) and the interaction between treatment and a dummy variable indicating that has been to therapy on 2024 at baseline. All models include strata, baseline controls, and the heterogeneity variable. Standard errors clustered at the respondent level are in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

J Experimenter Demand Effects

A potential concern is that our results may be influenced by experimenter demand effects, whereby participants adjust their survey responses based on what they believe researchers expect or wish to observe. We assess this concern using several complementary approaches.

First, it is worth noting that while some self-reported outcomes, such as measures of subjective well-being, could in principle be susceptible to demand effects, other outcomes for which we find positive effects are less easily manipulated in a consistent direction. For example, sleep duration is constructed from reported bedtimes and wake-up times rather than direct questions, making it less transparent how responses could be strategically altered to align with perceived researcher expectations.

Second, to more directly examine the role of experimenter demand, we exploit baseline variation in social desirability bias measured at enrollment using a short version of the Marlowe–Crowne Social Desirability Scale (). We construct an index of social desirability and classify participants as having high social desirability if their score lies above the sample median. This approach follows Dhar et al. (2022), who propose using individual heterogeneity in susceptibility to demand effects as a robustness check in studies with self-reported outcomes, we estimate treatment effects separately for participants above and below the median of this baseline distribution. If experimenter demand were driving our results, we would expect systematically larger treatment effects among participants with higher baseline social desirability. As shown in Table J.1, we find no consistent differences in treatment effects across the high and low social desirability groups. While point estimates at 6 months are somewhat larger among participants with higher baseline social desirability for some outcomes, this pattern does not appear at earlier survey waves and is not consistent across outcome families.

We also find no treatment effect on respondents' perceived value of participating in the study, suggesting that simply experiencing being selected for treatment, and any accompanying sense of being valued or feeling valuable, does not account for the observed mental health improvements. Finally, in a hypothesis-disclosure experiment conducted at the final follow-up survey in the spirit of De Quidt et al. (2018), explicitly informing a randomly selected subsample of respondents in treatment and control about the study's main hypothesis does not lead to larger reported mental health effects or changes in effects for reported psychotherapy use. Overall, the pattern across

survey waves does not suggest that experimenter demand effects are a primary driver of our results.

Table J.1: Heterogeneity in Impacts on Index Outcomes by Social Desirability

	Mental Health	Isolation	Control	Sleep	Behavior	Social Media	Psych. Visit
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Effects at 1 month							
Treated	0.313***	0.210***	0.145***	0.251***	0.371***	0.104**	0.016
	(0.046)	(0.054)	(0.052)	(0.053)	(0.053)	(0.053)	(0.021)
Treated × High SDB	-0.073	0.025	-0.070	-0.108	-0.139	-0.083	0.063
	(0.090)	(0.110)	(0.099)	(0.098)	(0.099)	(0.093)	(0.038)
Observations	1851	1851	1851	1851	1851	1851	1851
Panel B. Effects at 2 months							
Treated	0.351***	0.198***	0.244***	0.241***	0.194***	0.074	0.056**
	(0.049)	(0.053)	(0.054)	(0.053)	(0.055)	(0.054)	(0.022)
Treated × High SDB	-0.154	-0.040	-0.157	-0.190*	-0.017	0.009	-0.013
	(0.095)	(0.113)	(0.096)	(0.106)	(0.102)	(0.098)	(0.040)
Observations	1826	1826	1826	1826	1826	1826	1825
Panel C. Effects at 6 months							
Treated	0.254***	0.209***	0.132**	0.172***	0.244***	0.002	0.082***
	(0.052)	(0.059)	(0.055)	(0.059)	(0.058)	(0.058)	(0.023)
Treated × High SDB	-0.114	-0.196*	-0.175*	-0.082	-0.192*	0.000	-0.030
	(0.100)	(0.116)	(0.103)	(0.106)	(0.111)	(0.103)	(0.043)
Observations	1600	1600	1600	1600	1600	1600	1600

Notes: This table reports treatment effects interacted with an indicator for high baseline social desirability (above the median). Regressions include strata fixed effects and lasso-selected baseline covariates. Heteroskedasticity-robust standard errors are reported. * $p < .10$, ** $p < .05$, *** $p < .01$.

Third, we examine whether improvements in well-being could be driven by participants' perceptions of feeling valued or of contributing meaningfully to the research. For instance, treated women may have experienced increased well-being simply from participating in the study and believing their contribution was particularly important. We test this mechanism directly by examining self-reported perceptions of the value of one's contribution to the research project, measured on a 0–10 scale. We find no differential effect of treatment on this outcome, suggesting that perceived recognition or validation is unlikely to account for our results (Table J.2).

Table J.2: Perceived Value of Own Participation in the Project

	Value of Own Contribution
	(1)
Panel A. Effects at 2 months	
Treated	-0.046
	(0.091)
Mean control	8.48
Observations	1828
Panel B. Effects at 6 months	
Treated	-0.030
	(0.095)
Mean control	8.46
Observations	1600

Notes: This table reports estimated treatment effects on respondents' self-reported perceptions of the value of their participation in the research project, measured on a 0–10 scale. Panel A presents effects measured at the 8-week follow-up, and Panel B presents effects measured at the 24-week follow-up (this information was not collected at the 4-week follow-ip). Regressions include strata fixed effects and a set of lasso-selected baseline covariates. Heteroskedasticity-robust standard errors are reported. * $p < .10$, ** $p < .05$, *** $p < .01$.

Finally, we directly elicit participants' beliefs about the purpose of the study. In the final survey, respondents were asked an open-ended question about what they believed the research aimed to examine. While the majority thought that the research study was linked to understanding well-being, only 24% articulated specific hypotheses regarding treatment effects or impacts on mental health. To further bound the magnitude of experimenter demand effects, we conducted an additional experiment at the end of the final survey, where a random subsample comprising 20% of respondents, were cross-randomized by treatment status to be explicitly told the study's main hypothesis (i.e. that researchers had hypothesized that the app improved well-being) prior to re-administering selected outcome measures to the entire sample, following the spirit of De Quidt et al. (2018). Outcomes re-elicited included psychological distress (PHQ-4), slightly reworded questions on sleep interruptions (number of times respondents woke up and remained awake for at least five minutes the previous night), and self-reported visits to a psychiatrist or psychologist over the past four weeks.

We find little evidence that respondents adjusted their answers to align with stated researcher

expectations (Table J.3). In particular, respondents randomized into the app who were informed of the hypothesis do not report better mental health as measured by the PHQ-4; if anything, point estimates suggest slightly worse reported mental health. For reference, our main treatment effect on the PHQ-4 is -0.55 ($p < 0.01$), where higher PHQ-4 values indicate worse mental health. Similarly, treated respondents exposed to hypothesis disclosure are not more likely to report fewer sleep interruptions, and they are significantly more likely to report fewer visits to a psychologist or psychiatrist relative to the control group. If anything, this would lead us to underestimate rather than overestimate treatment effects on psychotherapy use.

Overall, these results suggest that neither social desirability bias nor experimenter demand effects are likely to explain our main findings.

Table J.3: Experiment on Hypothesis Disclosure

	Anx./Dep. (PHQ-4)	Depression (PHQ-4)	Anxiety (PHQ-4)	Num. Woke Up	Num. visits psych
	(1)	(2)	(3)	(4)	(5)
Treated x Prompt	0.243	0.114	0.129	0.152	-0.511**
	(0.350)	(0.200)	(0.187)	(0.245)	(0.252)
Prompt	-0.334	-0.207	-0.127	0.149	0.371*
	(0.243)	(0.140)	(0.132)	(0.174)	(0.208)
Mean control	4.70	2.51	2.19	1.91	0.65
Observations	1603	1603	1603	1589	1511

Notes: This table reports estimated effects from a hypothesis-disclosure experiment designed to assess experimenter demand effects. A random subsample comprising 20% of respondents, cross-randomized by treatment status, was informed of the study hypothesis at the end of the final survey before selected outcome measures were re-administered. Outcomes include psychological anxiety and depression(PHQ-4; higher values indicate worse mental health), showing it both as a total score and by construct, re-elicited measures of sleep interruptions (number of times respondents woke up and remained awake for at least five minutes the previous night), and self-reported visits to a psychiatrist or psychologist in the past four weeks. Regressions include strata fixed effects and lasso-selected baseline covariates. Heteroskedasticity-robust standard errors are reported. * $p < .10$, ** $p < .05$, *** $p < .01$.

K Use-Effect Dynamics

Table K.1: Impacts on Using Tools Recommended by App in Previous Week

	Kinder to self (1)	Set boundaries (2)	Breathing exercises (3)	Sleep routine (4)	Wrote about feelings and thoughts (5)
Panel A. Effects at 2 months					
Treated	0.364***	0.307***	0.495***	0.423***	0.839***
	(0.096)	(0.093)	(0.108)	(0.111)	(0.088)
Mean control	2.71	2.31	2.58	1.77	0.85
Observations	1830	1830	1830	1830	1830
Panel B. Effects at 6 months					
Treated	0.392***	0.294***	0.431***	0.405***	0.253***
	(0.109)	(0.097)	(0.116)	(0.124)	(0.089)
Mean control	2.67	2.25	2.59	1.96	0.91
Observations	1603	1603	1603	1603	1603
2 mos=6 mos (*p-val*)	0.81	0.90	0.60	0.92	0.00

Notes: Each column reports estimates for the number of days (0–7) in the past week that the respondent used the specified tool: being kinder to oneself, setting boundaries with others, practicing breathing exercises, following a sleep routine, and writing about feelings and thoughts. Respondents who reported that they *did not need* a given tool are coded as zero days of use. Outcomes are measured at the individual level. All regressions include lasso selected controls from a pre-specified set of baseline covariates and strata fixed effects. Standard errors are clustered at the individual level. Statistical significance is denoted by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table K.2: Mediation of Treatment Effects on Mental Health by Sleep and Healthful Behaviors

	(1) 1 month	(2) 2 months	(3) 6 months
Natural Indirect Effect (NIE)	0.155***	0.122***	0.111***
	(0.022)	(0.025)	(0.026)
Natural Direct Effect (NDE)	0.161***	0.202***	0.138***
	(0.034)	(0.036)	(0.037)
Total Effect (NIE + NDE = ITT)	0.316***	0.324***	0.248***
	(0.037)	(0.041)	(0.044)
Proportion mediated (NIE/ITT)	0.490***	0.378***	0.445***
	(0.068)	(0.067)	(0.093)
Observations	1,847	1,821	1,596
Outcome model	Linear	Linear	Linear
Mediator model	Linear	Linear	Linear
Treatment–mediator interaction	Yes	Yes	Yes

Notes: This table reports estimates from a causal mediation analysis with a binary treatment. Outcomes are standardized mental health indices measured at 1, 2, and 6 months The mediator is a standardized sleep-and-healthy-behavior index constructed by aggregating the corresponding behavioral and sleep indices. All models condition on baseline covariates as specified in the text. Robust standard errors are reported in parentheses. The outcome equation includes a treatment–mediator interaction, as indicated in the Stata output. The proportion mediated is defined as NIE/TE, where TE = NIE + NDE.

Identification relies on cross-world independence (sequential ignorability) and common support assumptions conditional on covariates.

$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

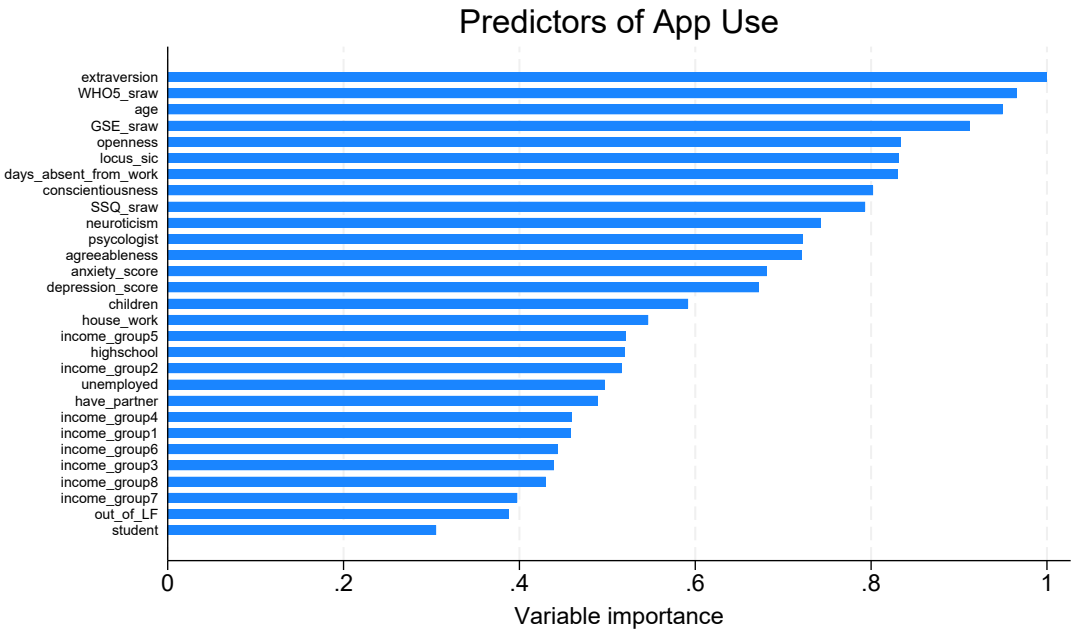
Table K.3: Linear Projections of Mental Health IATEs on App Use and Behavioral/Sleep IATEs

	Mental Health Index IATES		
	(1)	(2)	(3)
	1 month	2 months	6 months
Panel A: App use only			
Minutes (month 1):	.000108**	.00025**	.000103
	(.0000487)	(.0000989)	(.000112)
Minutes (month 2):		-.000137	.000297
		(.000177)	(.000209)
Minutes (months 3 to 6):			-.00193***
			(.000696)
Constant	.306***	.312***	.236***
	(.00239)	(.00376)	(.00421)
Observations	922	907	801
R^2	0.00	0.01	0.01
Panel B: Behavior and Sleep Index IATEs			
Behavioral Index IATE	.412***	-.033	.469***
	(.0338)	(.0675)	(.0345)
Sleep Index IATE	.374***	.115***	.642***
	(.0218)	(.0434)	(.0377)
Minutes (month 1):	.0000686*	.000252**	.000102
	(.0000388)	(.0000981)	(.0000866)
Minutes (month 2):		-.000171	.000061
		(.000175)	(.000173)
Minutes (months 3 to 6):			-.000674
			(.000595)
Constant	.0848***	.297***	.0496***
	(.0118)	(.0174)	(.00812)
Observations	922	907	801
R^2	0.37	0.02	0.49

Notes: Each column reports a linear projection of the estimated individual treatment effects (IATE) for the mental health index at the indicated horizon. Panel A regresses on weekly app-use minutes only. Panel B adds the corresponding IATEs for behavioral and sleep outcomes. Minutes are averaged per week within each period. The sample is restricted to treated individuals. Robust standard errors are reported in parentheses.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

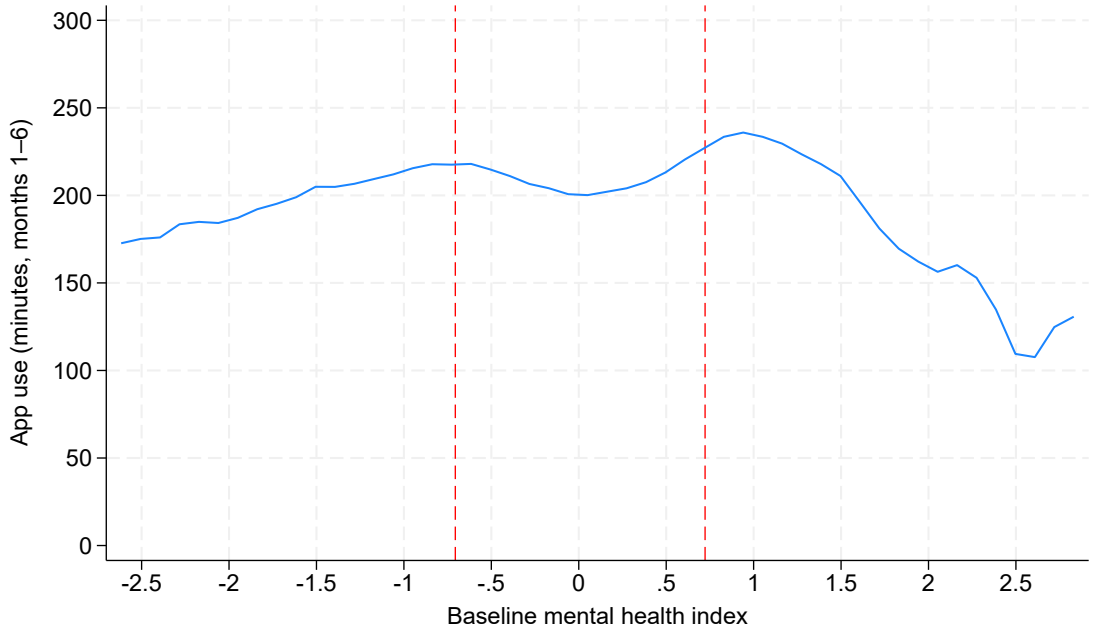
L App Use and Mental Health Impacts

Figure L.1: Predictors of Product Engagement



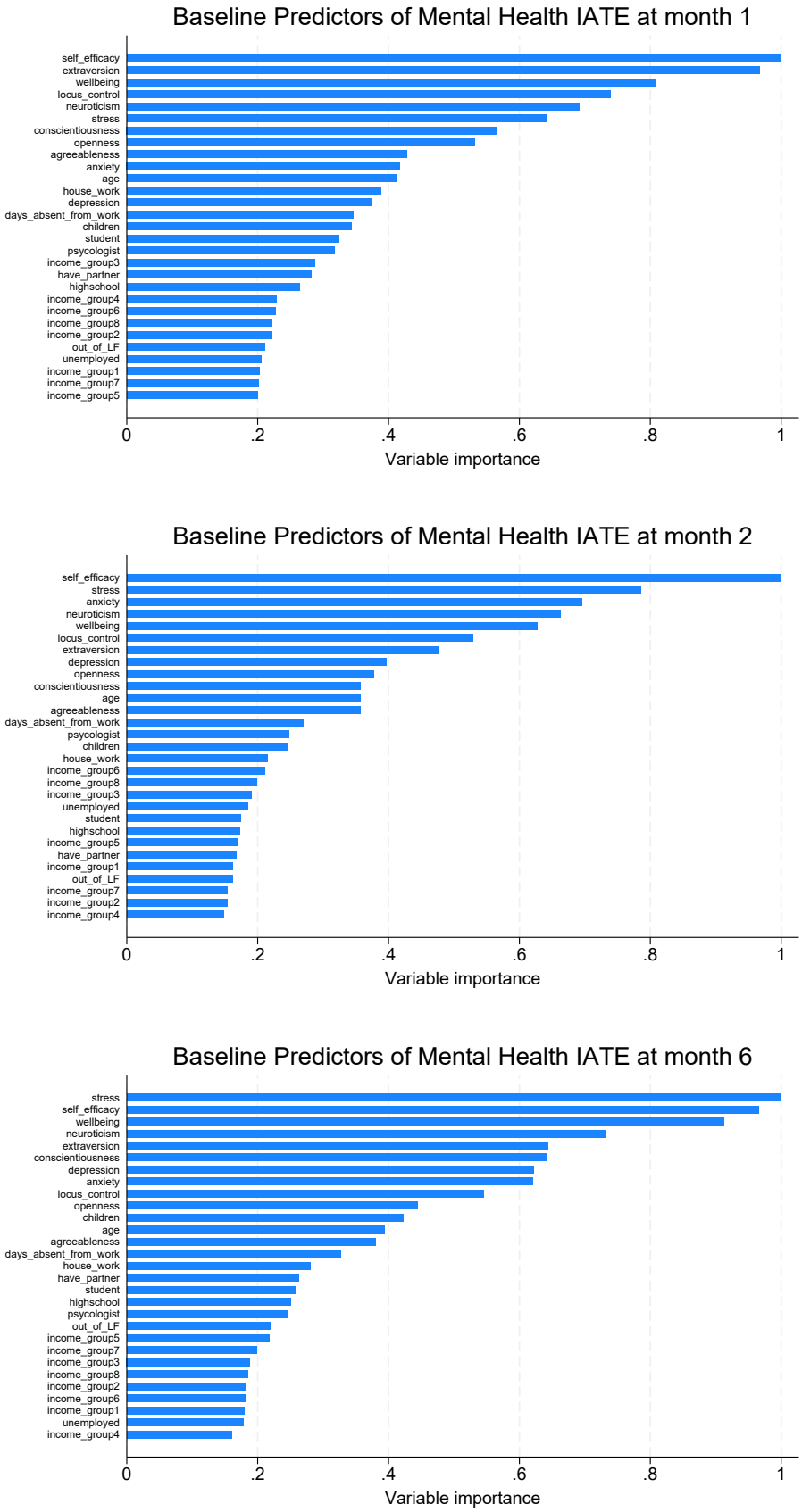
Notes: This figure ranks the baseline predictors of app use at 6 months (defined as the total number of minutes of app engagement over 6 months) in order of importance, using a random forest algorithm. Each predictors' importance is scaled relative to the most important predictor, whose importance is normalized to 1.

Figure L.2: Product Engagement by Baseline Mental Health



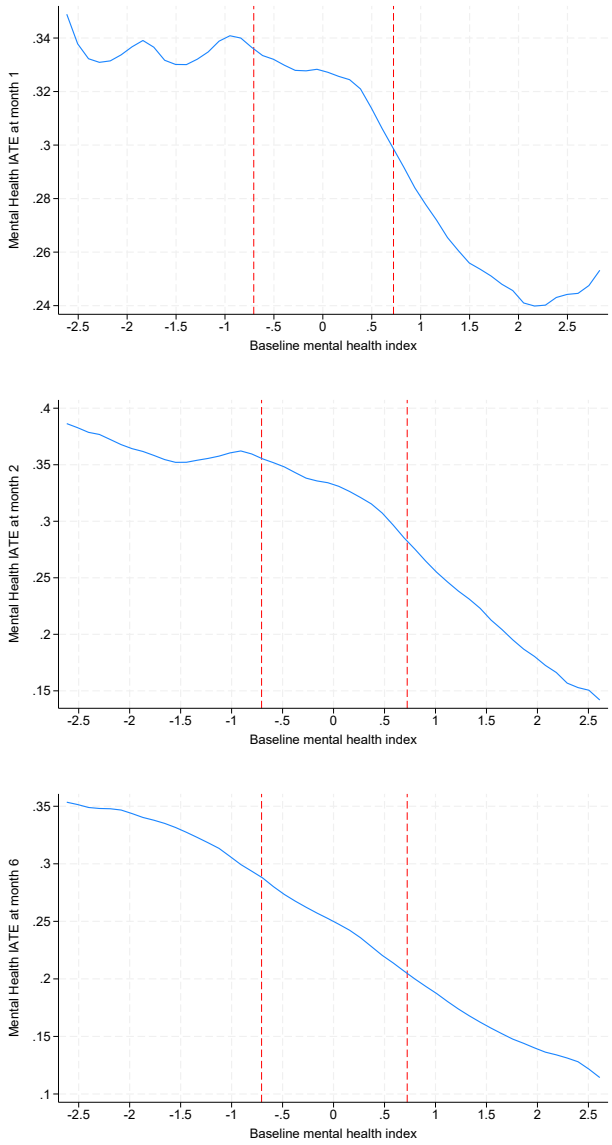
Notes: This figure shows the non-parametric relationship between app use (total minutes over 6 months) and baseline psychological distress (an index of PHQ-4, WHO-5, and PSS-4) in the treatment group using kernel-weighted local regression of degree 0 and an Epanechnikov kernel. The vertical bars mark the interquantile range of this index. A higher Index value means better mental health at baseline.

Figure L.3: Predictors of Mental Health Impacts at 1, 2, and 6 months.



Notes: This figure ranks the baseline predictors of mental health IATE estimates at 1, 2, and 6 months in order of importance, using a random forest algorithm. Each predictors' importance is scaled relative to the most important predictor, whose importance is normalized to 1.

Figure L.4: Mental Health Impacts at 1, 2, and 6 months by Baseline Psychological Distress.



Notes: This figure shows the non-parametric relationship between Mental Health IATEs estimated at 1, 2, and 6 months and baseline psychological distress (an index of PHQ-4, WHO-5, and PSS-4) in the treatment group using kernel-weighted local regression of degree 0 and an Epanechnikov kernel. The vertical bars mark the interquantile range of this index.

M Psychotherapy Use

Table M.1: Impacts on the Demand for Psychotherapy

	Visit Psychologist (1)	Num. Visits (2)
Panel A. Effects at 1 month		
Treated	0.033*	
	(0.018)	
Mean control	0.15	
Observations	1562	
Panel B. Effects at 2 months		
Treated	0.061***	0.155***
	(0.019)	(0.057)
Mean control	0.15	0.36
Observations	1562	1562
Panel C. Effects at 6 month		
Treated	0.079***	0.304***
	(0.019)	(0.075)
Mean control	0.14	0.49
Observations	1562	1562
1 mo=2 mo (*p-val*)	0.10	
2 mos=6 mos (*p-val*)	0.37	0.04
1 mo=6 mos (*p-val*)	0.03	
1 mo=2 mos=6 mos (*p-val*)	0.07	

*Notes:*psychotherapy utilization outcomes measured in the month prior to the survey. The first outcome, *Any psychotherapy use*, is an indicator equal to one if the respondent reports at least one visit with a psychologist or psychotherapist in the previous month. The second outcome, *Number of visits*, reports the total number of psychotherapy visits in the previous month. Regressions control for strata and lasso-selected baseline covariates. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table M.2: Impacts on First and Second Order Beliefs

	People with mental health conditions:				
	Are Lazy	Can't care for self	Have low willpower	Are dangerous	Seek attention/ are selfish
	(1)	(2)	(3)	(4)	(5)
First Order Beliefs					
Panel A. Effects at 2 months					
Treated	-0.003	-0.025	0.035	0.011	0.053
	(0.047)	(0.058)	(0.058)	(0.045)	(0.043)
Mean control	1.77	2.94	2.65	2.27	1.73
Observations	1523	1523	1523	1523	1523
Panel B. Effects at 6 months					
Treated	-0.025	-0.098*	-0.030	0.014	-0.010
	(0.045)	(0.057)	(0.058)	(0.046)	(0.040)
Mean control	1.69	2.90	2.50	2.22	1.67
Observations	1603	1603	1603	1603	1603
Second Order Beliefs					
Panel C. Effects at 2 months					
Treated	-0.060	-0.048	-0.097**	-0.073	-0.055
	(0.050)	(0.049)	(0.049)	(0.051)	(0.053)
Mean control	3.30	3.44	3.54	3.49	3.36
Observations	1521	1521	1521	1521	1521
Panel D. Effects at 6 months					
Treated	-0.188***	-0.080	-0.143***	-0.124**	-0.124**
	(0.048)	(0.050)	(0.048)	(0.050)	(0.054)
Mean control	3.46	3.52	3.62	3.65	3.44
Observations	1602	1603	1603	1602	1603

Notes: Outcome variables are measured on a five-point likert scale from 1 ("totally disagree") to 5 ("totally agree"). First-order beliefs capture whether the respondent personally agrees with each statement, while second-order beliefs reflect respondents' perceptions of whether people in their neighborhood agree or disagree. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table M.3: Impacts on Reported Channels and Perceptions of Psychotherapy

	Easy to find MH provider (1)	Know how to make appointment (2)	Willing to pay for therapy (3)	Can identify poor mental health (4)	I have contact information (5)	Embarrassed to see a MH provider (6)	No time for therapy (7)	MH providers are effective (8)
Treated	0.051	-0.030	-0.072	0.050	0.226***	0.001	0.056	0.075*
	(0.056)	(0.048)	(0.048)	(0.045)	(0.065)	(0.057)	(0.058)	(0.043)
Mean control	3.02	4.01	3.77	3.85	3.30	2.05	2.50	4.13
Observations	1603	1603	1603	1603	1603	1603	1603	1603

Notes: Each column reports treatment effects on beliefs or perceptions related to psychotherapy access and effectiveness. Outcome variables are measured on a five-point likert scale from 1 ("totally disagree") to 5 ("totally agree"). Effects measured at 6 months. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table M.4: Mediation of Treatment Effects by Psychotherapy Attendance

	(1) 1 month	(2) 2 months	(3) 6 months
Natural Indirect Effect (NIE)	0.004	0.016**	0.006
	(0.004)	(0.007)	(0.006)
Natural Direct Effect (NDE)	0.311***	0.309***	0.243***
	(0.037)	(0.041)	(0.044)
Total Effect (ITT)	0.316***	0.325***	0.248***
	(0.037)	(0.041)	(0.044)
Proportion mediated (NIE/ITT)	0.014	0.049**	0.022
	(0.011)	(0.021)	(0.025)
Observations	1,847	1,820	1,596
Treatment–mediator interaction	Yes	Yes	Yes

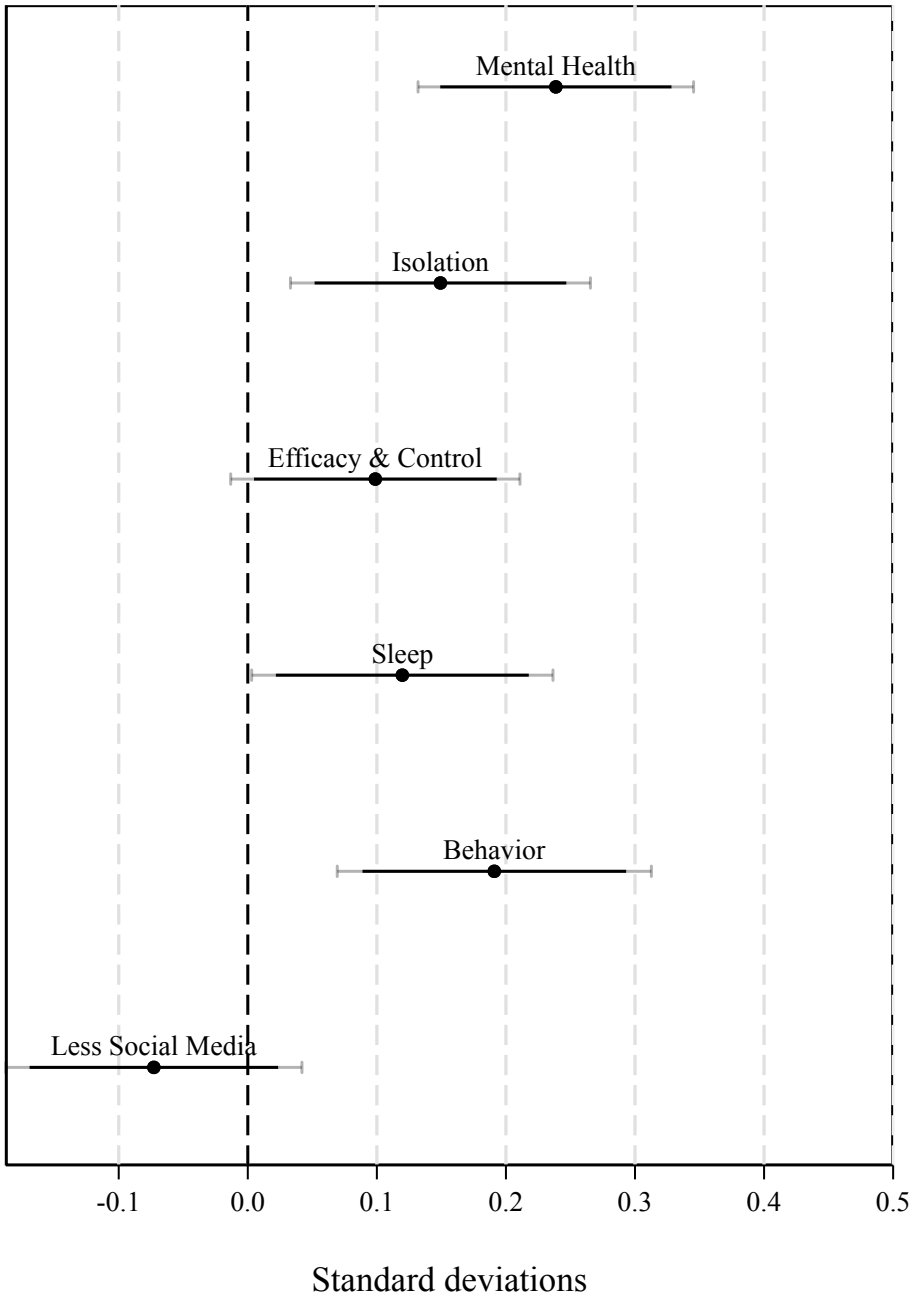
Notes: This table reports estimates from a mediation analysis. The outcome is the standardized mental health index at 1, 2, and 6 months. The mediator is an index of the demand for psychotherapy. Robust standard errors are reported in parentheses. The proportion mediated is defined as NIE/ITT, where ITT $=$ NIE $+$ NDE.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table M.5: Mental Health IATE by Psychotherapy Use

	(1) 1 month	(2) 2 months	(3) 6 months
Psychotherapy use	0.005	0.008	0.000
	(0.004)	(0.006)	(0.007)
Constant	0.313***	0.311***	0.246***
	(0.002)	(0.003)	(0.003)
Observations	922	906	801
R^2	0.002	0.002	0.000

Notes: Each column reports results from a linear regression of the estimated individual treatment effects (IATE) for the mental health index on an indicator for psychotherapy use in the previous month. The sample is restricted to treated individuals. Robust standard errors are reported in parentheses.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure M.1: Differences in outcomes (treatment - control) for people Who Did Not Attend Psychotherapy



Notes: This figure reports the differences in mental health outcomes betweer treatment and control groups restricting the sample to respondents who did not attend psychotherapy during the study period. Estimates are shown with 95% confidence intervals. Because psychotherapy attendance is a post-treatment outcome, this analysis conditions on an endogenous variable and should not be interpreted causally.

N Experimental Results from Extending App Access

Table N.1: Survey Completion Rates for App Extension Treatment

	6 months (1)	All 3 surveys (2)
App Extension	0.018	0.022
	(0.025)	(0.026)
Mean control	0.81	0.79
Observations	982	982

Notes: The table reports estimates from regressions of survey completion indicators on the app extension treatment assignment, restricting the sample to individuals assigned for the App extension randomization. Columns correspond to completion at 6 months and completion of all three follow-up surveys. Standard errors are clustered at the individual level. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table N.2: Balance Test for App Extension Randomization

	Control (1)	Treatment (2)	p-val (3)
Age	37.96	38.12	0.599
	(7.66)	(7.86)	
Income decile	4.15	4.01	0.292
	(2.15)	(2.02)	
Works	0.48	0.49	0.747
	(0.50)	(0.50)	
Number of children	1.03	0.93	0.119
	(0.99)	(0.95)	
Single	0.28	0.29	0.652
	(0.45)	(0.46)	
High School or more	0.64	0.64	1.000
	(0.48)	(0.48)	
Depression/Anxiety score (PHQ-4)	6.57	6.70	0.143
	(2.62)	(2.65)	
Well-being score (WHO-5)	7.65	7.84	0.404
	(4.02)	(4.12)	
Stress score (SSQ)	10.10	9.89	0.140
	(2.56)	(2.42)	
Self-efficacy score (GSE)	15.31	15.32	0.944
	(3.19)	(3.32)	
Locus of control score	11.77	11.75	0.889
	(2.75)	(2.86)	
Extroversion (Big5)	8.28	8.14	0.390
	(2.57)	(2.42)	
Agreeableness (Big5)	12.22	12.30	0.525
	(1.89)	(2.03)	
Conscientiousness (Big5)	10.40	10.37	0.879
	(2.70)	(2.80)	
Neuroticism (Big 5)	11.29	11.06	0.101
	(2.37)	(2.44)	
Openness to Experience (Big5)	10.73	10.78	0.727
	(2.07)	(1.97)	
Visit psych. (prior year)	0.32	0.31	0.841
	(0.47)	(0.46)	
Joint F-test (p-value)			0.466
Observations	491	491	

Note: The table reports baseline balance for the App extension randomization within treated group individuals that took place at 2 months of use. Column (3) reports p-values from regressions of each baseline covariate on the secondary treatment assignment, controlling for stratification fixed effects. Robust standard errors are used. The final row reports the p-value from an omnibus F-test of the null that baseline covariates do not jointly predict secondary treatment assignment.

O Expert Predictions

Table O.1: Predictions and Estimated Impacts

	Non-expert (1)	Expert (2)	Confident experts (3)	Study impacts (4)
Panel A. Continuous outcomes (medians)				
App use (% used at least once/week in wks 1–3)	55	61	71	.
App use (% used in week 8)	30	33	37	37
ITT: Mental health index (SD), 3 weeks	.12	.13	.18	.3
ITT: Mental health index (SD), 8 weeks	.11	.12	.12	.3
ITT: Healthful habits index (SD), 3 weeks	.1	.12	.13	.33
ITT: Without at least moderate depression (pp), 3 weeks	2.3	4.3	6.6	.
ITT: Slept $\geq$7 hours last night (pp), 3 weeks	2.4	3.2	6.9	.
Panel B. Correctly guessed sign of impact (%)				
Psychotherapy use	25	40	40	3
Social media use	41	44	40	.09
Social isolation	59	60	47	.22
N(predictions)	101	25	15	.

Notes: Columns 1–3 summarize expert predictions. For continuous outcomes, entries are medians from the survey of predictions for (i) non-experts, (ii) experts , and (iii) confident experts. For outcomes elicited at "3 weeks," the estimated impact reported is the study's 1-month effect; for "8 weeks," the estimated impact reported is the 2-month effect.
For categorical direction outcomes, columns 1–3 report the percentage of respondents whose predicted direction matches the study's estimated direction (increase for psychotherapy use; decrease for social media use and social isolation). Social media and isolation effects in the study are measured using standardized indices, where higher values correspond to *less* social media use and *lower* social isolation, respectively.