Discussion Paper Series

IZA DP No. 18538

April 2026

The Well-Being Effects of Digital Mental Health Care

Manuela Angelucci
University of Texas at Austin
and IZA@LISER

Raissa Fábregas
University of Texas at Austin

Antonia Vazquez
University of Texas at Austin

The Well-Being Effects of Digital Mental Health Care*

Abstract

AI-powered mental health apps have attracted growing interest as a low-cost way to expand care. Yet questions remain about their effectiveness, safety, and whether they may crowd out psychotherapy. We evaluate one such app in a randomized controlled trial among 1,964 Mexican women with mild to severe psychological distress. Over six months, app access improved mental health by 0.3 standard deviations with no evidence of harm, improved sleep quality, increased healthful behaviors, and reduced missed work, yielding considerably larger benefits than costs. Treated participants were also more likely to seek traditional psychotherapy, but this increase does not explain most of the mental health gains. App use was high in the first month but then declined, as is common in digital interventions. Despite this drop in use, treatment effects persisted. Participants continued to implement practices promoted by the app, suggesting that even short-term engagement can produce durable improvements through sustained behavioral change.

JEL classification

I12, O33, J24, C93, I15, I31

Keywords

digital mental health, AI-powered care, well-being, randomized controlled trial, Mexico, behavioral change, mental health apps, sleep quality, labor productivity, psychotherapy

Corresponding author

Manuela Angelucci
mangeluc@utexas.edu

* 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, Aalto University, Innovations and Regulatory Reforms in Healthcare IHS, Pacific Conference for Development Economics, and the University of Texas at Austin. We gratefully acknowledge financial support from The Agency Fund. The authors declare no financial relationships or other conflicts of interest. 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, and Vasco Andrade Baldini, Estefania Capriata, Monica Vargas, and Jiqing Zhao for excellent research assistance. We thank the Population Research Center at UT Austin for grant management support.

1 Introduction

Mental health disorders are a leading cause of disability worldwide, affecting nearly one billion people and generating an estimated welfare burden of 5 trillion USD (Arias et al., 2022). Mental health is both intrinsically valuable and an input into economic behavior, including labor-market outcomes (Grossman, 1972, Lund et al., 2024). Despite this, access to evidence-based care remains severely limited (World Health Organization, 2025): monetary and non-monetary costs constrain demand, while shortages of trained providers and underinvestment in the public system hinder its supply (Ridley et al., 2020).

Recent technological advances have created new ways to deliver mental health care at scale. Artificial Intelligence (AI) enables real-time, personalized care at low cost, and the widespread use of smartphones allows for private and on-demand delivery. This flexibility is valuable because people often experience fluctuations in symptom severity (Mohr et al., 2014, Kazantzis et al., 2010, Kleiman et al., 2017). Thus, app-based, AI-powered mental health care is an easily scalable, low-cost technology that can relax supply- and demand-side constraints to care. Indeed, one in ten people in the United States have mental health apps on their phones (Mikulic, 2025).

A central question is whether this technology is beneficial and, if so, how long its effects last. Uncertainty about effectiveness is heightened by rapid user disengagement from digital products, a common phenomenon often interpreted as evidence of limited effectiveness or value (Eysenbach, 2005, Smith et al., 2025). Thus, users may not engage long enough to realize benefits. At the same time, questions remain about safety and whether this technology crowds out existing evidence-based care, trading off more systematic and sustained benefits for short-term symptom alleviation (Torous et al., 2020a, Firth et al., 2022, Vaidyam et al., 2019).

To answer these questions, we conducted a large-scale experiment to estimate the mental health and behavioral impacts of an AI-powered mental health app. We study access to *Mindsurf*, a well-being app that combines an AI conversational agent trained on Cognitive Behavioral Therapy (CBT) techniques with features such as mood tracking and guided exercises. We recruited 1,964 digitally literate, high-need, high-constraint Mexican women interested in digital mental health support. We randomly assigned free access to the app for up to six months. We measured outcomes using complementary data sources: surveys at 1, 2, and 6 months; high-frequency affect data capturing weekly mood; and administrative data to study engagement. We also elicited

expert forecasts to benchmark our results against prevailing priors.

We report three main findings. First, app access improved mental health over six months by 0.29 standard deviations (SD). High-frequency affect data confirm this finding. This magnitude is comparable to psychotherapy and pharmacotherapy, with mental health benefits of 0.20–0.60 SD and 0.35 SD (Gartlehner et al., 2017, Singla et al., 2017, Cuijpers et al., 2023). The prevalence of severe symptoms also decreased by up to 27%, mitigating safety concerns. Experts under-predict these effects. Thus, the results are informative relative to prevailing priors. The intervention also improved sleep, healthful behaviors, and daily functioning, including a 0.10 SD improvement in an index of labor-market outcomes (absenteeism, employment, and hours worked). Back-of-the-envelope calculations suggest that app access can generate six-fold gains relative to costs through reduced absenteeism, and up to 400-fold gains when also accounting for forgone disability.

Second, disengagement need not imply ineffectiveness or low value, as benefits persist after use declines. Initial take-up is high but use declines: 82% of treated participants used the app in week one, but only 36% were active at two months and fewer than 10% at six months; total use averaged 242 minutes over six months of access. Yet the effects are highly persistent. A second experiment extended app access from 3 to 6 months and increased use by 39% but generated no additional benefits, indicating that continued use is not required for effectiveness. A plausible explanation is that people may adopt symptom-relieving skills and practices early on, reducing the need for continued use. Consistent with this interpretation, app access increased use of recommended practices; sleep and healthful behaviors mediated up to half of the mental health effects; and people with larger mental health gains also had larger sleep and behavioral impacts. Moreover, participants with the largest mental health gains did not have the highest app use, further suggesting that impacts need not track engagement.

Third, we study how access to digital mental health support interacts with traditional care. We model and estimate the effect of app access on psychotherapy use and find a 35% increase in the likelihood of attending psychotherapy in the previous month. In the model, this occurs either because the two technologies are complements in the production of mental health or because using the app decreases the costs of psychotherapy, for example by making symptoms more manageable. Consistent with complementarity or cost reduction, psychotherapy use increases among participants with higher baseline distress, for whom digital care likely makes psychotherapy more accessible or useful.

Taken together, the results show that this scalable technology can cost-effectively improve mental health and economic outcomes, without exacerbating severe symptoms or displacing skilled care, and that even brief use can generate sustained benefits.

This paper contributes to several strands of literature. First, we show that a low-marginal-cost, app-based intervention can generate large and sustained improvements in mental health in a non-clinical setting. Existing evidence on AI-based mental health products comes from small, short-duration efficacy trials in selected samples.[1] By conducting a large randomized evaluation with six-month follow-up in a policy-relevant population, we add a new technology to the set of evidence-based scalable delivery models, such as lay-provided psychotherapy, and speak to ongoing debates about whether meaningful mental health gains can be achieved at scale despite severe constraints on trained providers (see, e.g., Angelucci and Bennett (2026), Angelucci et al. (2026), Baranov et al. (2020), Bhat et al. (2022), Zadey (2023)).

Second, the paper contributes to the literature on digital technology adoption by showing that product use is an imperfect proxy for impact. Donors, international organizations, and policymakers often rely on engagement and retention metrics to evaluate performance of these tools (e.g., Gupta et al., 2004, Kumar et al., 2010, WHO, 2016, UNICEF, 2021, OECD, 2021). We show instead that effects persist even as use declines, consistent with behavioral change that does not require continued product use. When use and impact are decoupled in this way, engagement understates effectiveness, making downstream outcome measurement essential for evaluation. Our study shows that collecting such outcomes via text messages could be a promising, rapid, and low-cost approach for future evaluations. This decoupling also has implications for scale: if gains are largely generated in the first weeks of use while monetization depends on sustained engagement, private providers may capture only a fraction of the social value under subscription-based pricing. In that case, market provision may fall short of the social optimum, and alternative financing or delivery models may be needed to achieve efficient scale.

Third, the paper contributes to the broader literature on the economic consequences of mental health by providing experimental evidence on work-related outcomes. These results are consistent with frameworks in which mental health affects labor supply and productivity (Haushofer and Fehr, 2014, Ridley et al., 2020). They also help interpret mixed evidence across contexts:

[1] The median sample size in a meta-analysis of these products is 148, and the median study duration is four weeks (Li et al., 2023).

meta-analytic work finds positive average labor-market effects (Lund et al., 2024), but there is much unexplained heterogeneity (Nieuwenhuijsen et al., 2020). For example, several experimental studies from South Asia, with predominantly or all-female samples, report limited or null labor market effects (Angelucci and Bennett, 2024, Baranov et al., 2020, Bhat et al., 2022); our findings suggest that gender alone is unlikely to account for these null results.

Lastly, the paper contributes to the study of health care delivery when new care technologies may complement or substitute pre-existing ones (Cutler and McClellan, 2001, Chandra and Skinner, 2012). In particular, when technologies are complements in health production, or one lowers the other's non-monetary costs, the new technology may cause crowding in rather than displacement. In mental health care, this logic underlies models of collaborative and stepped care, in which low-intensity interventions facilitate initiation and escalation into specialist services rather than replace them (Unützer et al., 2002, Patel et al., 2010). Consistent with this view, we provide evidence of complementarity in our setting: app access increases psychotherapy use, particularly among higher-distress participants.

2 Conceptual Framework

This paper studies a new mental health technology, A. This technology provides symptom relief through practices that promote healthful behaviors and coping strategies, including better sleep hygiene, emotional regulation, and other routines and tools to manage distress. These behavioral changes may persist even if active use of the product later declines.

People value mental health (H) directly. Thus, A can directly increase utility. Moreover, because mental health affects day-to-day functioning, improvements in H may also translate into gains in labor-market outcomes. In this sense, the benefits of A may extend beyond mental health.

Technology A functions alongside existing in-person psychotherapy, T, and delivers mental health support at lower cost than traditional care. It can also be used privately, flexibly, and on demand; and it requires less time, planning, and coordination than a psychotherapy visit. These features lower both monetary and non-monetary barriers to care. Thus, its introduction may change the use of traditional psychotherapy in ways that are not obvious *ex ante*. On one hand, A may substitute for T. If A improves health, some people may need less psychotherapy than they otherwise would have. This substitution depends on whether and how effectively A

performs a therapeutic role comparable to in-person care. On the other hand, A may complement T. By improving sleep, concentration, emotional stability, or daily functioning, it may increase the productivity of psychotherapy. It may also lower the non-monetary costs of therapy by reducing stigma, improving knowledge about treatment, or making help-seeking more manageable. In that case, access to A can crowd in psychotherapy.

This framework yields four testable hypotheses that guide the analysis. First, if A is effective, access should improve average mental health, though the magnitude of these gains may differ across individuals. Second, some of these gains may operate through changes in healthful behaviors and coping practices. Third, improved mental health may lead to better labor-market outcomes. Fourth, the effect of A on psychotherapy use is theoretically ambiguous: it may be negative if digital care mainly substitutes for therapy, null if marginal users would not have sought psychotherapy otherwise, or positive if it lowers the cost of seeking therapy or complements it. The empirical analysis that follows tests these hypotheses.

3 Product, Experimental Design, and Data

3.1 The Mental Health App

We offered free access to a premium version of *Mindsurf*, a commercially available CBT-based mental health care app. CBT improves mental health through cognitive restructuring (reframing maladaptive thoughts) and behavioral activation (encouraging engagement in meaningful, mood-improving activities). The app has four core features: a mood diary, an AI-powered conversational agent, audio and guided exercises, and self-assessment tools. The AI agent is both a stand-alone feature for unstructured conversations and a content manager that directs users to relevant tools and exercises. The app can also connect users to external mental health resources, including psychotherapy, although take up of that component was minimal (four users in the sample). A free version of the app includes only the mood diary.

To promote safety, the company restricts the AI agent to mental health–related interactions, periodically stress-tests it using risk-related prompts, has trained mental health professionals to review a subset of de-identified conversations, and uses a separate AI-based system to detect text suggestive of acute distress for review and follow-up when appropriate. Appendix A describes the app, its training, and safeguards.

3.2 Experimental Design

The experiment ran from April to December 2025 (Figure 1). We recruited participants on a rolling basis in April–May through Facebook and Instagram advertisements framed around emotional well-being and designed to reach people experiencing psychological distress.[2] Interested individuals completed a brief screening survey. Eligibility required being a woman aged 18–50 living in Mexico, having no graduate education and a monthly household income below the 7th decile in national data (∼1,435 USD), and scoring at least 3 on the PHQ-4 depression/anxiety screener.[3] We focus on women in a middle-income country, a population that bears a large share of the global burden of common mental disorders and experiences high rates of depression and anxiety (WHO, 2023, GBD, 2022). The distress, income, and education criteria help us identify a high-need, high-barrier-to-care population. Of 6,690 screened respondents, 2,923 were eligible (44%); 2,184 completed a baseline survey two days later (75%), and 1,964 signed a participation pledge intended to reduce long-run attrition (90%).

We randomly assigned half of the 1,964 participants to the treatment group and the other half to the control group. Treated participants received access to the app for either three or six months via a second randomization done at three months; because this second randomization had no detectable effects on our outcomes of interest, we pool treatment arms in the main analysis. Controls received access after six months. All participants received US$15 for participation.[4]

3.3 Data

We draw on four data sources: (i) participant surveys at baseline and 1, 2, and 6 months post-randomization; (ii) weekly affect measures; (iii) administrative app use data; and (iv) expert predictions collected via the Social Science Prediction Platform (DellaVigna et al., 2019).

[2] In total, 825,727 users viewed the ads and 31,259 clicked to begin screening. Figure B.1 shows sample ads. Facebook is widely used in Mexico, with estimates suggesting reach of approximately 89% among individuals aged 13 and above (Kemp, 2025).

[3] Scores of 3–5 indicate mild distress, and 6–8 or 9–12 indicate moderate or severe distress. Scores ≥ 3 on either subscale indicate likely anxiety or depression and suggest the need for further assessment.

[4] We filed a pre-analysis plan before the first round of data collection, specifying outcome families and specifications, and later amended it (before the final survey wave) to pre-register outcomes for the six-month follow-up.

3.3.1 *Survey Data*

At baseline, we measured mental health (PHQ-4, WHO-5, PSS-4), socioeconomic characteristics, personality traits, healthful activities, locus of control, self-efficacy, social media use, and psychotherapy use in the prior 12 months. We discuss outcome data next.

Mental health. Mental health is our primary pre-registered family. We measured depression with the PHQ-8 (0–24), anxiety with the GAD-7 (0–21), subjective well-being with the WHO-5 (0–100; scores below 50 indicate poor well-being), and stress with the PSS-4 (0–16; higher scores indicate greater stress). Scores ≥ 10 on the PHQ-8 or GAD-7 indicate at least moderate symptoms and are strongly predictive of depressive or anxiety disorders. First-line treatments are CBT-based psychotherapy and pharmacotherapy (SSRIs), with symptom improvements often observed as early as 2–4 weeks (Kroenke et al., 2009, Löwe et al., 2008).

Healthful behaviors. We measured three indicators of daily *functioning* that are likely improved by better mental health: job absenteeism, helping others, and emotional regulation. We also measured three *wellness* practices known to benefit and be shaped by mental health: exercise, leisure outings, and self-care (Mahindru et al., 2023, Fancourt et al., 2021). Respondents reported how many days in the prior week they missed work, helped someone with homework, raised their voice in anger, exercised, went out for non-work/school reasons, and spent at least 10 minutes on themselves. We combined them into an index, along with functioning and wellness sub-indices. At 2 and 6 months, we also measured the frequency of app-recommended practices (e.g., self-kindness, boundary setting, breathing exercises, sleep routines, journaling) and constructed a corresponding index (higher values indicate more frequent practice).

Sleep quality. Poor sleep is both a cause and a consequence of psychological distress (Harvey, 2011, Baglioni et al., 2016). We measured sleep using the PHQ-8 sleep item, which asks whether respondents have had trouble falling asleep, staying asleep, or sleeping too much over the past two weeks, as well as self-reported bed and wake times (to compute sleep duration), and the number of nighttime awakenings, and combined these into a sleep quality index, with higher values indicating better sleep.

Social media use. Social media can be addictive and can reduce subjective well-being and mental health (Allcott et al., 2020, 2022, Braghieri et al., 2022, Mosquera et al., 2020). In our setting, app access could either reduce social media use (via improved well-being or time substitution) or increase it (via greater overall phone engagement). We measured days of use in the prior week

(Facebook, X, Instagram, TikTok) and average daily time spent on social media, from which we constructed an indicator for high use (more than two hours per day). We combined them into an index, with higher values indicating lower social media use.

Social isolation. Social isolation is both correlated with and predictive of worse mental health (e.g., Holt-Lunstad, 2022). Because digital care lacks in-person contact, it may increase disconnection (DHHS, 2023). We measured loneliness using the six-item De Jong Gierveld scale (De Jong-Gierveld and van Tilburg, 2006), reporting emotional and social subscales and an overall index, where higher values indicate less isolation.

Self-efficacy and locus of control. Self-efficacy and locus of control are linked to resilience and well-being (Botha and Dahmann, 2024). We measured self-efficacy using the General Self-Efficacy (GSE) scale and locus of control using the Internal–External Locus of Control Short Scale–4 (IE-4). We combined them into an *agency* index, with higher values indicating greater perceived agency.

Cognitive tasks. We implemented two incentivized tasks measuring cognitive function and effort. First, participants counted zeros in binary matrices under time pressure (Abeler et al., 2011). Second, we implemented an emotional Stroop task assessing cognitive control in the presence of emotionally salient stimuli (Williams et al., 1996). Participants identified word color in blocks of emotional versus neutral words; longer reaction times for emotional words are commonly interpreted as difficulty disengaging from salient content. Performance in both conditions can capture broader aspects of cognitive function and effort. We combined the tasks into an index, with higher values indicating better performance. Details in Appendix C.

Psychotherapy use. We measured recent psychotherapy use by asking whether respondents attended any sessions with a psychologist, psychiatrist, or therapist in the prior 30 days.

Non-monetary costs of therapy. At 2 and 6 months, we measured first- and second-order beliefs about mental health by collecting own perceptions about mental health and beliefs about neighbors' perceptions. At 6 months, we also measured other non-monetary access costs of psychotherapy: whether the respondent knows how to find a psychologist, knows how to make an appointment, and has contact information for a mental health professional. We created indices of *stigma* and *access costs*, with higher values indicating less negative beliefs.

3.3.2 *High-Frequency Affect Data*

A distinctive feature of the study is the high-frequency mood measurement. Beginning one week after app access, we sent weekly WhatsApp messages asking participants about their current mood; respondents selected a happy, neutral, or sad emoji. We use these responses as a high-frequency proxy for affect and construct a "happy affect" indicator. Appendix D describes these data and reports strong correlations with survey-based mental health measures (Table D.1).

3.3.3 *Administrative Data*

We also use administrative data on app engagement, including use frequency and duration, sessions completed, and feature-specific engagement. To protect privacy, we do not observe the content of the interactions between the participant and the AI-agent.

3.3.4 *Expert Predictions*

To gauge expert priors and how our findings may change the views of the scientific community, we elicited predictions from economists and mental health experts via the Social Science Prediction Platform (DellaVigna et al., 2019). The survey had 126 respondents.

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

We estimate average Intent to Treat (ITT) effects from the following equation:

$$Y_i = \alpha + \beta \mathrm{T}_i + \gamma X_i + \delta Y_{i0} + \mu_s + \varepsilon_i, \tag{1}$$

where Y_i is the outcome for person i, T_i is an indicator for assignment to immediate app access, X_i is a vector of baseline covariates, Y_{i0} is the dependent variable at baseline (when available), μ_s are strata fixed effects, and ε_i is the error term. We select covariates via LASSO following Chernozhukov et al. (2018).[5] The coefficient β identifies the ITT effect under random assignment and SUTVA. To address multiple testing, we construct outcome-family indices standardized relative to the control-group distribution (Anderson, 2008). We estimate effects separately at

[5]The full set of baseline covariates to be selected by LASSO consists of: PHQ-4, Stress, and well-being; self-efficacy and locus of control; visiting a psychotherapist in the previous year; 'Big 5' personality traits (extraversion, openness, neuroticism, agreeableness, and conscientiousness); indicators for being employed, having below sample median income and at least a high-school diploma; being single; age and number of children.

1, 2, and 6 months post-randomization using heteroskedasticity-robust standard errors. When pooling survey rounds, we cluster standard errors at the individual level.

5 Sample and Descriptive Statistics

5.1 Sample

Of the 6,690 individuals who completed the first screening survey, 2,923 met the eligibility criteria, and 1,964 completed all baseline surveys and were randomized. Table E.1 shows that sample characteristics are similar across recruitment stages, besides differences in dimensions used for screening. The final sample comprises adult women with at least mild psychological distress recruited across Mexico (Figure E.1). Participants are, on average, 38 years old, have 1 child, roughly half of them are employed, and 31% report a psychotherapy visit in the prior year. Overall, the sample captures women experiencing distress with some familiarity with formal mental health care.

For context, about 50% and 19.5% of Mexican adult women experienced anxiety or depression in the previous year, with higher rates among low-income women (ENBIARE, 2021). Despite this burden, access to treatment remains limited. Public mental health services are scarce and concentrated in urban areas, and the supply of trained professionals is insufficient: Mexico has one psychiatrist per 100,000 people, far below the 16 per 100,000 in the United States (Patel et al., 2025). At the same time, internet and cellphone penetration are high, with 81% of the population having access to both (INEGI, 2025). Thus, Mexico has a large user base of digital products, making digital mental health care easily scalable.

5.2 Balance and Attrition

Table F.1 shows no evidence of differential attrition by treatment status for the survey data. Survey completion was 95%, 94%, and 82% at 1, 2, and 6 months. Table F.2 shows that baseline characteristics are balanced: we cannot reject the null that coefficients are jointly equal by arm.

The weekly affect data response rate was 78%, decreasing from 87% in week 1 to 75% in week 26. Table F.3 pools an indicator for attrition over 26 weeks and regresses it on baseline covariates and their interaction with the treatment indicator, clustering standard errors by person. Baseline mental health does not predict attrition, either overall or by arm. However, the covariates jointly

predict attrition for both the treatment and control groups ($p < 0.05$), and unconditional attrition is 9 percentage points higher in the treatment group. To account for differential attrition by arm in the affect data, we estimate Lee bounds (Lee, 2009).[6]

6 The Effects of App Access

6.1 App Take-Up

Eighty-seven percent of the treatment group downloaded the app, and 95% of them used it during the study period. Total use averaged 242 after 6 months of access. Figure 2 shows that 82, 56, 36, and 10% of the treatment group used the app in weeks 1, 4, 8, and 26. The drop is more marked for the group whose access to the full suite of app features ended at three months (afterwards, they had access to the mood diary only). Declining use is not unique to digital delivery: in-person psychotherapy also exhibits premature dropout, with rates as high as 50% (McGovern et al., 2024).

6.2 Effects At a Glance

Figure 3 reports ITT estimates for standardized indices by outcome family, with positive coefficients denoting improvements. The mental health index (our primary outcome) increases by about 0.30 SD at 1 and 2 months and 0.25 SD at 6 months. We also find statistically significant improvements across other outcome families, with effect sizes typically between 0.10 and 0.30 SD and some point estimates attenuating over time. Results are qualitatively unchanged when restricting the sample to respondents who completed all three surveys (Figure G.1).

6.3 Effects on Mental Health

Effects on Depression, Anxiety, Subjective Well-Being, and Stress. Table 1 shows that app access significantly improves psychological well-being for up to six months, with a pooled effect of 0.29 SD on the mental health index and no evidence of effect change over time. We find statistically significant improvements in each outcome in all follow-ups, and effects are broadly stable over time: depression and anxiety impacts, the clinically salient index components, are largely unchanged through month six. By contrast, impacts on well-being and stress attenuate by month six, although this decline is statistically significant only for stress.

[6] Using also inverse propensity weights did not change the estimates. Thus, we do not report those estimates.

To benchmark impacts against other evidence-based treatments for anxiety and depression, Figure G.2 shows the standardized effect sizes. The severity of symptoms of depression declined by 0.31 at 1 month to 0.27 SD at 6 months, of anxiety by 0.20 SD throughout the 6 months, and of stress by 0.30 to 0.15 SD. Well-being increased by 0.32 to 0.24 SD. These magnitudes are comparable to effects reported in the literature on digital self-help tools. Efficacy studies of AI-based interventions report effects of 0.3 SD for well-being at one month (Li et al., 2023). However, this evidence is based on highly selected samples (often college students), small sample sizes, and short follow-up horizons, which limit external validity and inferences about persistence. For instance, the studies reviewed in Li et al. (2023) have a median sample of 148, a median duration of 1 month, and primarily recruit college students from the USA, the UK, or Japan. *Therabot*, a generative AI agent, reduces depression by 1.5 points on the PHQ-9 (Heinz et al., 2024), similar to our 1.5–1.6 point reductions for the PHQ-8.

These effect sizes are also comparable to the impacts of psychotherapy and pharmacotherapy. Meta-analytic evidence (primarily from high-income settings) shows that CBT reduces depressive symptoms by about 0.22 SD (Gartlehner et al., 2017), with more recent estimates ranging from 0.47 SD (after publication-bias adjustments) to 0.60 SD (in low-bias studies) (Cuijpers et al., 2023). The magnitudes we document are thus notable, given the intervention's light-touch, short-duration, and non-clinical delivery. Standard in-person CBT typically lasts 1.5–5 months and is substantially more time- and resource-intensive than app-based care. A closer comparison is with scalable alternatives such as brief, lay-delivered, adiagnostic group therapy: Angelucci and Bennett (2026) and Angelucci et al. (2026) report improvements in a mental health index of 0.2–0.4 SD among adult women in Mexico and India after 5 sessions, similar to the effects we estimate here.

Distributional Impacts. A concern about using large language models (LLMs) for psychological support is that AI-generated responses could provide inappropriate guidance or exacerbate distress for some users. Although we do not observe the content of AI-agent interactions, we can assess risk indirectly by comparing outcome distributions across treatment and control groups. Figure 4 shows that app access shifts the entire distribution of our mental health outcomes in the pooled data. App access substantially reduces clinically relevant symptoms: it lowers the prevalence of at least moderate depression by 12 pp (50% vs. 62%) and at least moderate anxiety by 9 pp (34.0% vs. 43.0%). We observe similar patterns for perceived stress and well-being. Moreover, treated participants are no more likely than control participants to experience severe

depression, severe anxiety, high stress, or very low subjective well-being. In fact, extreme distress is more prevalent in the control group.

To measure impacts on severe depression and anxiety, we estimate ITT effects on indicators for these outcomes. Table G.1 shows that the likelihood of having severe symptoms decreases by 2.7 pp ($p<0.01$) for depression (a 27% reduction) and by 2.3 pp ($p<0.05$) for anxiety (a 14% reduction), with effects that do not systematically vary over time. Thus, we find no evidence that app access worsens mental health outcomes.

High-Frequency Affect Data. High-frequency outcomes mirror the survey pattern: Figure 5 shows that treatment effects on reporting a happy emoji are largest in the first two months and then attenuate, while remaining positive and statistically significant throughout the study period. The weekly data also indicate a rapid onset of benefits: the probability of reporting "happy" is 10 pp higher in the treatment group in week 1 and peaks around 20 pp in weeks 4–6. These short-run gains are consistent with evidence that digital mental health interventions can generate improvements over very short horizons (Li et al., 2023).

6.4 Effects on Other Outcomes

Improved mental health can have downstream positive impacts on several families of outcomes: people may have more energy and productivity, take better care of themselves, and feel more in control of their lives. At the same time, the app may influence the adoption of behaviors such as sleep routines, exercise, social engagement, and social media use. Some of these actions have a circular effect: for example, sleeping and exercising both contribute and respond to mental health (Mahindru et al., 2023, Scott et al., 2021).

Effects on Healthful Behaviors. Table 2 shows improvements in all measured healthful behaviors. The pooled index increases by 0.23 SD ($p<0.01$), with impacts that decline from 0.32 SD at 1 month to 0.19 SD at 6 months. The functioning and wellness sub-indices show similar dynamics: effects peak at 1 month (0.21 and 0.30 SD) and decline by 40–50% at two and six months ($p<0.01$ throughout). All components are statistically significant in pooled estimates and move in the expected direction. For example, exercise increases by 0.23 days per week, suggesting increased physical activity as a potential mechanism.

We also find a 0.08 day reduction in missed work days per week ($p<0.05$), a 16% decline relative to the control mean. This estimate is consistent with evidence that psychotherapy can reduce absenteeism among severely depressed patients (Patel et al., 2017). Such effects are plausible, given that core symptoms captured by the GAD-7 and PHQ-8 include sleep disruption, low energy, impaired concentration, and restlessness or uncontrollable worry, which can impair functioning. Thus, improved functioning can translate into lower absenteeism and potentially higher labor supply and productivity.

To further investigate labor-market outcomes, Table 3 shows estimates of impacts on working for pay and hours worked in the previous week (not prespecified) at two and six months. Besides its impact on absenteeism, app access also increased the probability of working by 3.1 pp in pooled estimates ($p < 0.05$), from a control mean of 50%. The pooled effect on hours worked is positive but statistically insignificant. A labor market index combining these two outcomes with absenteeism shows pooled impacts of 0.10 SD ($p < 0.01$).

These effects are notable in light of a recent meta-analysis finding that psychotherapy for common mental disorders in low- and middle-income countries (LMICs) improves a composite labor market outcome by 0.16 SD on average (Lund et al., 2024). However, individual studies have found mixed results, with some finding persistent mental health improvements but no labor market impacts, particularly among female populations (Angelucci and Bennett, 2024, Baranov et al., 2020, Bhat et al., 2022). Thus, the 0.10 SD index effect from this intervention is encouraging, showing that this technology can translate into better labor-market outcomes for women.

Effects on Sleep. Table 4 shows sustained improvements of 0.15–0.22 SD in the sleep index. In pooled estimates, all components are statistically significant ($p<0.05$): treated participants sleep about 10 additional minutes per night, report 0.11 fewer nighttime awakenings, and 0.11 fewer sleep issues. While some effects lose statistical significance at six months, we cannot reject the null of constant impacts over time. These findings align with evidence that sleep has a causal effect on mental health (Scott et al., 2021), suggesting improved sleep as a plausible pathway. For comparison, Bessone et al. (2021) explicitly target sleep and find that providing sleep aids and encouragement to low-income adults in India increased objectively measured sleep duration by

27 minutes per night. We find gains of 12 minutes at one month, roughly half as large.[7]

Effects on Social Media Use. Table 5 shows modest pooled impact on the social media index, which increases by 0.06 SD reflecting a small reduction in use of Facebook, X, as well as a decrease in high-frequency use of social media. Most effects fade out at six months, although we generally cannot reject the hypothesis that the effects are identical at 1 and 6 months.

Effects on Perceived Isolation and Agency. Table 6 shows that app access provides sustained reductions in both measures of isolation and agency, with pooled impacts of about 0.2 SD for both indices and no clear evidence of fading out. The impacts on agency are consistent with a well-established negative correlation between depression and anxiety and self-efficacy and control (e.g., Derese et al., 2024, Tahmassian and Moghadam, 2011).[8]

Effects on Cognitive Tasks. Table 7 shows no statistically significant pooled effects in either task, but faster reaction times in the Stroop task at 1 month. This temporary increase in reaction time could be related to both effort and cognitive function.

6.5 Local Average Treatment Effects (LATEs)

Table G.2 provides estimates of the LATEs for all indices, using treatment assignment as an instrument for having downloaded the app. Since approximately 87% of the treatment group downloaded the app, the LATE and ITT estimates are of similar magnitude.

6.6 Experimenter Demand Effects

Our results are unlikely to be driven by experimenter demand effects. We provide four pieces of evidence (Appendix H provides the details). First, effects extend to outcomes that are relatively difficult to manipulate, including sleep duration constructed from reported bed and wake times and performance on the emotional Stroop task. Second, following Dhar et al. (2022), we show treatment effects are not larger among participants with greater baseline susceptibility to social desirability bias (Table H.1). Third, app access does not affect respondents' perceived value of

[7]Our sleep duration measure is self-reported, constructed from reported bed and wake times. Bessone et al. (2021) show that self-reported time in bed closely tracks actigraphy, whereas self-reported sleep duration overstates objectively measured sleep. Any level bias is unlikely to differ systematically by arm, and treatment effects in that study operated primarily through increased time in bed, a margin captured by our measure.

[8]However, the causal effect between these two sets of outcomes is likely bi-directional, with, e.g., depression being both a cause and a consequence of low efficacy and lack of control.

participating in the study (Table H.2). Fourth, in an additional experiment following De Quidt et al. (2018), informing a randomized subset of respondents about the study's main hypothesis does not amplify treatment effects (Table H.3).

6.7 Expert Predictions

Table I.1 shows that, while expert predictions are systematically more accurate than non-experts', and confident experts' predictions are the most accurate, all forecasts understate the magnitude of the estimated impacts. Experts (confident experts) predict mental health improvements of 0.13 (0.18) SD at 3 weeks and 0.12 (0.12) SD at 8 weeks, whereas the 1- and 2-month effects are about 0.30 SD. These underestimates occur despite generally not under-predicting app use in the first three weeks (61% and 71% predicted vs 55% observed) and at 8 weeks (33% and 37% vs 36%). Thus, experts likely believe the app to be less effective than observed. To conclude, our results are likely surprising relative to priors in the research community. Details in Appendix I.

7 Cost-Effectiveness and Cost-Benefit Analyses

We provide back-of-the-envelope estimates of cost-effectiveness and cost-benefit analysis. The marginal cost of this product is low: most expenditures for an AI-based platform are fixed (e.g., development and infrastructure) and can be spread across a larger user base, while variable costs (e.g., API usage and server capacity) are comparatively low. *Mindsurf*'s business model operates largely through institutional subscriptions, with a per-user cost of roughly 1 USD per month at scale. The marginal cost is likely lower.

To benchmark, Table L.1 shows costs per 0.1 SD improvement in mental health across selected recently implemented scalable interventions that report their costs. Per-person costs vary from 6 to 1,189 USD, with *Mindsurf* having the lowest cost. This is relevant if users are charged for this product, especially for people with liquidity constraints. The monthly cost of improving mental health by 0.1 SD varies from 0.03 to 10 USD. The app is relatively cost-effective, at 0.37 USD. Costs would be even lower if a government developed and deployed its own digital product (as, e.g., the U.S. government currently does for veterans with PTSD (Kuhn et al., 2017)).

In addition, we perform two cost-benefit analyses. First, we benchmark against labor-market benefits using our estimated impacts on work absenteeism. The pooled estimated ITT effect is

-0.078 missed-work episodes per week; over 26 weeks, this implies $0.078 \times 26 = 2.03$ fewer missed-work days during access. Valuing time at Mexico's minimum wage yields wage gains of $2.03 \times 17.23 = 35$ USD over six months. Relative to an estimated six-month program cost of 6 USD, this implies a benefit-cost ratio of $35/6 = 5.8$ from absenteeism alone, before accounting for other welfare gains or any persistence beyond the intervention period.

Second, we benchmark benefits using Disability-Adjusted Life Years (DALY) considering the reduced prevalence of at least moderate depression. The pooled estimated ITT effect is a 12 pp reduction over six months (from Figure 4). This implies $0.12 \times 0.5 = 0.06$ moderate-depression years averted per participant offered access, a likely underestimate given that there continue to be impacts at 6 months. Applying the Global Burden of Disease 2013 disability weight for moderate major depressive disorder, 0.396, yields $0.06 \times 0.396 = 0.0238$ DALYs averted (Salomon et al., 2015). Valuing a DALY at 100,000 USD implies benefits of $0.0238 \times 100{,}000 = 2{,}376$ USD per participant over six months (Favaloro and Berger, 2021). Relative to the six-month program cost of 6 USD, this implies a benefit-cost ratio of $2{,}376/6 = 396$. The combined benefit-cost ratio of absenteeism and DALY is $(2{,}376 + 35)/6 = 402$.[9]

8 Patterns of Product Engagement

Access to administrative data allows us to measure use without relying on self-reports and to describe patterns in detail. We use these data to characterize specific features of engagement.

Figure 6 reports daily minutes among active users. Conditional on positive use, intensity declines from 12-17 daily minutes in week 1 to about 7 minutes by 2 months and then remains relatively stable through month 6. Thus, the remaining active users continue to devote non-trivial time to the app. Composition also shifts over time. Self-assessment occurs primarily in the first 10 days, whereas AI conversations and guided exercises account for most use, though they gradually decline over the first three months and then increase slightly in the second three months. By contrast, mood-diary use is comparatively stable at around 1 minute per day throughout the study period. Overall, these patterns suggest a shift away from assessment and skill acquisition.

[9] A public health provider is interested in the above analysis. However, a private individual may want to include the cost of additional therapy. Psychotherapy use increases by 5.2 pp per month over six months (from Table 9). Assuming a monthly therapy cost of 100 USD, this implies an additional expected cost of $0.052 \times 6 \times 100 = 31.2$ USD per participant. Total cost thus increases to $6 + 31.2 = 37.2$ USD. The DALY-based benefit-cost ratio then becomes $2{,}376/37.2 = 64$, and the combined absenteeism-plus-DALY ratio becomes $(2{,}376 + 35)/37.2 = 65$.

Figure 7 describes additional dimensions of use. About 56% of active days begin between 9:00 pm and 8:00 am on weekdays, 37% of sessions start between 9:00 pm and 6:00 am, and 27% occur on weekends. Thus, a substantial share of use occurs outside of times when in-person psychotherapy would typically be available. Panel C summarizes within-user regularity using the circular standard deviation of session start times within a day and the coefficient of variation (CV) of days between use. Regular use equals no variation in both cases. In contrast to traditional psychotherapy, where appointment times are often fixed, app use is highly irregular: 58% of users exhibit high variability in days of use (CV>1) and 84% exhibit high variability in session start times (SD>3 hours). Thus, most app use occurs at varying times of day and on different days of the week, consistent with flexible, opportunistic engagement rather than rigid scheduling.

This flexibility may be important, given high within-person variability in symptom severity both within and between days (Ebner-Priemer and Trull, 2009). The app can provide immediate support during periods of distress without waiting for a scheduled session. Consistent with this, the psychotherapy literature notes that scheduled sessions alone may be insufficient and that between-visit tools can be valuable (Kazantzis et al., 2010); more broadly, conventional care can create a timing mismatch between the need for and availability of providers (Mohr et al., 2014).

9 Disengagement and Effect Persistence

Disengagement from digital tools is common in health and education interventions and is often interpreted as evidence of limited effectiveness or value (Bhattacherjee, 2001, Eysenbach, 2005, Lipschitz et al., 2023, Torous et al., 2020b, Smith et al., 2025). Investors, international organizations, and donors often use engagement and retention to assess product effectiveness and value across domains (e.g., Gupta et al., 2004, Kumar et al., 2010, WHO, 2016, UNICEF, 2021, OECD, 2021).

We argue that declining use may also reflect achieved effects: digital tools can deliver information, skills, and behavioral strategies that users internalize, reducing the need for continued interaction once those inputs have been absorbed. This mechanism is particularly plausible for mental health interventions that rely on skill adoption. CBT, on which this app is based, improves mental health by helping people recognize and modify maladaptive thoughts and behaviors that contribute to psychological distress. As these changes are internalized, the marginal return to continued engagement may fall. Consistent with this view, standard CBT courses typically last

1.5 to 5 months and can generate effects that persist for at least six months (Van Dis et al., 2020).

If continued app use is necessary for sustained mental health gains, we expect minimal gains at 2 and 6 months among people who stop using the app after initial exposure. If instead benefits persist after use declines, mental health gains remain even among people with minimal subsequent use. If behavioral change is part of the mechanism, we should also observe: (i) sustained gains in mental health, sleep, and healthful behaviors despite declining use; (ii) evidence that sleep and healthful behaviors mediate treatment effects on mental health; and (iii) persistent adoption of app-recommended tools, as users internalize cognitive and behavioral strategies.

9.1 Does Longer App Access Increase Mental Health Gains? Randomizing Access Duration

To test whether longer engagement causes larger mental health gains, and, more broadly, whether continued engagement is a reliable proxy for product effectiveness and value, at 3 months we randomly extended app access to half of the treatment group for an additional 3 months. Attrition does not differ across arms, and baseline covariates are balanced (Tables M.1 and M.2). Table 8 shows that participants randomly assigned to 6 vs 3 months of app access used the app for an additional 68 minutes at six months ($p < 0.01$) – a 39% increase which, however, did not lead to either incremental mental health gains or changes in sleep and behavior: at three months, app use is decoupled from impacts. This is consistent with our conjecture that continued engagement is not necessary for sustained benefits, once users have adopted beneficial habits.

9.2 Can App-induced Skill adoption Contribute to Mental Health Gains? Indirect Evidence

Our experiments are not designed to identify whether app-induced skill adoption and behavioral change cause improvements in mental health. Nevertheless, the persistence of mental health gains despite declining app use suggests that continued use is not required for sustained benefits. Below, we present additional pieces of evidence consistent with this interpretation.

First, we have already established that impacts on mental health, 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; instead, impacts on key outcomes continue to be positive even when app use is minimal, e.g., in months 5 and 6, when only 10-20% of treated participants use the app.

Second, treated participants report greater use of CBT-consistent practices, including self-compassion, boundary setting, breathing exercises, writing about thoughts and feelings, and maintaining a sleep routine (Table J.1). This pattern is consistent with sustained adoption of app-recommended strategies, which could reduce the need for continued app use.[10]

Third, a mediation analysis shows that improvements in sleep and healthful behaviors account for 34-48% of the average ITT effect (Table J.2). Fourth, we estimate Conditional Average Treatment Effects (CATEs) for mental health and study their correlations with app use and CATEs for sleep and healthful behavior (Athey and Wager, 2019, Chernozhukov et al., 2025). Mental health CATEs are positive for all treatment members despite declining use. Also, people with larger impacts on sleep and healthful behavior also have larger mental health impacts, while people who use the app more do not have larger mental health impacts (Table J.3). Fifth, people with the worst baseline mental health have the largest mental health impacts from app access, especially at 2 and 6 months, despite not having the most intensive use (Figure K.1).

Taken together, these findings are consistent with skill- or behavior-adoption mechanisms: the app improves mental health, in part, by promoting the adoption of practices that persist beyond initial use. This interpretation aligns with the CBT literature, which emphasizes cognitive restructuring, behavioral activation, and skills practice as drivers of sustained symptom reduction (Cuijpers et al., 2013, 2019, 2023, Kazantzis et al., 2018). The notion that app-induced behavioral change yields durable mental health gains after discontinuation is also consistent with psychology and behavioral science research: behaviors are more likely to become habitual when benefits are immediate and salient, as early rewards strengthen learning and repetition through reinforcement and motivation (Lattal, 2010, Woolley and Fishbach, 2017, 2018). Benefits occur rapidly in our study. This may facilitate behavioral change without requiring continued use of the app.

10 Does Digital Therapy Crowd Out Psychotherapy?

A common concern about digital mental health tools is whether they crowd out in-person care. However, it is theoretically unclear whether this product acts as a complement to or a substitute for psychotherapy, and the relationship may vary across people. The ambiguity arises because digital tools can affect both the need for professional care and the propensity to seek it.

[10] Because these outcomes were not pre-specified, we exclude them from the primary behavior index.

10.1 Conceptual Framework

To illustrate how AI-powered digital care, A, affects in-person psychotherapy, T, consider individuals who value mental health directly and face a generalized resource constraint:

$$U = u(H,R), \qquad H = H(A,T), \qquad R = \bar{R} - q_A A - q_T T,$$

where $\bar{R}$ denotes total available resources and encompasses financial means, time, attention, logistical capacity, and the ability to seek and sustain treatment. Digital care is less costly than psychotherapy along both monetary and non-monetary dimensions ($q_A < q_T$).

This setup lets us make the following two points. First, some people may optimally choose $T = 0$ even when psychotherapy has positive health returns, because its cost exceeds their available resources. This can describe people with low income, high stigma, or severe distress that impairs functioning and makes treatment difficult to initiate or sustain. Second, the introduction of A may decrease, increase, or not change psychotherapy use. Crowding out may occur when A and T are substitutes in health production ($H_{AT} < 0$).

Crowding in may occur through either of two channels. One is technological complementarity ($H_{AT} > 0$), so that app use raises the marginal return to psychotherapy. The other is lower cost of psychotherapy. Psychotherapy entails both monetary and cognitive, stigma, and logistical costs. These non-monetary costs are likely higher when symptoms and functional impairment are more severe (Sweetman et al., 2021, Schnyder et al., 2017). To capture that, suppose that the cost of psychotherapy depends on mental health, so that $q_T = q_T(H)$ with $q_T'(H) < 0$. By improving mental health, A may also lower the cost of T, making psychotherapy newly feasible when it was previously out of reach. The effect of digital care on psychotherapy use is thus *ex ante* ambiguous. Appendix N describes the model.

Lastly, because digital care is less costly than psychotherapy, the app may expand access to mental health support without changing optimal psychotherapy use. This occurs when uptake is concentrated among individuals who would otherwise have remained untreated because psychotherapy was too costly, stigmatized, or difficult to access.

These channels generate distinct predictions with respect to baseline psychological distress. Substitution predicts that psychotherapy use is more likely to decrease among participants with mild baseline distress, who may no longer need in-person care once the app improves their health. Complementarity predicts that the app may crowd in psychotherapy among participants

with moderate or severe baseline distress, if app access increases the productivity of therapy. Instead, if the app decreases therapy's non-monetary costs, crowd-in should be strongest among participants whose non-monetary costs decrease the most. Finally, the access-expansion channel predicts that the app may improve mental health even when psychotherapy use does not change.

10.2 Empirical Evidence

Table 9 shows a pooled 35% increase in the likelihood of visiting a psychologist over the control mean. The impacts of 3, 5, and 8 pp in this likelihood at 1, 2, and 6 months ($p<0.1$, <0.01, and <0.01) denote gradual increases over time, with proportionally comparable impacts on the number of sessions. Thus, in this case app access acted as a complement rather than a substitute for traditional psychotherapy.[11] Although the app included referral options for online therapy (which involved out-of-pocket costs), only 4 participants contacted therapists through the app.

Next, we examine how impacts on psychotherapy use, mental health, beliefs, and other non-monetary costs of therapy vary by baseline 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. Table 10 pools observations across all available 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 person. The increase in psychotherapy is concentrated among people with moderate and severe baseline distress (column 1), for whom we estimate effects of 0.07 ($p<0.01$) and 0.08 ($p<0.01$), larger than the small and insignificant effects on people with mild distress. App access reduces symptoms of depression and anxiety for all groups (columns 2 and 3), but especially for people with moderate baseline distress, and the results are qualitatively unchanged for the mental health index (column 4).[12] First-order beliefs about mental illness do not change, possibly because negative views of mental illness are uncommon in this sample (column 5). Conversely, second-order beliefs improve (column 6), but by a small amount and not more so among people with moderate and severe baseline distress. Thus, the higher demand for psychotherapy is unlikely to be driven primarily by less negative views about mental illness. Lastly, ease of access to psychotherapy increases by 0.22 SD for people with high baseline distress, although not differently from the other groups (column 7). These results are consistent

[11] Restricting the sample to non-attriters across the 3 surveys leaves the results qualitatively unchanged (Table N.1).

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

with the complementarity and cost-reduction predictions: the app likely increases the use of psychotherapy by making symptoms of depression and anxiety more manageable for some people. This, in turn, may both increase the marginal return to psychotherapy and lower their non-monetary costs of psychotherapy.[13] The idea that non-monetary costs act as demand-side barriers to psychotherapy is consistent with evidence from Breza et al. (2026).

Does Psychotherapy explain the mental health impacts? Because app access increases psychotherapy use, we assess whether psychotherapy mediates the estimated mental health effects. A mediation analysis using psychotherapy attendance in the prior month as the mediator (Table N.2) shows that psychotherapy accounts for at most 5% of the ITT effect at the 2-month follow-up ($p<0.05$), with smaller, statistically insignificant mediation shares at 1 and 6 months. Thus, increased psychotherapy explains only a limited portion of the overall mental health gains.[14]

As an additional check, we compare mental health CATEs across participants who did versus did not attend psychotherapy in the prior month. The CATEs are approximately 0.30 SD for both groups, with small, statistically insignificant differences of $0.002 - 0.005$. This analysis is methodologically distinct from the mediation exercise but supports the same conclusion.

Lastly, Figure N.1 re-estimates the main treatment effects for participants who did not attend psychotherapy during the study period. This exercise conditions on a post-treatment outcome and should 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 Potential to Scale and Concluding Thoughts

AI-enabled apps may be a valuable new mental health care tool, but their effectiveness, safety, and value has remained unclear. In this study, access to a CBT-based app improved mental health by 0.3 SD over six months with no evidence of an increase of severe cases; improved sleep, healthful behaviors, daily functioning, and labor market outcomes; and increased psychotherapy

[13]The results are qualitatively unchanged when grouping 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, if diagnosed, are offered psychotherapy or pharmacotherapy. There is no screening for PSS-4 and WHO-5, the other two outcomes that we use to create the baseline mental health index.

[14]In unreported regressions, we also rejected the hypothesis that app access affects the use of antidepressants, anxiolytics, and mental health supplements.

use rather than crowding it out. Gains persisted well beyond the period of highest use, indicating that continued engagement was not necessary for sustained benefits. Given the intervention's low cost, the implied economic benefits are sizable. These findings suggest that digital mental health support can be an effective complement to existing care.

The results are also encouraging through the lens of scalability, considering the five threats to scale in List (2024). First, concerns about false positives are minimal: app access was randomized, and effects are both consistent across related outcomes and with signs aligned with expert predictions. Second, the study population is policy relevant, although it would be important to replicate findings in different high-need populations.[15] Third, we find no evidence of adverse spillovers within the outcomes we observe; if anything, we document improvements in sleep, healthful behavior, labor-market outcomes, and psychotherapy use. Conceptually, we do not expect adverse spillovers (e.g., from equilibrium effects) at scale. Fourth, the technology exhibits economies of scale, as marginal costs are likely to decline with broader deployment. Fifth, the study did not take place under unusual circumstances, suggesting results are unlikely to differ systematically over time, although replication in other contexts would be valuable. Overall, these patterns point to strong potential for scale.

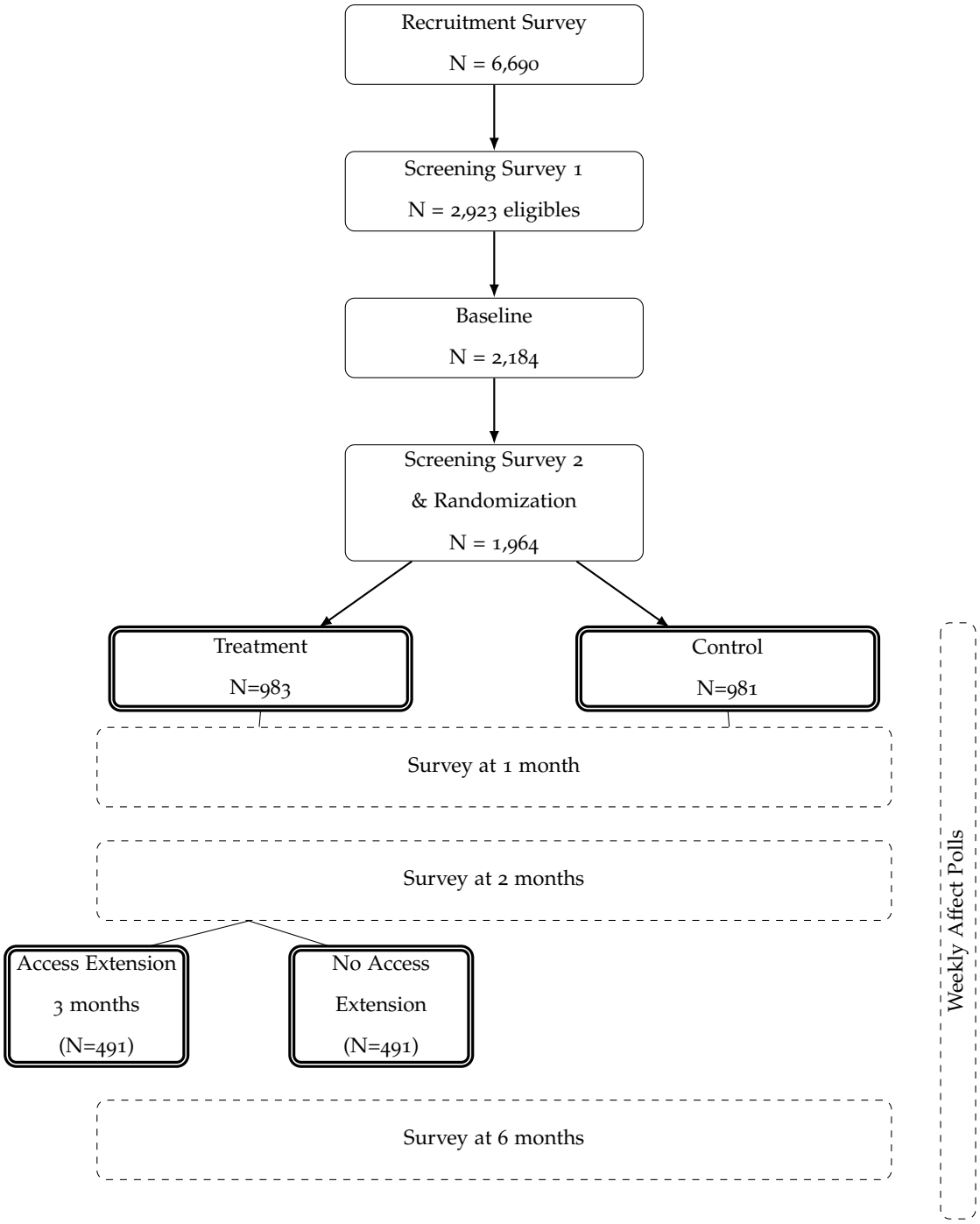
More broadly, our findings speak to the adoption and evaluation of digital technologies across sectors beyond health. One implication is that short-lived interactions can generate persistent behavioral change, decoupling engagement and impacts. As a result, usage metrics may systematically understate effectiveness and value when technologies build skills, habits, or knowledge that persist beyond active use. This has implications for how digital tools are evaluated across a range of domains, such as education, financial services, and agriculture, where interventions may operate by shifting behavior rather than requiring continuous interaction.

A second implication concerns how AI technologies interact with skilled services. Despite concerns about substitution, AI may instead complement human expertise by augmenting productivity and changing task allocation. In such cases, digital tools can reduce frictions, expand access, and facilitate engagement with higher-quality services rather than displace them.

[15] For instance, among pregnant women, depression risk is elevated, while antidepressant use declines and psychotherapy use increases (Boone et al., 2025). Similar considerations apply to other groups facing access or stigma-related barriers, such as adolescents or frontline workers.

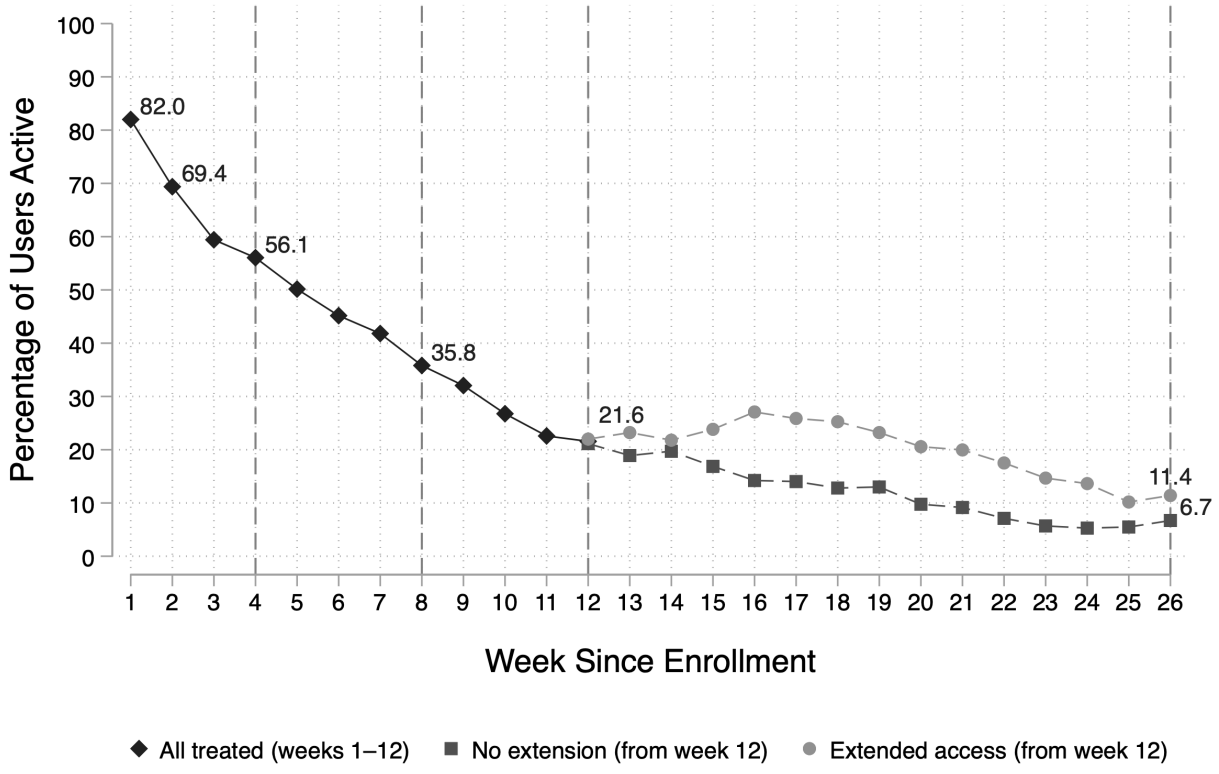
12 Tables and Figures

Figure 1: Experimental Design and Timeline



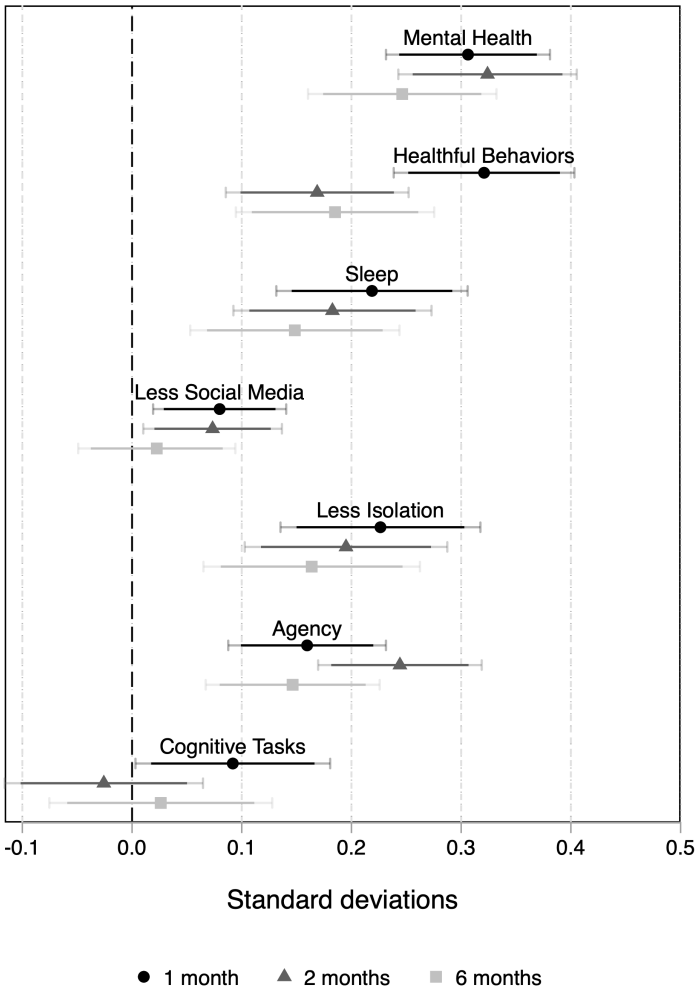
Notes: The figure shows a flowchart with the experimental design and timeline of the randomized controlled trial. The process began with the recruitment of N=6,690 participants, followed by Screening Survey 1, which yielded N=2,923 eligible participants (3,767 were screened out). After baseline data 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-, 2-, and 6-month marks. At three months, half of the people in the treatment group were randomly assigned to receive App access for an additional three 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



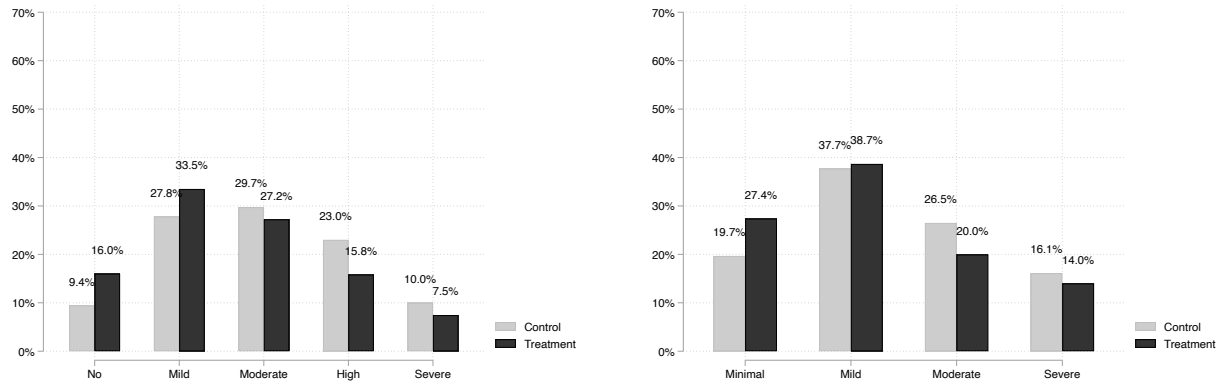
Notes: Engagement is defined as the percentage of individuals assigned to the treatment group who used the application at least once during a given week. Individuals who never downloaded the app are coded as zero use. Vertical dashed lines indicate the timing of the follow-up surveys conducted at 1,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

Figure 3: Summary Impacts by Family Indices



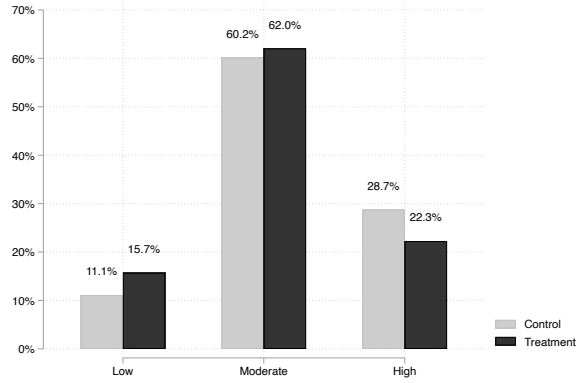
Notes: Each point reports the estimated treatment effect on a standardized outcome index, constructed following Anderson (2008). *Higher values indicate better outcomes*, and all variables are oriented so that positive effects reflect improvements (i.e., outcomes are reverse-coded when necessary). The *Mental Health Index* combines the WHO-5 well-being score, PHQ-8, GAD-7, and PSS-4 (the latter three reversed). The *Healthful Behaviors Index* summarizes the days in the past week during which participants reported engaging in positive daily activities (exercising, not missing work, practicing self-care, going out, not raising one's voice in anger, helping someone with homework). The *Sleep Index* aggregates hours of sleep, number of interruptions (reversed), and sleep difficulties (reversed). The *Less Social Media Index* captures platform use (Facebook, X/Twitter, Instagram, TikTok) and frequency of use (more than 2 hours per day, reversed). The *Less Isolation Index* is based on the six-item De Jong Gierveld Loneliness Scale (reversed), combining the social and emotional subscales. The *Agency Index* combines scores from the General Self-Efficacy (GSE) Scale and locus of control measures. The *Cognitive tasks index* combines performance in two incentivized tasks measuring cognitive function and effort. All regressions control for randomization strata and LASSO-selected baseline covariates.

Figure 4: Mental health outcomes by treatment status and symptom severity

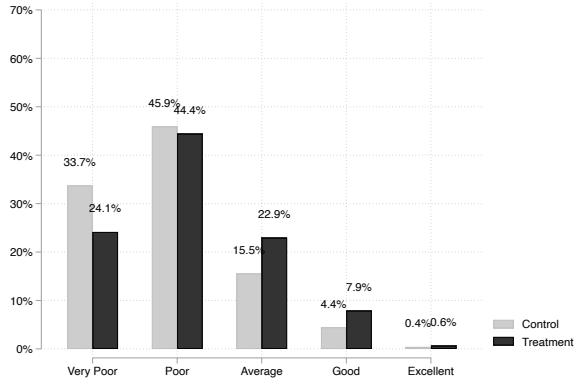


(a) Depression (PHQ-8)

(b) Anxiety (GAD-7)



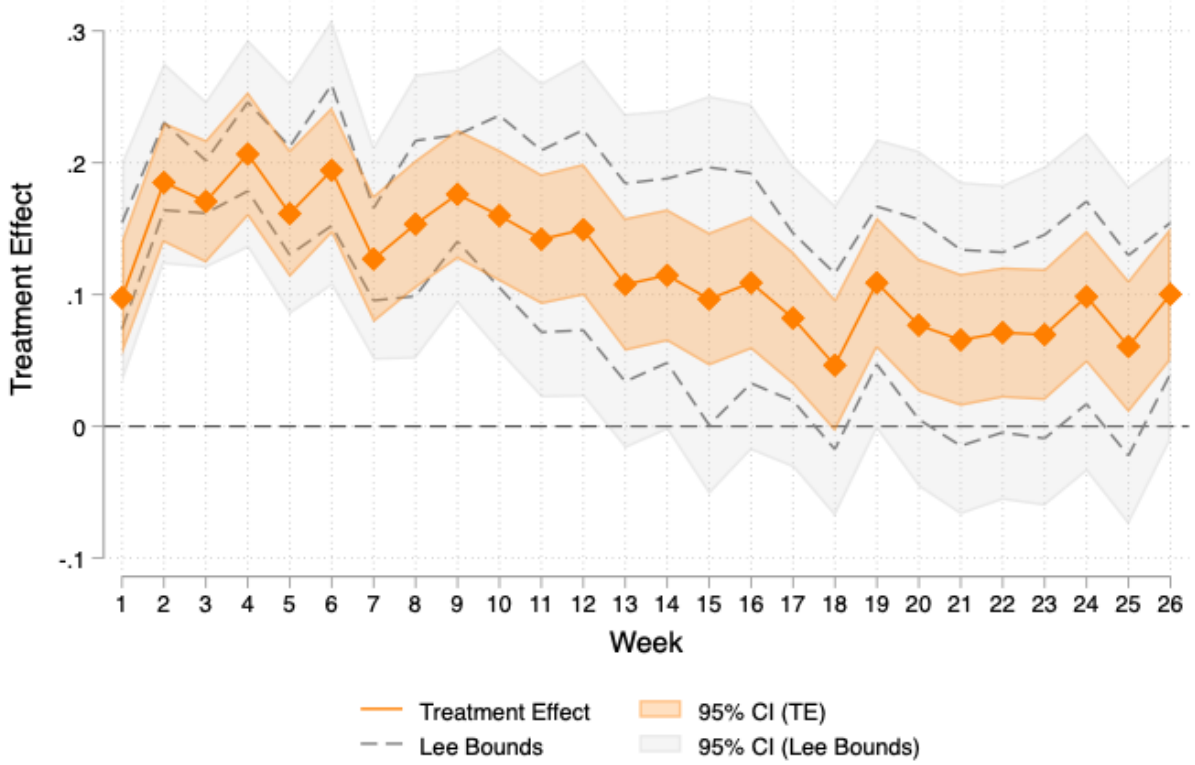
(c) Stress (PSS-4)



(d) Well-being (WHO-5)

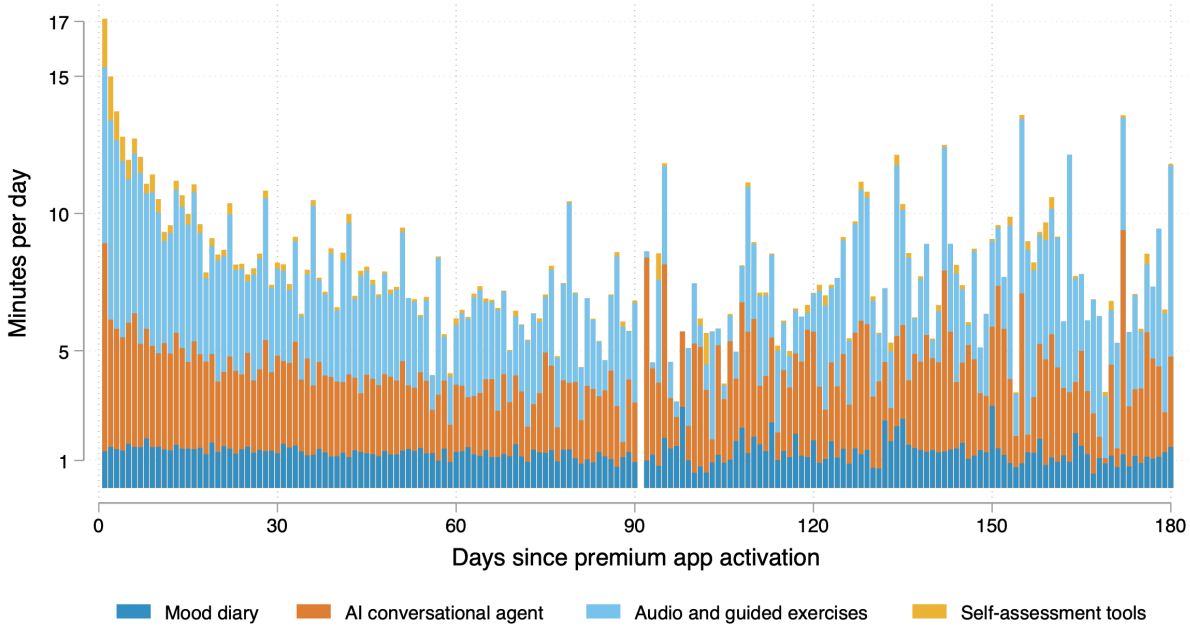
Notes: Each panel shows the percent by symptom severity and arm for the pooled data. The cutoffs used were: PHQ-8: 0–4 = No, 5–9 = Mild, 10–14 = Moderate, 15–19 = High, 20–24 = Severe; GAD-7: 0–4 = Minimal, 5–9 = Mild, 10–14 = Moderate, 15–21 = Severe; PSS-4: 0–5 = Low, 6–10 = Moderate, 11–16 = High;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

Figure 5: Impacts on Happy Affect from Weekly Data



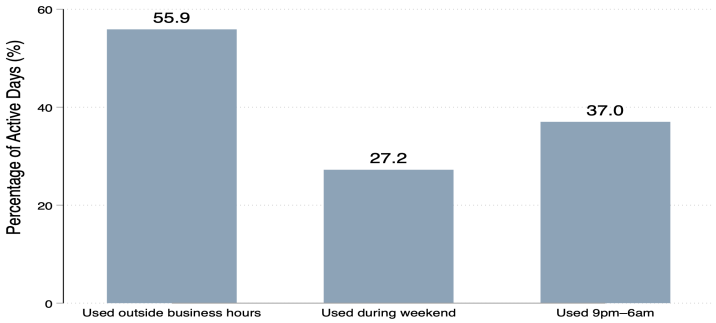
Notes: This figure displays the estimated treatment effect on the probability of having a "Happy" affect over 26 weeks. The orange line represents the point estimate with 95% confidence intervals (shaded orange area), calculated using a regression with baseline controls and strata fixed effects, with 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 6: Daily App Use by Activity Type

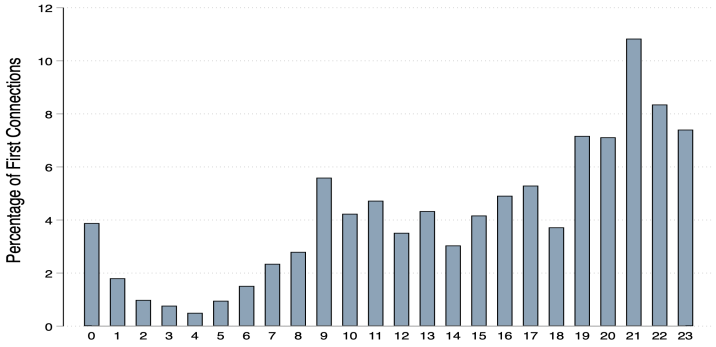


Notes: The figure plots average daily minutes of app use among active users in the treatment group, decomposed into the four core app functions: AI conversational agent, audio/guided exercises, mood diary, and self-assessment tools. We define event duration from consecutive timestamps within sessions, accounting for background inactivity. There are no active users on day 91. This is when the premium subscription expires, and before it is renewed for an additional 90 days for randomly selected participants in the treatment group.

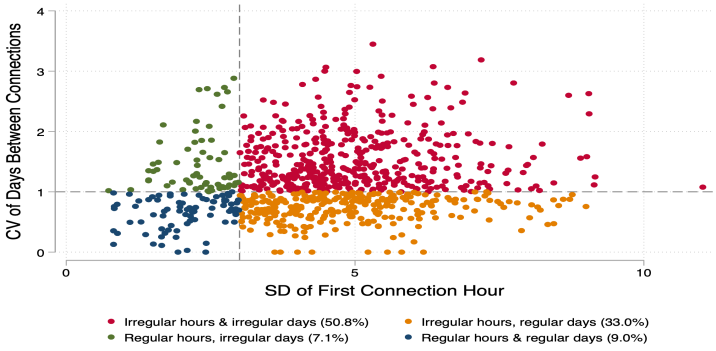
Figure 7: App Use Patterns by Hour and Variability



(a) Use outside regular hours



(b) Hour of first connection



(c) App use variability by weekday and hours

Notes: Use patterns for the treatment users who downloaded the app. Panel A shows the share of app use occurring outside regular hours (weekends or before 8:00 and after 21:00). We calculate percentages as the proportion of users with at least one connection outside regular hours in a given week. Panel B displays the distribution of the first connection time of the day. We calculate percentages as (number of days with first connection at a given hour / total active days) × 100, based on n=23,016 active days. Panel C classifies users based on usage regularity. The X axis represents the circular standard deviation (SD) of the first connection hour (SD $\leq$ 3 hours defined as regular hours). The Y axis represents the circular coefficient of variation (CV) of days between connections (CV $\leq$ 1 defined as regular days). Both measures use circular statistics to account for the periodicity of time. We compute the CV (=SD/mean) for days of the week because weekdays are coded 1–7. Two users who connect regularly but on different day pairs (e.g., days 1 and 3 versus 1 and 5) can have different variances; the CV normalizes by the mean to yield a more comparable measure of regularity across users with different use frequencies. We calculate percentages based on 884 users with three or more active days.

Table 1: Impacts on Mental Health

	Depression (PHQ-8)	Anxiety (GAD-7)	Stress (PSS-4)	Well-being (WHO-5)	**Mental Health Index (std)**
	(1)	(2)	(3)	(4)	(5)
	[−]	[−]	[−]	[+]	[+]
Panel A. Pooled Effects					
Treated	-1.627***	-0.981***	-0.627***	5.254***	0.294***
	(0.141)	(0.131)	(0.070)	(0.465)	(0.025)
Mean control	11.82	9.06	8.94	35.59	0.00
Observations	5272	5274	5273	5273	5274
Panel B. Effects at 1 month					
Treated	-1.690***	-1.006***	-0.559***	5.504***	0.306***
	(0.214)	(0.203)	(0.101)	(0.681)	(0.038)
Mean control	12.30	9.49	9.16	33.79	0.00
Observations	1849	1850	1850	1850	1850
Panel C. Effects at 2 months					
Treated	-1.604***	-0.968***	-0.834***	5.684***	0.324***
	(0.231)	(0.217)	(0.116)	(0.768)	(0.041)
Mean control	11.91	9.11	9.00	35.14	0.00
Observations	1824	1825	1825	1824	1825
Panel D. Effects at 6 months					
Treated	-1.544***	-0.997***	-0.441***	4.619***	0.246***
	(0.248)	(0.232)	(0.130)	(0.859)	(0.044)
Mean control	11.17	8.51	8.62	38.20	0.00
Observations	1599	1599	1598	1599	1599
1 mo=2 mos (*p-val*)	0.72	0.93	0.08	0.90	0.79
2 mos=6 mos (*p-val*)	0.82	0.99	0.02	0.32	0.18
1 mo=6 mos (*p-val*)	0.53	0.95	0.41	0.36	0.27
1 mo=2 mos=6 mos (*p-val*)	0.84	0.98	0.06	0.57	0.36

Notes: Each cell reports the estimated treatment effect on the specified mental health outcome using each scale's 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 (0–24; $\geq$ 10 indicates moderate to severe depression), the GAD-7 for anxiety (0–21; $\geq$ 10 indicates moderate to severe 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. * $p < .10$, ** $p < .05$, *** $p < .01$.

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) [+]	**Healthful Behaviors Index (std)** (9) [+]
Panel A. Pooled Effects									
Treated	-0.078**	0.124**	-0.239***	0.153***	0.227***	0.339***	0.199***	0.197***	0.226***
	(0.031)	(0.052)	(0.047)	(0.025)	(0.047)	(0.049)	(0.061)	(0.026)	(0.026)
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	-0.090*	0.175**	-0.346***	0.211***	0.335***	0.506***	0.269***	0.299***	0.321***
	(0.051)	(0.081)	(0.073)	(0.041)	(0.073)	(0.079)	(0.097)	(0.042)	(0.042)
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	-0.065	0.094	-0.181**	0.122***	0.190**	0.211***	0.104	0.127***	0.169***
	(0.047)	(0.086)	(0.079)	(0.041)	(0.076)	(0.081)	(0.100)	(0.043)	(0.043)
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	-0.082	0.101	-0.182**	0.125***	0.138	0.284***	0.207*	0.156***	0.185***
	(0.061)	(0.093)	(0.083)	(0.045)	(0.088)	(0.086)	(0.107)	(0.046)	(0.046)
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*)	0.67	0.44	0.11	0.12	0.12	0.01	0.21	0.01	0.01
2 mos=6 mos (*p-val*)	0.78	1.00	0.95	1.00	0.63	0.59	0.58	0.71	0.86
1 mo=6 mos (*p-val*)	0.95	0.45	0.13	0.13	0.06	0.06	0.60	0.03	0.02
1 mo=2 mos=6 mos (*p-val*)	0.93	0.67	0.21	0.20	0.12	0.02	0.47	0.01	0.01

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) *Missed Work*, did not go to work (unconditional on whether they work); (ii) *Helped*, helped their children or someone else with schoolwork, (iii) *Angry*, whether they raise their voice to someone else in anger, (iv) *Functioning Subindex (std)*, a standardized composite index combining Missed Worked, Helped, and Angry measures, oriented so that higher values indicate better daily functioning. (v) *Exercised*, (vi) *Go out*, whether the respondent went out for activities unrelated to school or work; (vii) *Self Care*, spent at least 10 minutes on themselves, (viii) *Wellness Subindex (std)*, a standardized composite index combining Exercised, Go out, and Self Care measures, oriented so that higher values indicate improved wellness. (ix) *Healthful Behavior Index (std)*, a standardized composite index combining all measures oriented so that higher values indicate more healthful behaviors. All standardized composite indexes are build using the inverse-covariance weighting method following Anderson (2008). Regressions control for strata and lasso-selected baseline covariates. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 3: Impacts on Labor Market Outcomes

	Works	Hours Worked	Days Absent	**Labor Index (std)**
	(1)	(2)	(3)	(4)
	[+]	[+]	[−]	[+]
Panel A. Pooled Effects (2 and 6 months)				
Treated	0.031**	0.795	-0.075*	0.096***
	(0.014)	(0.792)	(0.038)	(0.031)
Mean control	0.50	30.04	0.49	0.00
Observations	3423	3420	3424	3424
Panel B. Effects at 2 months				
Treated	0.022	1.695	-0.065	0.102**
	(0.019)	(1.050)	(0.047)	(0.042)
Mean control	0.50	28.86	0.44	0.00
Observations	1824	1821	1825	1825
Panel C. Effects at 6 months				
Treated	0.044**	-0.053	-0.082	0.090**
	(0.021)	(1.142)	(0.061)	(0.045)
Mean control	0.50	31.40	0.54	0.00
Observations	1599	1599	1599	1599
2 mos=6 mos (*p-val*)	0.48	0.22	0.78	0.78

Notes: Each cell shows the estimated treatment effect on one of four labor market outcomes. Robust standard errors, clustered by individual in the pooled data. 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: *Works*, a binary indicator equal to one if the respondent is currently working; *Hours worked*, the total number of hours worked in the past week; *Days absent*, the self-reported number of days absent from work in the past 7 days; and the *Labor 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 labor market outcomes. Panel A pools effects measured at 2 and 6 months only. The coefficient estimate of 'days absent' in Panel A of this table differs from the analogous coefficient in Panel A of Table 2 because that table pools data at 1, 2, and 6 months. The results are qualitatively unchanged. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 4: Impacts on Sleep Outcomes

	Hours Slept (1) [+]	Interruptions (2) [−]	Sleep Issues (3) [−]	**Sleep Index (std)** (4) [+]
Panel A. Pooled Effects				
Treated	0.166***	-0.110**	-0.106***	0.189***
	(0.050)	(0.043)	(0.013)	(0.028)
Mean control	7.05	2.81	0.50	0.00
Observations	5269	5274	5274	5274
Panel B. Effects at 1 month				
Treated	0.205**	-0.118*	-0.129***	0.219***
	(0.081)	(0.070)	(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	0.164*	-0.109	-0.096***	0.183***
	(0.086)	(0.069)	(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	0.096	-0.085	-0.092***	0.148***
	(0.090)	(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*)	0.69	0.87	0.33	0.65
2 mos=6 mos (*p-val*)	0.60	0.68	0.95	0.53
1 mo=6 mos (*p-val*)	0.35	0.75	0.34	0.25
1 mo=2 mos=6 mos (*p-val*)	0.62	0.96	0.51	0.57

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 Slept*, 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-8 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 5: Impacts on Social Media Use

	Facebook	X	Instagram	TikTok	Use Freq.	**Less Social Media Index (std)**
	(1)	(2)	(3)	(4)	(5)	(6)
	[-]	[-]	[-]	[-]	[-]	[+]
Panel A. Pooled Effects						
Treated	-0.105**	-0.106***	-0.049	0.074	-0.028**	0.061***
	(0.050)	(0.031)	(0.048)	(0.050)	(0.011)	(0.020)
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	-0.122	-0.139***	-0.184**	0.037	-0.016	0.080***
	(0.076)	(0.048)	(0.075)	(0.076)	(0.018)	(0.031)
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	-0.133	-0.064	-0.044	0.008	-0.045**	0.073**
	(0.082)	(0.051)	(0.078)	(0.081)	(0.019)	(0.032)
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	-0.040	-0.116**	0.093	0.177*	-0.024	0.023
	(0.091)	(0.057)	(0.088)	(0.093)	(0.020)	(0.037)
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*)	0.95	0.24	0.14	0.89	0.28	0.80
2 mos=6 mos (*p-val*)	0.49	0.58	0.28	0.21	0.43	0.32
1 mo=6 mos (*p-val*)	0.46	0.60	0.02	0.26	0.79	0.19
1 mo=2 mos=6 mos (*p-val*)	0.77	0.53	0.05	0.41	0.53	0.44

Notes: Each cell shows the estimated treatment effect on 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 of social media use for two hours or more per day. Column (6) reports *Social Media Index (std)*, which 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 6: Impacts on Feelings of Isolation and Personal Agency

	Isolation			*Agency*		
	Emotional Score	Social Score	**Less Isolation Index (std)**	Self-Efficacy (GSE)	Locus of Control	**Agency Index (std)**
	(1)	(2)	(3)	(4)	(5)	(6)
	[−]	[−]	[+]	[+]	[+]	[+]
Panel A. Pooled Effects						
Treated	-0.165***	-0.120***	0.198***	0.408***	0.520***	0.185***
	(0.026)	(0.026)	(0.029)	(0.078)	(0.069)	(0.023)
Mean control	2.21	2.48	0.00	15.86	12.10	0.00
Observations	5274	5274	5274	5273	5272	5274
Panel B. Effects at 1 month						
Treated	-0.170***	-0.149***	0.226***	0.367***	0.407***	0.160***
	(0.040)	(0.042)	(0.046)	(0.121)	(0.107)	(0.037)
Mean control	2.30	2.51	0.00	15.66	12.15	0.00
Observations	1850	1850	1850	1850	1850	1850
Panel C. Effects at 2 months						
Treated	-0.187***	-0.098**	0.195***	0.569***	0.651***	0.244***
	(0.043)	(0.042)	(0.047)	(0.130)	(0.112)	(0.038)
Mean control	2.23	2.48	0.00	15.81	12.02	0.00
Observations	1825	1825	1825	1824	1824	1825
Panel D. Effects at 6 months						
Treated	-0.125***	-0.122**	0.164***	0.252*	0.495***	0.146***
	(0.048)	(0.048)	(0.050)	(0.143)	(0.125)	(0.040)
Mean control	2.09	2.45	0.00	16.15	12.13	0.00
Observations	1599	1599	1599	1599	1598	1599
1 mo=2 mos (*p-val*)	0.78	0.34	0.56	0.32	0.14	0.14
2 mos=6 mos (*p-val*)	0.34	0.58	0.71	0.10	0.39	0.08
1 mo=6 mos (*p-val*)	0.44	0.70	0.38	0.54	0.64	0.78
1 mo=2 mos=6 mos (*p-val*)	0.64	0.60	0.67	0.22	0.25	0.24

Notes: Each cell reports the estimated treatment effect on isolation and agency outcomes. Isolation is measured using the six-item De Jong Gierveld Loneliness Scale, which includes three items capturing emotional loneliness and three capturing social loneliness; higher values indicate greater loneliness. We report effects separately for emotional and social subscales, as well as a standardized index reoriented so that higher values indicate less isolation.
Agency outcomes include self-efficacy, measured using the General Self-Efficacy (GSE) scale, and locus of control; we also report a standardized agency index. Indices are constructed following Anderson (2008). Regressions control for strata and lasso-selected baseline covariates. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table 7: Impacts on Cognitive Tasks

	Effort Task	*Emotional Stroop Task*		
	Correct/minute	Emotional words	Neutral words	**Cognitive Tasks Index (std)**
	(1)	(2)	(3)	(4)
	[−]	[−]	[−]	[+]
Panel A. Pooled Effects				
Treated	0.003	-8.409*	-6.530	0.031
	(0.061)	(4.858)	(4.608)	(0.029)
Mean control	7.22	993.43	999.57	-0.00
Observations	5274	5274	5274	5274
Panel B. Effects at 1 month				
Treated	0.022	-24.518***	-15.427*	0.092**
	(0.104)	(7.806)	(7.967)	(0.045)
Mean control	6.72	1011.62	1026.21	-0.00
Observations	1850	1850	1850	1850
Panel C. Effects at 2 months				
Treated	-0.011	5.973	4.676	-0.026
	(0.099)	(7.770)	(7.274)	(0.046)
Mean control	7.49	988.00	988.70	-0.00
Observations	1825	1825	1825	1825
Panel D. Effects at 6 months				
Treated	-0.020	-5.185	-9.421	0.026
	(0.102)	(9.011)	(8.078)	(0.052)
Mean control	7.49	978.50	981.05	-0.00
Observations	1599	1599	1599	1599
1 mo=2 mos (*p-val*)	0.71	0.00	0.04	0.04
2 mos=6 mos (*p-val*)	0.81	0.26	0.17	0.42
1 mo=6 mos (*p-val*)	0.66	0.10	0.51	0.27
1 mo=2 mos=6 mos (*p-val*)	0.88	0.01	0.16	0.14

Notes: Column (1) reports the average number of correctly solved matrices per minute in the effort task. Column (2) presents results from the emotional block of the Emotional Stroop test. Column (3) presents results from the neutral block, pooling the two neutral word sets (presented before and after the emotional block) by stacking observations and including a block fixed effect. [+] indicates that higher values represent better outcomes; [−] indicates that higher values represent worse outcomes. The Stroop measures capture reaction time in milliseconds, so positive coefficients indicate slower responses (greater emotional interference). The pooled specification combines observations across all three follow-up waves (1, 2, and 6 months post-intervention) and includes wave fixed effects. All regressions control for lasso-selected baseline covariates and strata fixed effects. Robust standard errors are reported in parentheses and clustered by person for pooled effects. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 8: Effects of Randomized Extended Access to the App

		Mental Health					Other Indexes				
	Total Use (minutes)	Depression (PHQ-8)	Anxiety (GAD-7)	Stress (PSS-4)	Well-Being (WHO-5)	Mental Health Index	Healthful Behaviors	Sleep	Less Social Media	Less Isolation	Agency
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Extended 6 months	241.789***	-1.400***	-0.967***	-0.414**	3.785***	0.231***	0.141**	0.113*	0.066	0.122**	0.153***
	(21.808)	(0.307)	(0.287)	(0.164)	(1.055)	(0.055)	(0.058)	(0.058)	(0.046)	(0.062)	(0.050)
No extension	173.572***	-1.692***	-1.027***	-0.470***	5.471***	0.261***	0.230***	0.185***	-0.022	0.206***	0.140***
	(12.577)	(0.305)	(0.288)	(0.162)	(1.081)	(0.054)	(0.055)	(0.062)	(0.044)	(0.064)	(0.050)
Difference (Ext. − No ext.)	68.217***	0.292	0.060	0.056	-1.685	-0.030	-0.090	-0.072	0.089*	-0.083	0.012
	(25.216)	(0.359)	(0.341)	(0.195)	(1.270)	(0.066)	(0.067)	(0.070)	(0.053)	(0.077)	(0.058)
Mean control	0.42	11.17	8.51	8.62	38.20	0.00	0.00	0.00	0.00	0.00	0.00
Observations	1599	1599	1599	1598	1599	1599	1599	1599	1599	1599	1599

Notes. This table reports estimates from the second randomization at the 6-month endline. The sample includes individuals assigned to treatment in the first randomization. "Extended 6 months" indicates individuals randomized to continued access; "No extension" indicates individuals whose access ended after the initial period. The omitted category is the control group. Column (1) reports effects on total platform usage (minutes) over the study period. Columns (2)–(6) report effects on individual mental health measures and a summary index. Columns (7)–(11) report effects on other outcome family indexes. Regressions control for strata and lasso-selected baseline covariates. Robust standard errors in parentheses. The difference row reports the coefficient and standard error from a linear combination test (Extended − No extension). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 9: Impacts on the Demand for Psychotherapy

	Visit Psychologist (1)	Num. Visits (2)
Panel A. Pooled Effects		
Treated	0.052***	
	(0.010)	
Mean control	0.15	
Observations	5273	
Panel B. Effects at 1 month		
Treated	0.032**	
	(0.016)	
Mean control	0.15	
Observations	1850	
Panel C. Effects at 2 months		
Treated	0.054***	0.148***
	(0.017)	(0.052)
Mean control	0.16	0.38
Observations	1824	1824
Panel D. Effects at 6 months		
Treated	0.076***	0.289***
	(0.018)	(0.073)
Mean control	0.15	0.51
Observations	1599	1599
1 mo=2 mo (*p-val*)	0.36	
2 mos=6 mos (*p-val*)	0.29	0.10
1 mo=6 mos (*p-val*)	0.07	
1 mo=2 mos=6 mos (*p-val*)	0.17	

Notes: The first outcome, *Visit Psychologist*, is an indicator variable equal to 1 if the respondent reports at least one visit to 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 10: Impacts on Therapy Use, Depression, Anxiety, Related Beliefs, and Therapy Access Costs by Baseline Distress

	Visit Psychologist	Depression (PHQ-8)	Anxiety (GAD-7)	Mental Health Index (std)	1st-order Beliefs Index (std)	2nd-order Beliefs Index (std)	Access Costs Index (std)
	(1)	(2) [+]	(3) [−]	(4) [−]	(5) [+]	(6) [+]	(7) [+]
Panel A. Mild baseline distress (PHQ-4 = 3–5)							
Treatment	0.019	-1.092***	-0.496***	0.208***	0.043	0.125**	0.064
	(0.015)	(0.198)	(0.181)	(0.036)	(0.055)	(0.054)	(0.072)
Panel B. Moderate baseline distress (PHQ-4 = 6–8)							
Treatment	0.065***	-2.565***	-1.817***	0.420***	0.068	0.187***	0.110
	(0.018)	(0.249)	(0.227)	(0.042)	(0.063)	(0.063)	(0.086)
Panel C. Severe baseline distress (PHQ-4 = 9–12)							
Treatment	0.084***	-1.390***	-0.642**	0.280***	-0.083	0.069	0.223**
	(0.021)	(0.311)	(0.303)	(0.055)	(0.076)	(0.076)	(0.097)
Panel D. Overall ITT							
Treatment	0.052***	-1.627***	-0.981***	0.294***	0.019	0.120***	0.102**
	(0.010)	(0.141)	(0.131)	(0.025)	(0.036)	(0.036)	(0.048)
Mean control	0.15	11.82	9.06	0.00	0.00	0.00	0.00
Observations	5273	5272	5274	5274	3117	3115	1599
Panel E. Tests of equality of treatment effects (p-values)							
Mild = Moderate	0.051	0.000	0.000	0.000	0.762	0.455	0.679
Mild = Severe	0.011	0.419	0.680	0.273	0.183	0.545	0.188
Moderate = Severe	0.466	0.003	0.002	0.043	0.127	0.229	0.384
Observations	5273	5272	5274	5274	3117	3115	1599
Measured at 1 mo	Yes	Yes	Yes	Yes	No	No	No
Measured at 2 mos	Yes	Yes	Yes	Yes	Yes	Yes	No
Measured at 6 mos	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: Baseline mental health distress is *Mild* (PHQ-4 = 3–5), *Moderate* (PHQ-4 = 6–8), or *Severe* (PHQ-4 = 9–12). Panel D reports the ITT estimate on the full sample. Panel E reports p-values from tests of equality of subgroup-specific treatment effects. All specifications use LASSO selected control variables and include randomization strata and baseline dependent variable when available. Standard errors are clustered by person. The *1st-order Beliefs Index* is a Likert-scale index comprising items that measure the respondent's perceived stigma toward mental health. The *2nd-order Beliefs Index* measures the respondent's perceptions of stigma toward mental health in their neighborhood. The *Access Costs Index* is a three-item index measuring: whether the respondent knows how to find a psychologist, how to make an appointment, and has contact information for a mental health professional. All standardized indices have a mean of zero in the control group.$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

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.

Allcott, H., Gentzkow, M., and Song, L. (2022). Digital addiction. American Economic Review, 112(7):2424–2463.

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.

Arias, D., Saxena, S., and Verguet, S. (2022). Quantifying the global burden of mental disorders and their economic value. EClinicalMedicine, 54.

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.

Bhattacherjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. MIS Quarterly, 25(3):351–370.

Boone, C., Colina, C., and Pope, D. (2025). Antidepressant use before, during, and after pregnancy. JAMA Network Open, 8(1):e2457324.

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.

Breza, E., Carney, K., Raghavan, V., Rajah, K., Rangaswamy, T., Rao, G., Schilbach, F., Shadbar, S., and Stratton, J. (2026). Financial incentives, health screening, and selection into mental health care: Experimental evidence from college students in india. NBER Working Paper 34819, National Bureau of Economic Research.

Chandra, A. and Skinner, J. (2012). Technology growth and expenditure growth in health care. Journal of Economic Literature, 50(3):645–80.

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.

Crowne, D. P. and Marlowe, D. (1960). Marlowe-crowne social desirability scale. Journal of Consulting Psychology.

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.

Cutler, D. M. and McClellan, M. (2001). Is technological change in medicine worth it? Health affairs, 20(5):11–29.

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.

Derese, A., Gebreegzhiabhere, Y., Medhin, G., Sirgu, S., and Hanlon, C. (2024). Impact of depression on self-efficacy, illness perceptions and self-management among people with type 2 diabetes: A systematic review of longitudinal studies. Plos one, 19(5):e0302635.

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.

ENBIARE (2021). Encuesta nacional de bienestar autorreportado (enbiare), 2021. Instituto Nacional de Estadística y Geografía (INEGI). Collected June–July 2021. Data and documentation available from INEGI.

Eysenbach, G. (2005). The law of attrition. Journal of medical Internet research, 7(1):e402.

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.

Favaloro, P. and Berger, A. (2021). Technical updates to our global health and wellbeing cause prioritization framework. Coefficient Giving (formerly Open Philanthropy). Originally published under the name Open Philanthropy.

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.

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.

GBD (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.

Grossman, M. (1972). On the concept of health capital and the demand for health. Journal of Political economy, 80(2):223–255.

Gupta, S., Lehmann, D. R., and Stuart, J. A. (2004). Valuing customers. Journal of Marketing Research, 41(1):7–18.

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.

Haushofer, J., Mudida, R., and Shapiro, J. P. (2020). The comparative impact of cash transfers and a psychotherapy program on psychological and economic well-being. NBER Working Paper 28106, National Bureau of Economic Research.

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.

INEGI (2025). Encuesta nacional sobre disponibilidad y uso de tecnologías de la información en los hogares (ENDUTIH) 2024: Reporte de resultados. Technical Report 9/25, Instituto Nacional de Estadística y Geografía, México. Datos de cobertura temporal 2024; levantamiento del 10 de junio al 9 de agosto de 2024.

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.

Kemp, S. (2025). Digital 2026: Mexico. https://datareportal.com/reports/digital-2026-mexico. DataReportal, We Are Social, and Meltwater.

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.

Kuhn, E., Kanuri, N., Hoffman, J. E., Garvert, D. W., Ruzek, J. I., and Taylor, C. B. (2017). A randomized controlled trial of a smartphone app for posttraumatic stress disorder symptoms. Journal of consulting and clinical psychology, 85(3):267.

Kumar, V., Aksoy, L., Donkers, B., Venkatesan, R., Wiesel, T., and Tillmanns, S. (2010). Undervalued or overvalued customers: Capturing total customer engagement value. Journal of Service Research, 13(3):297–310.

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

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

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).

Lipschitz, J. M., Pike, C. K., Hogan, T. P., Murphy, S. A., and Burdick, K. E. (2023). The engagement problem: a review of engagement with digital mental health interventions and recommendations for a path forward: Lipschitz et al. Current treatment options in psychiatry, 10(3):119–135.

List, J. A. (2024). Optimally generate policy-based evidence before scaling. Nature, 626(7999):491–499.

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., and Herzberg, P. Y. (2008). Validation and standardization of the generalized anxiety disorder screener (gad-7) in the general population. Medical care, 46(3):266–274.

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.

Meffert, S. M., Neylan, T. C., McCulloch, C. E., Blum, K., Cohen, C. R., Bukusi, E. A., Verdeli, H., Markowitz, J. C., Kahn, J. G., Bukusi, D., Thirumurthy, H., Rota, G., Rota, R., Oketch, G., Opiyo, E., and Ongeri, L. (2021). Interpersonal psychotherapy delivered by nonspecialists for depression and posttraumatic stress disorder among kenyan hiv-positive women affected by gender-based violence: Randomized controlled trial. PLOS Medicine, 18(1):e1003468.

Mikulic, M. (2025). Digital health in mental healthcare – statistics and facts. Statista. Accessed March 25, 2026.

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.

Mosquera, R., Odunowo, M., McNamara, T., Guo, X., and Petrie, R. (2020). The economic effects of facebook. Experimental Economics, 23(2):575–602.

Nieuwenhuijsen, K., Verbeek, J. H., Neumeyer-Gromen, A., Verhoeven, A. C., Bültmann, U., and Faber, B. (2020). Interventions to improve return to work in depressed people. Cochrane Database of Systematic Reviews, (10).

OCHA (2016). Mexico – subnational administrative boundaries. Humanitarian Data Exchange (HDX). Sourced from Instituto Nacional de Estadística y Geografía (INEGI). Accessed: March 2026.

OECD (2021). OECD Digital Education Outlook 2021: Pushing the Frontiers with Artificial Intelligence, Blockchain and Robots. OECD Publishing, Paris.

Patel, N. M., Savaliya, G. V., Mehta, P. J., and Kataria, L. R. (2025). Global disparities in mental health systems: A comparative cross-sectional study of ten countries with different income levels. Indian Journal of Psychological Medicine, page 02537176251379999.

Patel, V., Weiss, H. A., Chowdhary, N., Naik, S., Pednekar, S., Chatterjee, S., De Silva, M. J., Bhat, B., Araya, R., King, M., et al. (2010). Effectiveness of an intervention led by lay health counsellors for depressive and anxiety disorders in primary care in goa, india (manas): a cluster randomised controlled trial. The Lancet, 376(9758):2086–2095.

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.

Salomon, J. A., Haagsma, J. A., Davis, A., Maertens de Noordhout, C., Polinder, S., Havelaar, A. H., Cassini, A., Devleesschauwer, B., Kretzschmar, M., Speybroeck, N., Murray, C. J. L., and Vos, T. (2015). Disability weights for the Global Burden of Disease 2013 study. The Lancet Global Health, 3(11):e712–e723.

Schnyder, N., Panczak, R., Groth, N., and Schultze-Lutter, F. (2017). Association between mental health-related stigma and active help-seeking: systematic review and meta-analysis. The British Journal of Psychiatry, 210(4):261–268.

Scott, A. J., Webb, T. L., Martyn-St James, M., Rowse, G., and Weich, S. (2021). Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials. Sleep medicine reviews, 60:101556.

Sikander, S., Ahmad, I., Atif, N., Zaidi, A., Vanobberghen, F., Weiss, H. A., Nisar, A., Tabana, H., Fuhr, D. C., Price, L. N., and Rahman, A. (2019). Delivering the thinking healthy programme for perinatal depression through volunteer peers: A cluster randomised controlled trial in pakistan. The Lancet Psychiatry, 6(2):128–139.

Singla, D. R., Kohrt, B. A., Murray, L. K., Anand, A., Chorpita, B. F., and Patel, V. (2017). Psychological treatments for the world: lessons from low-and middle-income countries. Annual review of clinical psychology, 13:149–181.

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.

Sweetman, J., Knapp, P., Varley, D., Woodhouse, R., McMillan, D., and Coventry, P. (2021). Barriers to attending initial psychological therapy service appointments for common mental health problems: A mixed-methods systematic review. Journal of affective disorders, 284:44–63.

Tahmassian, K. and Moghadam, N. J. (2011). Relationship between self-efficacy and symptoms of anxiety, depression, worry and social avoidance in a normal sample of students. Iranian journal of psychiatry and behavioral sciences, 5(2):91.

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.

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., Michalak, E. E., and O'Brien, H. L. (2020b). Digital health and engagement—looking behind the measures and methods. JAMA network open, 3(7):e2010918.

UNICEF (2021). Monitoring distance learning during school closures. Technical report, UNICEF Regional Office for South Asia, Kathmandu.

Unützer, J., Katon, W., Callahan, C. M., Williams Jr, J. W., Hunkeler, E., Harpole, L., Hoffing, M., Della Penna, R. D., Noël, P. H., Lin, E. H., et al. (2002). Collaborative care management of late-life depression in the primary care setting: a randomized controlled trial. Jama, 288(22):2836–2845.

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.

WHO (2016). Monitoring and Evaluating Digital Health Interventions: A Practical Guide to Conducting Research and Assessment. World Health Organization, Geneva.

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

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 (2025). World mental health today: latest data. Technical report, World Health Organization, Geneva. Accessed 2026-03-04.

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 51

B Recruitment 53

C Cognitive Tasks: Matrices and Emotional Stroop Games 54

D High-Frequency Affect Data 56

E Sample Characteristics 57

F Attrition and Balance 59

G Additional Impact Results 62

H Experimenter Demand Effects 66

I Expert Predictions 70

J App Engagement and Effect Persistence 71

K App Use and Mental Health Impacts 73

L Cost-Effectiveness 76

M Impacts from Extending App Access 77

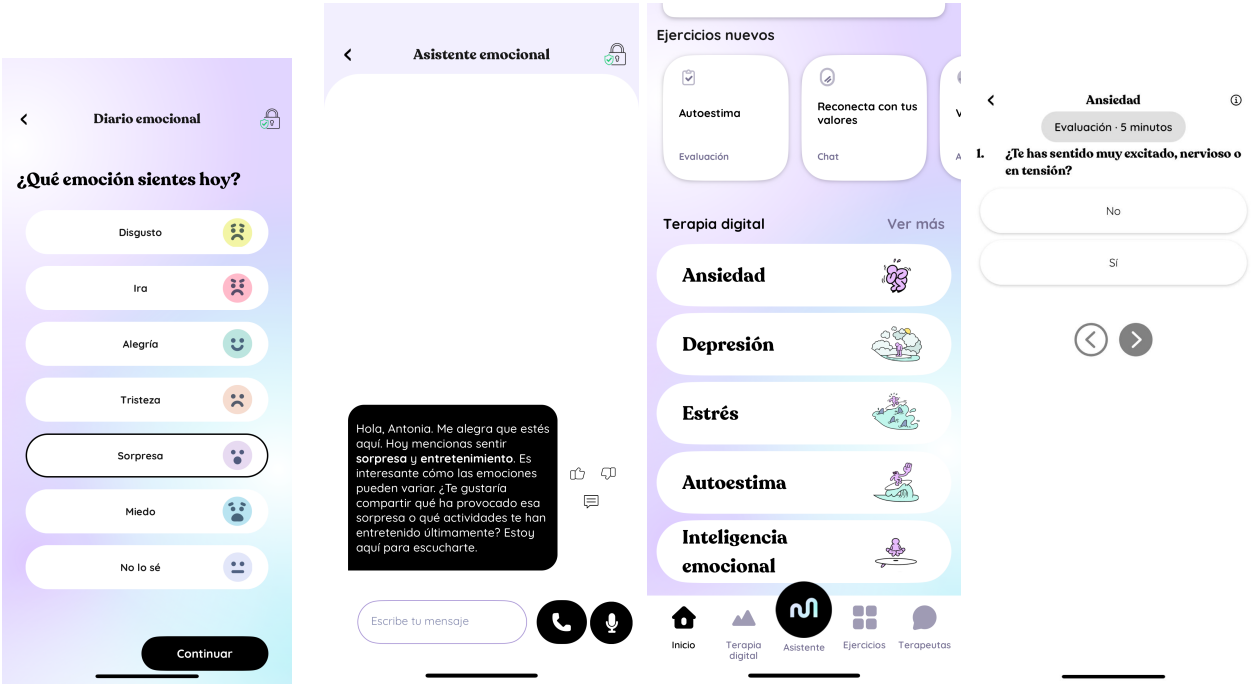
N Psychotherapy Use 79

O Supplemental material - Not for Publication 85

A The Mindsurf App

The four main app features

Figure A.1: App Interface



(a) Mood diary (b) AI agent (c) Exercise menu (d) Self Assessment

Notes: This figure presents screenshots showing the app interface. Panel (a) displays the mood diary with the question "¿How do you feel today?" and six options. Panel (b) shows the AI agent where users can engage in real-time dialogue with the AI assistant. Here 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."* Panel (c) 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. Panel (d) shows an example of the self-assessment tool for anxiety with a question to assess anxiety levels (*"Have you felt very excited, nervous, or tense?"*

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 body 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 material 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 material 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.

Safety Protocols

The app has multiple protocols to ensure the safety of its users. First, the AI assistant is 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 redirected 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.

Second, the model is trained to detect both self-harm and harm to others based on conversation content. In case of detected self-harm, the agent acknowledges the user's distress and encourages her to contact an emergency number, help lines, or a next-of-kin emergency contact provided upon registration, depending on the severity of the distress.

Third, human oversight supervises the model to detect errors, flagging and reviewing alarming content. 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 agent responds to harmful or unusual content. Detected errors are used to improve the accuracy of algorithmic detection.

B Recruitment

We recruited via social media advertisements targeted at 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

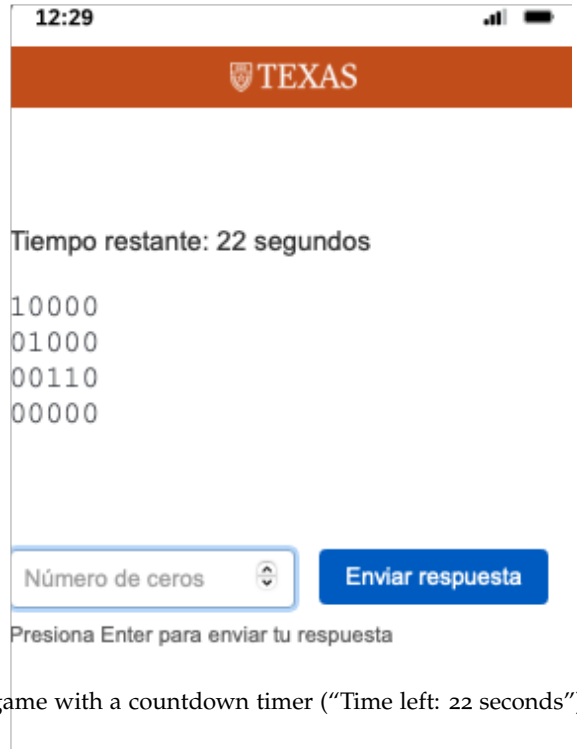
(c) Version 3

C Cognitive Tasks: Matrices and Emotional Stroop Games

Matrix Counting

To measure cognitive function, 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), as in Figure C.1. Participants had one minute to solve up to 10 matrices. The interface showed the remaining time and provided feedback on correct answers. Each correct answer received a ticket for a final prize draw among all participants. Our outcome is the correctly completed matrices. Higher scores indicate greater cognitive function.

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



Note: Participant's view of the game with a countdown timer ("Time left: 22 seconds").

Emotional Stroop Task

To assess attentional bias and cognitive control in the presence of emotionally salient stimuli, we used the emotional Stroop task (Williams et al., 1996). In this task, participants identified the color of emotionally charged or neutral words, presented in separate blocks. Longer reaction times to emotional words reflect poorer cognitive control.

Figure C.2 shows a neutral word, "Televisión" (television), in red text, requiring participants to select "rojo" (red) from the color options below and an emotional word, "Depresión" (depression), also requiring color identification. We adapted the task to Spanish using lexicographically similar words, and incentivized participants to complete as many trials as possible within a fixed time window by offering tickets for a final prize draw based on performance.

The task consisted of a neutral block, followed by an emotional block containing negative valence words, and by a final neutral block. Our outcome measures are the total response times in milliseconds for each block and overall. Longer times in the emotional block indicate emotional interference or attentional bias toward negative emotional content. Overall speed indicates reaction time, an indicator of cognitive function.

Figure C.2: Examples of Emotional Stroop Task Trials

(a) Neutral word (b) Emotional word

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

D High-Frequency 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 variable "Happy" (mean frequency of reported happiness) and our primary outcomes in the control group. The reasonably strong and statistically significant correlations ($p<0.01$) across these measures offer supportive evidence for the validity of the emoji-based check-in as a proxy for emotional well-being. The correlation between all outcomes becomes slightly stronger over time.

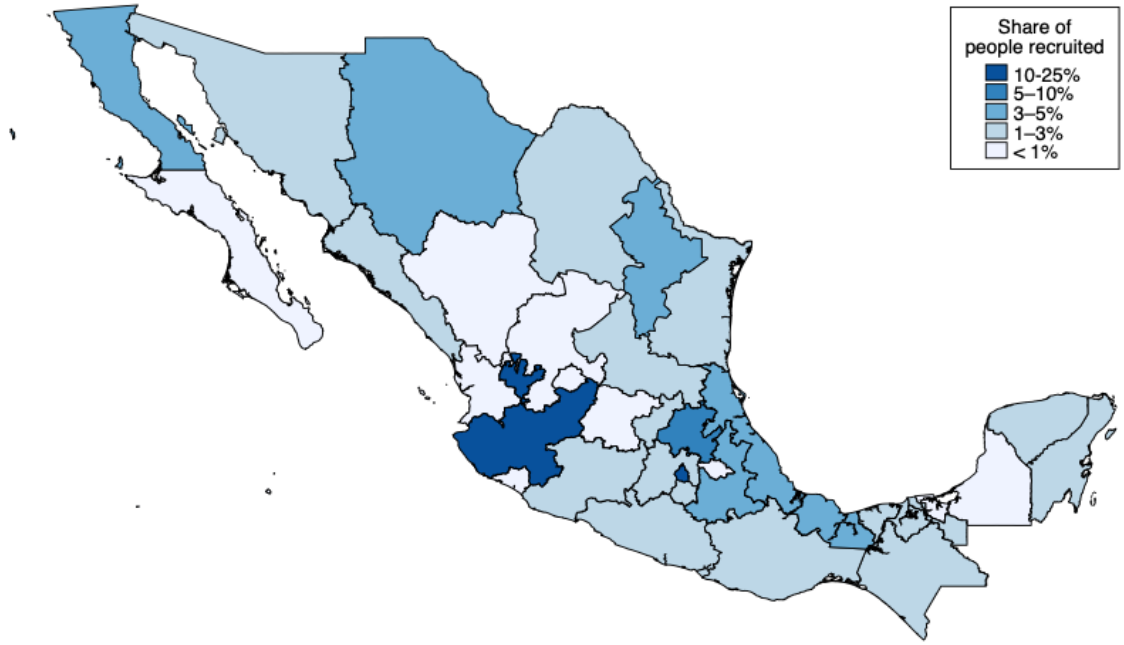
Table D.1: Correlation between Affect Mental Health Outcomes in the Control Group

	1 month					2 months					6 months				
	Happy	WHO-5	PHQ-8	GAD-7	PSS-4	Happy	WHO-5	PHQ-8	GAD-7	PSS-4	Happy	WHO-5	PHQ-8	GAD-7	PSS-4
Happy	1					1					1				
WHO-5	0.256	1				0.329	1				0.384	1			
PHQ-8	-0.253	-0.672	1			-0.324	-0.724	1			-0.349	-0.728	1		
GAD-7	-0.251	-0.579	0.718	1		-0.287	-0.598	0.738	1		-0.304	-0.594	0.753	1	
PSS	-0.216	-0.584	0.619	0.578	1	-0.274	-0.627	0.657	0.649	1	-0.324	-0.636	0.661	0.637	1

Note: The table shows correlations in the control group at 1, 2, and 6 months. These correlations are all statistically significant at the 99% confidence level ($p<0.01$). The "happy" variable is the mean frequency of reported happiness in the weeks of the corresponding survey and the previous week. That is, "happy" at one month is the average number of times that a person reported being happy in weeks 3 and 4 (this mean can be 0, 0.5, or 1). Similarly, "happy" at 2 (6) months is the average number of times that a person reported being happy in weeks 7 and 8 (25 and 26). We do this to harmonize event timing: the four mental health outcomes have a two-week recall period. Thus, the PHQ-8 at one month shows people's symptoms of depression during weeks 3 and 4.

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 share of women recruited from each state. We obtain waw map files from OCHA (2016).

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)
Below median income	0.35	0.45	0.45	0.45	0.42
	(0.48)	(0.50)	(0.50)	(0.50)	(0.49)
Some tertiary education	0.72	0.63	0.63	0.63	0.64
	(0.45)	(0.48)	(0.48)	(0.48)	(0.48)
Depression/Anxiety (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)
Stress (PSS-4)	-	-	10.02	10.02	10.04
			(2.47)	(2.48)	(2.46)
Well-being (WHO-5)	-	-	31.10	30.96	31.13
			(16.31)	(16.13)	(15.95)
Self-efficacy (GSE)	-	-	15.41	15.41	15.43
			(3.30)	(3.25)	(3.26)
Locus of control	-	-	11.85	11.78	11.72
			(2.79)	(2.75)	(2.76)
Visited psych.	-	-	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 Attrition and Balance

Table F.1: Survey Completion Rates by Treatment

	1 month (1)	2 months (2)	6 months (3)	All 3 surveys (4)
Treated	-0.014	-0.017	-0.002	0.003
	(0.010)	(0.011)	(0.017)	(0.018)
Mean control	0.95	0.94	0.82	0.79
Observations	1963	1963	1963	1963

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

Table F.2: Balance Table Across Survey Samples

Variable	Randomized		1 month		2 months		6 months		All surveys	
	Control	Δ Treated	Control	Δ Treated	Control	Δ Treated	Control	Δ Treated	Control	Δ Treated
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Age	38.00	0.05	38.07	0.07	37.96	0.10	37.86	0.23	37.90	0.21
	(7.78)	(0.22)	(7.70)	(0.23)	(7.72)	(0.23)	(7.81)	(0.25)	(7.77)	(0.25)
Below median income	0.45	-0.00	0.44	-0.00	0.43	0.01	0.43	0.00	0.42	-0.00
	(0.50)	(0.02)	(0.50)	(0.02)	(0.50)	(0.02)	(0.49)	(0.02)	(0.49)	(0.02)
Works	0.48	0.01	0.48	0.01	0.48	-0.00	0.50	-0.01	0.50	-0.01
	(0.50)	(0.02)	(0.50)	(0.02)	(0.50)	(0.02)	(0.50)	(0.02)	(0.50)	(0.03)
Number of children	0.93	0.05	0.93	0.05	0.93	0.05	0.92	0.05	0.92	0.04
	(1.01)	(0.04)	(1.01)	(0.04)	(1.01)	(0.05)	(1.02)	(0.05)	(1.03)	(0.05)
Single	0.32	-0.03	0.32	-0.03	0.32	-0.03	0.33	-0.03	0.33	-0.04
	(0.47)	(0.02)	(0.47)	(0.02)	(0.47)	(0.02)	(0.47)	(0.02)	(0.47)	(0.02)
Some tertiary education	0.63	0.01	0.63	0.02	0.64	0.01	0.63	0.01	0.63	0.02
	(0.48)	(0.02)	(0.48)	(0.02)	(0.48)	(0.02)	(0.48)	(0.02)	(0.48)	(0.02)
Depression/Anxiety (PHQ-4)	6.57	0.07	6.54	0.10	6.57	0.10	6.52	0.13**	6.49	0.13*
	(2.61)	(0.06)	(2.59)	(0.06)	(2.59)	(0.06)	(2.58)	(0.07)	(2.57)	(0.07)
Stress (PSS-4)	10.05	-0.06	10.02	0.02	10.06	-0.01	10.04	-0.00	10.02	0.04
	(2.47)	(0.10)	(2.47)	(0.11)	(2.47)	(0.11)	(2.45)	(0.11)	(2.44)	(0.11)
Well-being (WHO-5)	30.92	0.06	31.05	-0.17	30.80	0.03	31.24	-0.18	31.23	-0.20
	(15.99)	(0.67)	(16.02)	(0.68)	(16.01)	(0.68)	(16.01)	(0.74)	(15.95)	(0.74)
Self-efficacy (GSE)	15.50	-0.19	15.54	-0.25*	15.49	-0.23	15.61	-0.38**	15.63	-0.41***
	(3.25)	(0.14)	(3.23)	(0.15)	(3.24)	(0.15)	(3.24)	(0.16)	(3.23)	(0.16)
Locus of control	11.79	-0.03	11.78	-0.04	11.76	-0.04	11.79	-0.11	11.80	-0.15
	(2.70)	(0.12)	(2.70)	(0.13)	(2.72)	(0.13)	(2.72)	(0.14)	(2.72)	(0.14)
Extroversion (Big5)	8.37	-0.15	8.37	-0.16	8.32	-0.11	8.36	-0.14	8.36	-0.16
	(2.53)	(0.11)	(2.54)	(0.11)	(2.53)	(0.12)	(2.55)	(0.12)	(2.55)	(0.13)
Agreeableness (Big5)	12.21	0.05	12.21	0.02	12.20	0.05	12.19	0.04	12.18	0.04
	(1.96)	(0.09)	(1.95)	(0.09)	(1.96)	(0.09)	(1.97)	(0.10)	(1.97)	(0.10)
Conscientiousness (Big5)	10.34	0.05	10.31	0.06	10.29	0.11	10.29	0.10	10.30	0.06
	(2.80)	(0.12)	(2.79)	(0.13)	(2.82)	(0.13)	(2.78)	(0.14)	(2.78)	(0.14)
Neuroticism (Big 5)	11.08	0.10	11.08	0.11	11.13	0.08	11.11	0.19*	11.12	0.18
	(2.36)	(0.10)	(2.35)	(0.10)	(2.34)	(0.10)	(2.32)	(0.11)	(2.31)	(0.11)
Openness to Experience (Big5)	10.83	-0.07	10.82	-0.07	10.80	-0.06	10.82	-0.06	10.82	-0.07
	(2.20)	(0.09)	(2.20)	(0.10)	(2.23)	(0.10)	(2.27)	(0.11)	(2.27)	(0.11)
Visited psych.	0.31	0.01	0.31	0.01	0.31	0.00	0.30	-0.00	0.30	0.00
	(0.46)	(0.02)	(0.46)	(0.02)	(0.46)	(0.02)	(0.46)	(0.02)	(0.46)	(0.02)
Observations	1,964		1,856		1,830		1,603		1,562	
Joint F-test (p-value)	0.728		0.689		0.737		0.193		0.222	

Note: The table reports baseline characteristics by treatment arm across subsamples defined by the corresponding survey completion. Odd columns report the control group mean with standard deviations in parentheses. Even columns report the coefficient on the treatment indicator from a regression controlling for stratification fixed effects, with robust standard errors in parentheses. The joint F-test p-value is from a regression of the treatment indicator on all balance variables with stratification fixed effects and robust standard errors. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table F.3: Predictors of Attrition in the Affect Data

	(1) Covariates	(2) Interaction with treatment
Treatment	0.102	
	(0.157)	
Below median income	0.048***	-0.057**
	(0.015)	(0.023)
Works	0.003	0.001
	(0.015)	(0.023)
Number of children	0.006	0.013
	(0.008)	(0.013)
Single	0.001	0.011
	(0.017)	(0.027)
Some tertiary education	-0.026*	-0.029
	(0.015)	(0.025)
Well-being (WHO-5)	-0.000	-0.001
	(0.001)	(0.001)
Stress (PSS-4)	0.001	-0.003
	(0.004)	(0.006)
Self-efficacy (GSE)	-0.007**	0.013***
	(0.003)	(0.004)
Locus of control	0.005*	-0.000
	(0.003)	(0.005)
Extroversion (Big5)	-0.003	-0.001
	(0.003)	(0.005)
Agreeableness (Big5)	-0.004	0.005
	(0.004)	(0.006)
Conscientiousness (Big5)	0.004	-0.009*
	(0.003)	(0.005)
Neuroticism (Big 5)	-0.004	-0.003
	(0.004)	(0.006)
Openness to Experience (Big5)	0.007**	-0.007
	(0.004)	(0.006)
Visited psych.	-0.001	0.039
	(0.016)	(0.025)
Joint F-test, p-value	0.012	0.037
Joint F-test all, p-value	0.000	

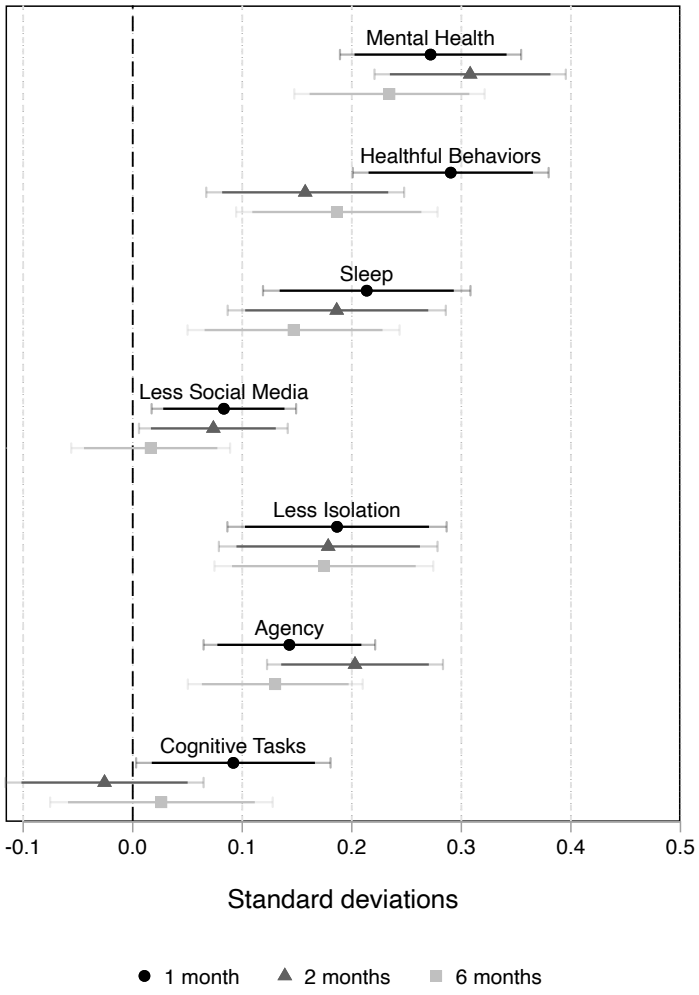
Note: Unconditional rates of attrition are 9 percentage points higher in the treatment group ($p < 0.10$). The table reports OLS estimates from the regression:

$$\mathrm{A}_{it} = \alpha + \beta T_i + \gamma X_i + \delta(T_i \times X_i) + \mu_s + \varepsilon_i$$

where $\mathrm{A}_{it} = 1$ if participant i did not respond in week t, and 0 otherwise. T_i is an indicator for treatment assignment. X_i are baseline covariates. μ_s are strata fixed effects. Column (1) reports $\hat{\gamma}$ and Column (2) reports $\hat{\delta}$. Standard errors are clustered by person. 50,908 observations and 1,958 clusters. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

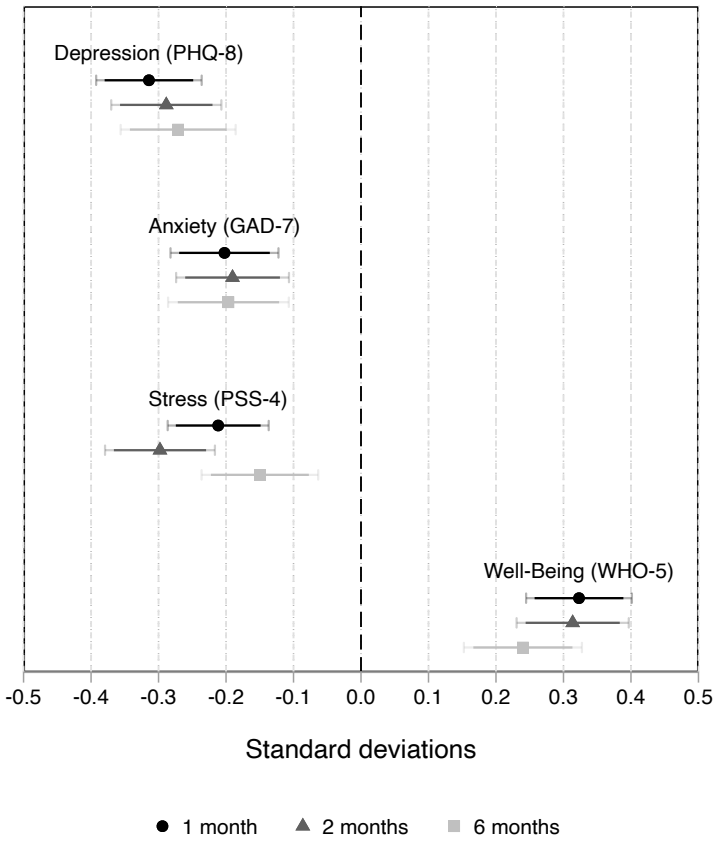
G Additional Impact Results

Figure G.1: Impacts on Indices for the Fixed Sample Across Survey Rounds



Notes: We restrict the sample to participants with data at 1, 2, and 6 months, excluding those who do not respond to at least one of the three follow-up surveys. 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 *Behavioral Index* summarizes the days in the past week during which 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 *Sleep Index* aggregates hours of sleep, number of interruptions (reversed), and sleep difficulties (reversed). The *Less Social Media Index* captures platform use (Facebook, X/Twitter, Instagram, TikTok) and frequent use (more than two hours per day, reversed). The *Less Isolation Index* is based on the six-item De Jong Gierveld Loneliness Scale (reversed), combining the social and emotional subscales. The *Agency Index* combines scores from the General Self-Efficacy (GSE) Scale and locus of control measures. The *Cognitive Tasks index* combines performance in two incentivized tasks measuring cognitive function and effort. All regressions control for randomization strata and LASSO-selected baseline covariates. Circles, triangles, and squares denote 1, 2, and 6-month estimates. Thick and thin bars indicate 90% and 95% confidence intervals. Robust standard errors, clustered by individual in the pooled data.

Figure G.2: Standardized Mental Health Impacts



Notes: Each point reports the estimated treatment effect on a standardized mental health outcome. Outcomes are standardized around the control group mean. *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. *Depression (PHQ-8)* is the eight-item Patient Health Questionnaire. *Anxiety (GAD-7)* is the seven-item Generalized Anxiety Disorder scale. *Stress (PSS-4)* is the four-item Perceived Stress Scale. The *Well-Being (WHO-5)* score is the five-item World Health Organization Well-Being Index. Thick and thin bars indicate 90% and 95% confidence intervals. All regressions control for randomization strata and lasso-selected baseline covariates and adjust for robust standard errors.

Table G.1: Impacts on the Likelihood of Severe Depression and Anxiety

	Severe Depression (1) [−]	Severe Anxiety (2) [−]
Panel A. Pooled Effects		
Treated	-0.027***	-0.023**
	(0.008)	(0.010)
Mean control	0.10	0.16
Observations	5272	5274
Panel B. Effects at 1 month		
Treated	-0.032**	-0.021
	(0.013)	(0.016)
Mean control	0.11	0.17
Observations	1849	1850
Panel C. Effects at 2 months		
Treated	-0.015	-0.020
	(0.013)	(0.016)
Mean control	0.10	0.16
Observations	1824	1825
Panel D. Effects at 6 months		
Treated	-0.030**	-0.036**
	(0.013)	(0.016)
Mean control	0.09	0.14
Observations	1599	1599
1 mo=2 mos (*p-val*)	0.37	0.97
2 mos=6 mos (*p-val*)	0.41	0.58
1 mo=6 mos (*p-val*)	0.92	0.61
1 mo=2 mos=6 mos (*p-val*)	0.60	0.80

Notes: Each cell reports the estimated treatment effect on the likelihood of being severely depressed or anxious, where severe depression is defined as PHQ-8 $\geq$ 20 and severe anxiety as GAD-7 $\geq$ 15. Regressions control for strata and lasso-selected baseline covariates. Robust standard errors in parenthesis, clustered by individual in the pooled data. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table G.2: Local Average Treatment Effects on Index Outcomes

	Mental Health	Behavior	Sleep	Less Social Media	Less Isolation	Agency
	(1)	(2)	(3)	(4)	(5)	(6)
	[+]	[+]	[+]	[+]	[+]	[+]
Panel A. Pooled Effects						
Downloaded app	0.327***	0.256***	0.217***	0.070**	0.229***	0.204***
	(0.037)	(0.036)	(0.039)	(0.029)	(0.044)	(0.034)
Mean control	0.00	0.00	0.00	0.00	0.00	0.00
First stage coef.	.894***	.894***	.894***	.894***	.894***	.894***
First stage *F*	8162.6	8150.2	8164.1	8162.8	8164.1	8164.1
Observations	5274	5274	5274	5274	5274	5274
Panel B. Effects at 1 month						
Downloaded app	0.351***	0.370***	0.260***	0.099***	0.262***	0.186***
	(0.042)	(0.047)	(0.050)	(0.035)	(0.052)	(0.041)
Mean control	0.00	0.00	0.00	0.00	0.00	0.00
First stage coef.	.886***	.886***	.886***	.886***	.886***	.886***
First stage *F*	7282.1	7282.1	7282.6	7282.9	7282.6	7282.6
Observations	1850	1850	1850	1850	1850	1850
Panel C. Effects at 2 months						
Downloaded app	0.356***	0.191***	0.210***	0.079**	0.229***	0.268***
	(0.046)	(0.047)	(0.051)	(0.036)	(0.052)	(0.042)
Mean control	0.00	0.00	0.00	0.00	0.00	0.00
First stage coef.	.893***	.893***	.893***	.893***	.893***	.893***
First stage *F*	7698.0	7683.1	7700.1	7707.1	7700.1	7700.1
Observations	1825	1825	1825	1825	1825	1825
Panel D. Effects at 6 months						
Downloaded app	0.267***	0.204***	0.179***	0.027	0.194***	0.151***
	(0.048)	(0.051)	(0.054)	(0.040)	(0.055)	(0.044)
Mean control	0.00	0.00	0.00	0.00	0.00	0.00
First stage coef.	.906***	.906***	.906***	.906***	.906***	.906***
First stage *F*	7782.0	7767.5	7781.4	7768.4	7781.4	7781.4
Observations	1599	1599	1599	1599	1599	1599
1 mo=2 mos (*p-val*)	0.75	0.00	0.49	0.67	0.39	0.05
2 mos=6 mos (*p-val*)	0.05	0.88	0.53	0.20	0.55	0.01
1 mo=6 mos (*p-val*)	0.12	0.00	0.20	0.11	0.20	0.58
1 mo=2 mos=6 mos (*p-val*)	0.14	0.00	0.44	0.27	0.43	0.03

Notes: Estimates of LATE on each outcome index, using random assignment as an instrument for app download. [+] indicates that higher values represent better outcomes; [−] indicates that higher values represent worse outcomes. Regressions control for strata and lasso-selected baseline covariates. Robust standard errors in parenthesis, clustered by individual in the pooled data. * $p < .10$, ** $p < .05$, *** $p < .01$.

H 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, 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, we exploit baseline variation in social desirability bias measured at enrollment using a short version of the Marlowe–Crowne Social Desirability Scale (Crowne and Marlowe, 1960). We construct an index of social desirability and classify participants as having high social desirability if their score exceeds the sample median. Following **?**, who propose using heterogeneity in susceptibility to demand effects as a robustness check in studies with self-reported outcomes, we estimate treatment effects separately above and below the median of this baseline distribution. If experimenter demand were driving our results, we would expect larger treatment effects among participants with higher baseline social desirability. Table H.1 shows no consistent differences in treatment effects across groups. If anything, the interaction estimates suggest that participants with higher social desirability traits tend to display somewhat smaller treatment effects, which goes against the concern that our results are driven by over-reporting.

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 H.2).

Finally, we directly elicit participants' beliefs about the study's purpose. In the final survey, respondents were asked an open-ended question about what they believed the research aimed to examine. While the majority believed the research study was aimed at understanding well-being,

Table H.1: Impacts Heterogeneity by Social Desirability Traits

	Mental Health (1)	Healthful Behaviors (2)	Sleep (3)	Less Social Media (4)	Less Isolation (5)	Agency (6)	Visit Psychologist (7)
Panel A. Pooled							
Treated	0.323***	0.252***	0.222***	0.070***	0.203***	0.208***	0.052***
	(0.029)	(0.031)	(0.033)	(0.024)	(0.033)	(0.028)	(0.012)
Treated × High SDB	-0.104*	-0.094	-0.116*	-0.035	-0.032	-0.080	0.001
	(0.057)	(0.059)	(0.062)	(0.043)	(0.069)	(0.052)	(0.022)
Observations	5274	5274	5274	5274	5274	5274	5273
Panel B. Effects at 1 month							
Treated	0.321***	0.363***	0.247***	0.107***	0.207***	0.169***	0.018
	(0.043)	(0.049)	(0.053)	(0.037)	(0.053)	(0.043)	(0.019)
Treated × High SDB	-0.057	-0.134	-0.100	-0.099	0.049	-0.026	0.048
	(0.087)	(0.096)	(0.098)	(0.068)	(0.110)	(0.083)	(0.035)
Observations	1850	1850	1850	1850	1850	1850	1850
Panel C. Effects at 2 months							
Treated	0.349***	0.173***	0.229***	0.074*	0.190***	0.275***	0.060***
	(0.048)	(0.051)	(0.053)	(0.038)	(0.053)	(0.045)	(0.020)
Treated × High SDB	-0.123	-0.015	-0.163	-0.001	0.005	-0.109	-0.022
	(0.095)	(0.094)	(0.106)	(0.071)	(0.113)	(0.083)	(0.037)
Observations	1825	1825	1825	1825	1825	1825	1824
Panel D. Effects at 6 months							
Treated	0.265***	0.235***	0.166***	0.024	0.206***	0.170***	0.082***
	(0.051)	(0.054)	(0.059)	(0.044)	(0.058)	(0.048)	(0.022)
Treated × High SDB	-0.095	-0.178*	-0.064	-0.007	-0.149	-0.093	-0.020
	(0.097)	(0.103)	(0.106)	(0.080)	(0.116)	(0.089)	(0.040)
Observations	1599	1599	1599	1599	1599	1599	1599

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. Standard error clustered at the participant level in Panel A. Robust standard errors in Panels B, C, and D. * $p < .10$, ** $p < .05$, *** $p < .01$.

only 24% articulated specific hypotheses regarding treatment effects or mental health impacts. 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 about sleep interruptions (the number of times respondents woke up and remained awake for at least 5 minutes the previous night), and self-reported visits to a psychiatrist or psychologist in the past 4 weeks. We find little evidence that respondents adjusted their answers to align with stated researcher expectations (Table H.3).

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

	Value of Own Contribution (1)
Panel A. Effects at 2 months	
Treated	-0.026
	(0.090)
Mean control	8.48
Observations	1823
Panel B. Effects at 6 months	
Treated	-0.027
	(0.095)
Mean control	8.46
Observations	1596

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 2-month follow-up, and Panel B presents effects measured at the 6-month follow-up (this information was not collected at the 1-month follow-up). Regressions include strata fixed effects and a set of lasso-selected baseline covariates. Robust standard errors in parentheses. * $p < .10$, ** $p < .05$, *** $p < .01$.

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.

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

Table H.3: Experiment on Hypothesis Disclosure

	Anx./Dep. (PHQ-4) (1)	Depression (PHQ-4) (2)	Anxiety (PHQ-4) (3)	Num. Woke Up (4)	Num. visits psych (5)
Treated x Prompt	0.173	0.087	0.102	0.158	-0.464*
	(0.341)	(0.194)	(0.185)	(0.245)	(0.251)
Prompt	-0.277	-0.182	-0.121	0.167	0.332*
	(0.235)	(0.136)	(0.130)	(0.171)	(0.202)
Mean control	4.71	2.51	2.19	1.92	0.65
Observations	1599	1599	1599	1585	1507

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, after which 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 5 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. Robust standard errors. * $p < .10$, ** $p < .05$, *** $p < .01$.

I Expert Predictions

Table I.1: Predictions and Estimated Impacts

	Non-experts	Experts	Confident experts	Study impacts
Panel A. Continuous outcomes (medians)				
App use (% used at least once/week in wks 1–3)	55.00	61.00	71.00	54.53
App use (% used in week 8)	30.00	33.00	37.00	36.32
ITT: Mental health index (SD), 3 weeks	0.12	0.13	0.18	0.30
ITT: Mental health index (SD), 8 weeks	0.11	0.12	0.12	0.31
ITT: Healthful habits index (SD), 3 weeks	0.10	0.12	0.13	0.33
ITT: Without at least moderate depression (pp), 3 wks	2.30	4.30	6.60	13.44
ITT: Slept $\geq$7 hours last night (pp), 3 weeks	2.40	3.20	6.90	5.60
Panel B. Correctly guessed sign of impact (%)				
Psychotherapy use	24.75	40.00	40.00	0.03
Social media use	40.59	44.00	40.00	0.09
Social isolation	59.41	60.00	46.67	0.23
N (predictions)	101	25	15	–

Notes: The first three columns 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. The last column reports ITT estimates from double-selection LASSO regressions, controlling for strata and baseline covariates. For Panel A, app use outcomes are the percentages of treated participants who meet each use threshold. Depression and sleep entries are in percentage points, obtained by regressing binary indicators on treatment. Panel B reports 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). In column 4, the psychotherapy use estimate is the treatment coefficient on a binary indicator (proportion); the social media and social isolation estimates are in standard deviations, with higher values corresponding to *less* social media use and *lower* social isolation.

We compare expert forecasts with the realized impacts to assess how well experts anticipate the intervention's effects. All forecasts understate the magnitude of the estimated impacts. Experts (confident experts) predict increases in mental health of 0.13 (0.18) SD at 3 weeks and 0.12 (0.12) SD at 8 weeks, whereas the 1- and 2-month average effects are approximately 0.30 SD. Similarly, they predict a reduction in depression prevalence of 4.3 (6.6) pp, whereas the estimated reduction is 13.4 pp. A similar pattern holds for healthful habits, with forecasts of a 0.12 (0.13) SD improvement at 3 weeks, compared to an estimated effect of 0.33 SD. Conversely, expert predictions on the frequency of sleep improvements (3.2 and 6.9 pp) are closer to the estimated changes (5.6 pp). Lastly, most people do *not* correctly predict the sign of observed impacts on psychotherapy use and social media use, and only 60% (47%) of experts (confident experts) correctly predict reductions in social isolation.

J App Engagement and Effect Persistence

Impacts on App-recommended CBT Tool Use

Table J.1: Impacts on Using CBT Tools Recommended by App in the Previous Week

	Kinder to self	Set boundaries	Breathing exercises	Sleep routine	Wrote about feelings and thoughts	Tools Index (std)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Pooled Effects (2 and 6 months)						
Treated	0.392***	0.323***	0.465***	0.413***	0.567***	0.346***
	(0.073)	(0.068)	(0.081)	(0.084)	(0.064)	(0.036)
Mean control	2.69	2.28	2.58	1.85	0.88	0.00
Observations	3424	3424	3424	3424	3424	3424
Panel B. Effects at 2 months						
Treated	0.390***	0.310***	0.501***	0.423***	0.841***	0.424***
	(0.095)	(0.092)	(0.109)	(0.110)	(0.088)	(0.050)
Mean control	2.71	2.30	2.57	1.76	0.85	0.00
Observations	1825	1825	1825	1825	1825	1825
Panel C. Effects at 6 months						
Treated	0.419***	0.347***	0.438***	0.423***	0.262***	0.270***
	(0.107)	(0.097)	(0.117)	(0.124)	(0.088)	(0.048)
Mean control	2.67	2.25	2.59	1.96	0.91	0.00
Observations	1599	1599	1599	1599	1599	1599
2 mos=6 mos (*p-val*)	0.84	0.93	0.60	0.93	0.00	0.02

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. All regressions include lasso-selected covariates and strata fixed effects. Robust standard errors in parentheses, and clustered by individual for the pooled effects. Statistical significance is denoted by $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Sleep and Behaviors Mediate Mental Health Impacts

We conduct a mediation analysis to quantify the extent to which improvements in sleep and healthful behaviors account for the treatment effect on mental health. We construct a standardized mediator index that combines relevant sleep and behavioral outcomes, following Anderson (2008), and estimate the average Natural Indirect Effect (NIE), the Natural Direct Effect (NDE), and the share NIE/ITT, where $NIE + NDE = ITT$. Identification relies on standard assump-

tions of cross-world independence and common support (Imai et al., 2010a,b).[16] Improvements in sleep and healthful behaviors account for 48%, 34%, and 43% of the average ITT effect at one, two, and six months.

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

	Natural Indirect Effect (NIE)	Natural Direct Effect (NDE)	Total Effect (TE) (NIE+NDE)	Proportion Mediated (NIE/TE)	Observations
1 month	0.148^{***}	0.164^{***}	0.312^{***}	0.475^{***}	1,850
	(0.021)	(0.034)	(0.038)	(0.067)	
2 months	0.107^{***}	0.211^{***}	0.318^{***}	0.338^{***}	1,825
	(0.023)	(0.036)	(0.041)	(0.064)	
6 months	0.103^{***}	0.139^{***}	0.242^{***}	0.425^{***}	1,599
	(0.025)	(0.038)	(0.043)	(0.094)	

Notes: This table reports estimates from a mediation analysis with a binary treatment. The outcome is a standardized mental health index measured at 1, 2, and 6 months. The mediator is a standardized composite index constructed in a single step following Anderson (2008), aggregating all individual sleep and healthful behavior variables directly. All models condition on baseline covariates and strata as specified in the main specifications. Both the outcome and mediator equations are estimated using linear regression. The outcome equation includes a treatment–mediator interaction. 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. Robust standard errors in parentheses.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Correlations between Mental Health Impacts and Sleep and Healthful Behavior Impacts

We estimate Conditional Average Treatment Effects (CATEs) for mental health and study how they correlate with app use, sleep, and healthful behavior CATEs (Athey and Wager, 2019, Chernozhukov et al., 2025). The estimates are stable when using different baseline covariates and number of folds in the cross-fitting algorithm. Mental health CATEs range from 0 to 0.5 SD, behavioral CATEs from 0 to 0.5 SD, and sleep CATEs from –0.05 to 0.4 SD. That mental health effects remain predominantly positive even among non-active users is consistent with eventual decoupling between use and benefits. Table J.3 shows that app use in the first month positively correlates with mental health impacts, but this relationship is clinically insignificant: people who used the app for 151 minutes in the first month (the 75% percentile of use) have a mental health

[16]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.

CATE only 0.01 SD bigger than people who used the app only for 20 minutes (the 25% percentile of use). Conversely, people with bigger sleep and behavioral CATEs also have bigger mental health impacts at 1, 2, and 6 months. For example, people with a 0.10 SD bigger sleep CATE also have a 0.03 SD bigger mental health CATE at one month.

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

	Mental Health Index CATEs					
	1 month		2 months		6 months	
	(1)	(2)	(3)	(4)	(5)	(6)
Minutes (100s, month 1)	0.009*	0.005	0.013	0.014	0.020**	0.008
	(0.004)	(0.003)	(0.009)	(0.010)	(0.011)	(0.007)
Minutes (100s, month 2)			-0.011	-0.014	-0.006	0.001
			(0.013)	(0.013)	(0.015)	(0.010)
Minutes (100s, months 3–6)					-0.054*	-0.011
					(0.030)	(0.015)
Behavioral Index CATE		0.274***		-0.041		0.663***
		(0.025)		(0.055)		(0.037)
Sleep Index CATE		0.339***		0.120***		0.695***
		(0.020)		(0.045)		(0.030)
Observations	922	922	907	907	801	801

Notes. Each column reports a linear projection of estimated Conditional Average Treatment Effects (CATEs) for the mental health index at the indicated horizon, where CATEs are estimated at the individual level (IATEs) using causal forests. Odd columns regress mental health CATEs on weekly app-use minutes only. Even columns add the corresponding CATEs for the behavioral and sleep indices. Minutes are averaged per week within each period and rescaled to units of 100 minutes. The sample is restricted to treated individuals. Robust standard errors in parentheses.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

K App Use and Mental Health Impacts

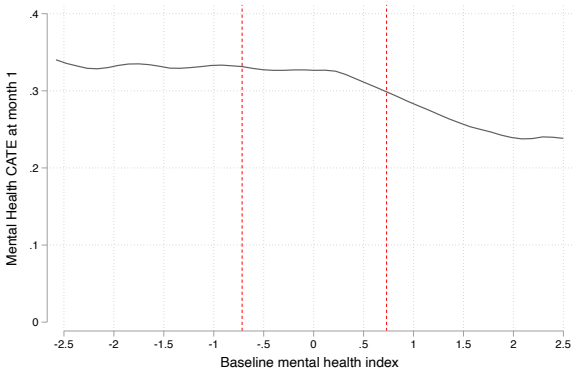
A priori, it is unclear how baseline psychological distress correlates with impacts. This is because it is difficult to predict both who will use the app the most and who will experience the largest mental health impacts conditional on use. For example, people with the most severe baseline mental needs mechanically have the biggest scope for benefiting from the app. However, they likely also face the highest behavioral barriers to use. At the causal level, the features of users who benefit the most from a mental health care app are not well-established.

We estimate the relationship between baseline mental health and both mental health CATEs and app use non-parametrically. Panels a-c of Figure K.1 show that people with the highest baseline distress have the largest mental health impacts from app access, especially at 2 and 6

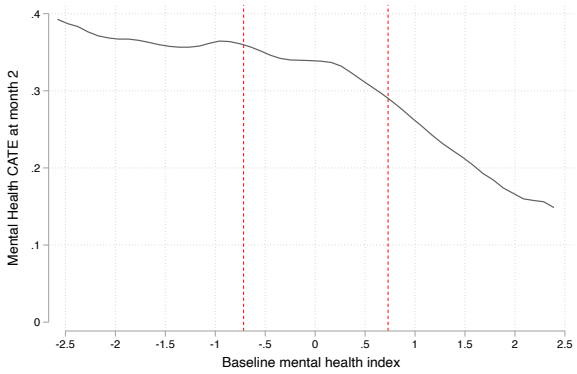
months. A meta-analysis finds similar effects of in-person care (Bower et al., 2013). Panel d shows that app use varies non-monotonically from 100 to 230 total minutes with respect to baseline mental health: use is lowest for people in the top quartile of the mental health index, who have the least need 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).

Overall, these patterns suggest that the app is effective at engaging users with the highest needs, who also experience the largest mental health improvements. These findings also reiterate our conclusion that there is no strong dose-response relationship between product use and mental health impacts: the people with the largest impacts are not the most active users.

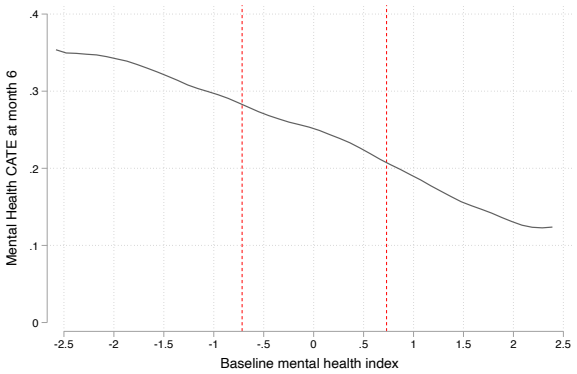
Figure K.1: Mental Health CATEs and App Use by Baseline Mental Health



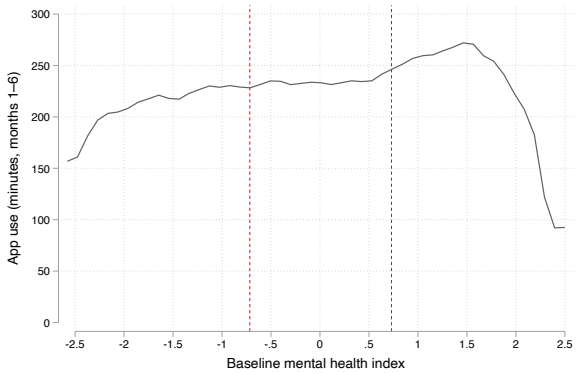
(a) Mental Health CATE at 1 Month



(b) Mental Health CATE at 2 Months



(c) Mental Health CATE at 6 Months



(d) App Use (Total Minutes at 6 Months)

Notes: Subfigures a, b, and c show the non-parametric relationship between Mental Health CATEs estimated at 1, 2, and 6 months and baseline mental health (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. Subfigure d shows the non-parametric relationship between app use (total minutes over 6 months) and baseline mental health (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 dashed red lines mark the inter-quartile range of this index. Higher Index values mean better baseline mental health.

L Cost-Effectiveness

Table L.1: Cost per SD at a Fixed Horizon and Cost per SD-month in Low- and Middle-Income Countries (LMICs)

Study / arm	(1) Cost (USD)	(2) $t^\star$ (mo.)	(3) Effect (SD)	(4) $/SD	(5) AUC (mo.)	(6) $/ 0.1SD-mo.	(7) Source
AI-powered Digital Apps in LMICs							
Digital App, *Mindsurf* (Mexico)	6	6	0.24	26	0–6	0.37	This paper
Psychotherapy / psychosocial support in LMICs							
HAP (India)	66	12	0.23	285	0–12	1.67	Patel et al. (2017)
HAP (India)	66	60	0.23	287	0–60	0.44	Bhat et al. (2022)
Thinking Healthy Program (Pakistan)	10	84	0.18	56	0–84	0.03	Baranov et al. (2020)
Thinking Healthy Program (Pakistan)	8.88	6	0.13	68	0–6	0.81	Sikander et al. (2019)
Thinking Healthy Program (Pakistan)	6.56	3	0.30	22	0–3	1.46	Sikander et al. (2019)
PM+ psychotherapy, Kenya	1189	12	-0.01	—	0–12	—	Haushofer et al. (2020)
Group IPT (Kenya)	36	3	0.74	49	0–3	3.24	Meffert et al. (2021)
Pharmacotherapy in LMICs							
Pharmacotherapy + livelihoods (India)	232	26	0.24	967	0–26	4.19	Angelucci and Bennett (2024)
Pharmacotherapy only (India)	221	26	0.04	5525	0–26	10.13	Angelucci and Bennett (2024)

Notes. Column 1: cost of the intervention per treated participant in USD as reported in the study. Column 2: $t^\star$ is time of endline measurement. Column 3: "Effect (SD) at $t^\star$" is the standardized impact on a mental health-related outcome (typically depression) at $t^\star$. For studies reporting treatment effects in SD units, we use the reported SD effect at the relevant follow-up time point. For studies reporting effects in outcome units, we standardize by the baseline pooled SD of the outcome (or the paper's stated standardization, when provided). We code improvements as positive (reductions in symptom severity); when a paper reports the outcome so that higher values are worse, we flip the sign accordingly. Column 4: "$/SD at $t^\star$" is Cost / Effect at $t^\star$ when Effect > 0; otherwise it is not reported (—). Column 5: AUC window is the months for which we compute Area Under the Curve (AUC). Column 6: "$/0.1 SD-mo." is the cost to improve mental health by 0.1 SD per month. It is computed as $Cost/\left(\int_0^T |\Delta_t|\,dt\right)$ over an AUC window $[0,T]$ in months. When multiple effect estimates are available, we approximate $\int_0^T |\Delta_t|\,dt$ using trapezoids between measurement points and set $\Delta_0 = 0$ (no pre-baseline effect). Unless otherwise stated, SD effects are taken as reported in the cited paper. Bhat et al. (2022) Figure A.7 reports cumulative *depression-months averted* for HAP over five years (PHQ-9$\geq$ 10) of 9 months. We convert depression months to SD months using Bhat et al.'s long-run mapping between SD changes in PHQ-9 and changes in depression prevalence. AUC integrates over 0, 3, 12, and 60 months. For Baranov et al. (2020), we use the treatment effects on the "depression severity" index at 6 months, 1 year, and 7 years reported in Table H. 26. AUC integrates over 0, 6, 12, and 84 months. For Angelucci and Bennett (2024) we AUC uses the paper's own PHQ-9×months construction: "during" (8 months) plus "after" (18 months) contributions. AUC integrates over 0, 8, and 26 months.

For each study, we report two quantities. First, "cost per SD at a fixed horizon" equals the intervention cost per participant divided by the standardized treatment effect on a depression-related outcome at a specified follow-up horizon, $t^\star$. Second, "cost per SD-month" adjusts for duration by dividing cost by the area under the treatment-effect curve, measured in SD-months. We approximate this area using the reported treatment effects at available follow-up points and linear interpolation between them. This second measure is useful because interventions may differ not only in the size of their effects at a point in time, but also in how long those effects persist. Both measures should be interpreted with caution, since the underlying studies differ in price year, whether costs are incremental or total, and whether they include supervision, training, or overhead.

M Impacts from Extending App Access

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

	6 months (1)	All 3 surveys (2)
App Extension	0.016	0.020
	(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 to the App extension randomization. Columns correspond to completion at 6 months and completion of all three follow-up surveys. Robust standard errors in parenthesis. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table M.2: Balance Table for Extension Experiment

Variable	Randomized Sample		Completed 6 mo. survey	
	Control (1)	Δ Treated (2)	Control (3)	Δ Treated (4)
Age	37.96	0.18	38.04	0.04
	(7.66)	(0.31)	(7.90)	(0.35)
Below median income	0.43	0.02	0.41	0.03
	(0.50)	(0.03)	(0.49)	(0.04)
Works	0.48	0.01	0.48	0.02
	(0.50)	(0.03)	(0.50)	(0.04)
Number of children	1.03	-0.10	0.99	-0.06
	(0.99)	(0.06)	(0.96)	(0.07)
Single	0.28	0.02	0.29	0.00
	(0.45)	(0.03)	(0.46)	(0.03)
Some tertiary education	0.64	-0.01	0.64	-0.00
	(0.48)	(0.03)	(0.48)	(0.03)
Depression/Anxiety (PHQ-4)	6.57	0.15*	6.59	0.10
	(2.62)	(0.09)	(2.62)	(0.10)
Stress (PSS-4)	10.10	-0.16	10.16	-0.16
	(2.56)	(0.15)	(2.52)	(0.16)
Well-being (WHO-5)	30.58	0.64	30.70	0.73
	(16.06)	(0.97)	(15.89)	(1.06)
Self-efficacy (GSE)	15.31	-0.08	15.28	-0.17
	(3.19)	(0.20)	(3.15)	(0.23)
Locus of control	11.77	-0.07	11.74	-0.19
	(2.75)	(0.18)	(2.79)	(0.20)
Extroversion (Big5)	8.28	-0.13	8.24	-0.04
	(2.57)	(0.16)	(2.61)	(0.18)
Agreeableness (Big5)	12.22	0.10	12.22	0.05
	(1.89)	(0.13)	(1.91)	(0.14)
Conscientiousness (Big5)	10.40	-0.00	10.33	0.18
	(2.70)	(0.17)	(2.71)	(0.19)
Neuroticism (Big 5)	11.29	-0.21	11.41	-0.23
	(2.37)	(0.14)	(2.27)	(0.15)
Openness to Experience (Big5)	10.73	0.03	10.74	0.03
	(2.07)	(0.13)	(2.04)	(0.14)
Visited psych.	0.32	-0.01	0.31	-0.01
	(0.47)	(0.03)	(0.46)	(0.03)
Observations	983		801	
Joint F-test (p-value)		0.519		0.691

Note: The table reports baseline characteristics by extension assignment for the treated sample. "Full Sample" includes all individuals randomized in the second randomization. "Endline (6 months)" restricts to those who completed the 6-month endline survey. Odd columns report the control group mean with standard deviations in parentheses. Even columns report the coefficient on the extension indicator from a regression with stratification fixed effects, and robust standard errors are reported in parentheses. The joint F-test p-value is from a regression of the extension indicator on all balance variables with stratification fixed effects and robust standard errors. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

N Psychotherapy Use

Digital Care and Psychotherapy: A Simple Model

This appendix presents a simple model of digital care and in-person psychotherapy to clarify why access to digital care may leave psychotherapy unchanged, crowd it out, or crowd it in. People choose AI-powered digital care, $A \geq 0$, and in-person psychotherapy, $T \geq 0$. Utility depends on mental health and residual resources, $U = u(H,R)$, where $u_H > 0$, $u_R > 0$, $u_{HH} \leq 0$, and $u_{RR} \leq 0$. Mental health is produced according to $H = H(A,T)$, with $H_A > 0$ and $H_T > 0$. Residual resources satisfy $R = \bar{R} - q_A A - q_T T$, where $\bar{R}$ denotes total available resources and $q_A, q_T > 0$ are the generalized unit costs of digital care and psychotherapy. We assume that digital care has lower monetary and non-monetary costs than psychotherapy, i.e., $q_A < q_T$.

Total resources, $\bar{R}$, encompass income, financial means, time, attention, privacy, logistical capacity, and the ability to seek and sustain treatment. As a result, a person may have low effective resources not only because she is poor, but also because stigma, distress, or impaired functioning make treatment harder to initiate or sustain. Substituting the resource constraint into the utility, the individual solves

$$\max_{A,T \geq 0} \; u(H(A,T), \bar{R} - q_A A - q_T T)\,.$$

Interior and corner solutions. If the optimum is interior, so that $A > 0$ and $T > 0$, the first-order conditions are

$$u_H H_A = u_R q_A, \qquad u_H H_T = u_R q_T.$$

At a point with $T = 0$, psychotherapy is not used if $u_H H_T(A,0) \leq u_R q_T$. Psychotherapy may therefore not be chosen even if it has positive health returns, because its cost is too high relative to available resources. Similarly, at a point with $A = 0$, digital care is not used if $u_H H_A(0,T) \leq u_R q_A$. The individual chooses neither input if $u_H H_A(0,0) \leq u_R q_A$ and $u_H H_T(0,0) \leq u_R q_T$.

Because $q_A < q_T$, it is possible that digital care is used while psychotherapy is not. A sufficient local condition is that $u_H H_A(0,0) > u_R q_A$ while $u_H H_T(A,0) \leq u_R q_T$ for the relevant level of A. In this case, digital care expands access to health care without inducing psychotherapy use.

Counterfactual analysis: introducing technology A. Before digital care is available, the individual chooses psychotherapy only:

$$\max_{T \geq 0} \; u(H(0,T), \bar{R} - q_T T) \,.$$

If the optimum is interior, it satisfies $u_H H_T(0,T) = u_R q_T$; if instead $u_H H_T(0,0) \leq u_R q_T$, then the individual remains at the corner $T = 0$. After digital care is introduced, the problem becomes

$$\max_{A,T \geq 0} \; u(H(A,T), \bar{R} - q_A A - q_T T) \,.$$

The introduction of A expands the set of feasible ways to improve mental health. Because $q_A < q_T$, some people may adopt digital care even if psychotherapy remains out of reach. Others may adjust psychotherapy use upward or downward depending on whether digital care acts mainly as a substitute, as a complement, or as a way to lower the effective cost of therapy by improving mental health.

Counterfactual analysis: no impact on T. A first possibility is that psychotherapy does not change in equilibrium. This occurs when uptake of digital care is concentrated among individuals who would otherwise have remained untreated. Formally, a person may satisfy $u_H H_T(0,0) \leq u_R q_T$ before the app is available, so that psychotherapy is not used, and also satisfy $u_H H_A(0,0) > u_R q_A$ after the app is introduced, so that digital care is adopted. In this case, $T^{\text{after}} = T^{\text{before}} = 0$ while $A^{\text{after}} > 0$: psychotherapy use is unchanged, yet access to mental health support expands.

Counterfactual analysis: crowding out of T. A second possibility is crowd-out. Digital care may crowd psychotherapy out if the two inputs are substitutes in mental health production, that is, if $H_{AT} < 0$. In this case, app use lowers the marginal health return to psychotherapy. For an interior choice of therapy, the relevant first-order condition is $u_H H_T = u_R q_T$. A rise in A then lowers the left-hand side through H_T; if the left-hand side is decreasing in T, restoring equality requires a lower value of T. Thus, substitution creates a force toward lower psychotherapy use at the intensive margin. The same logic applies at the extensive margin: a person at $T = 0$ enters therapy only if $u_H H_T(A,0) > u_R q_T$, and if $H_{AT} < 0$, higher app use makes this inequality less likely to hold. A second crowd-out force may arise even when $H_{AT} = 0$. If the app improves mental health directly, then the marginal utility of further health gains, u_H, may fall, reducing the marginal value of psychotherapy and thereby lowering therapy use.

Counterfactual analysis: crowding in of T. A third possibility is crowd-in through technological complementarity or cost reduction. Consider technological complementarity first. If $H_{AT} > 0$, app use increases the marginal health return to psychotherapy.[17] For an interior choice of therapy, the relevant first-order condition is $u_H H_T = u_R q_T$. A rise in A then increases the left-hand side through H_T. If the left-hand side is decreasing in T, restoring equality requires a higher level of psychotherapy. Thus, technological complementarity tends to raise T at the intensive margin.

The same logic applies at the extensive margin. A person at $T = 0$ enters therapy when $u_H H_T(A,0) > u_R q_T$. If $H_{AT} > 0$, then higher app use raises $H_T(A,0)$, making this inequality more likely to hold. Thus, complementarity can crowd psychotherapy in both by increasing therapy use among existing users and by inducing entry into therapy among non-users.

Next, crowd-in can occur through cost reduction. If worse mental health raises the effective cost of psychotherapy, then app use may crowd therapy in by improving mental health and thereby lowering those costs. A simple reduced-form way to represent this is to let the unit cost of psychotherapy depend on mental health, so that $q_T = q_T(H)$ with $q_T'(H) < 0$. The problem then becomes

$$\max_{A,T \geq 0} \; u(H(A,T), \bar{R} - q_A A - q_T(H(A,T))T) \, .$$

For an interior choice of psychotherapy, the first-order condition is

$$u_H H_T = u_R \left[q_T(H) + q_T'(H) H_T T \right] .$$

If app use improves mental health, then it lowers $q_T(H)$. This reduces the effective marginal cost of psychotherapy and tends to raise the optimal level of therapy. Intuitively, participants in better mental health may find it easier to schedule sessions, travel to them, concentrate during them, tolerate the emotional demands of treatment, and sustain attendance over time.

At the extensive margin, a person who initially chose $T = 0$ may switch into therapy after app access if

$$u_H H_T(A,0) > u_R q_T(H(A,0)).$$

An improvement in mental health lowers $q_T(H(A,0))$, making this inequality more likely to hold. Thus, app-induced cost reduction can crowd psychotherapy in at both the intensive and extensive margins.

[17] This argument isolates the role of technological complementarity, $H_{AT} > 0$. Because app use may also improve mental health directly, it may reduce the marginal utility of further health gains, u_H, partially offsetting the complementarity effect. Thus, $H_{AT} > 0$ creates a force toward crowd-in, but does not mechanically imply a larger equilibrium value of T in every case.

Additional Empirical Results for Psychotherapy Use

Table N.1: Impacts on the Demand for Psychotherapy (Fixed Sample)

	Visit Psychologist (1)	Num. Visits (2)
Panel A. Effects at 1 month		
Treated	0.032*	
	(0.017)	
Mean control	0.15	
Observations	1558	
Panel B. Effects at 2 months		
Treated	0.059***	0.149***
	(0.018)	(0.055)
Mean control	0.15	0.36
Observations	1558	1558
Panel C. Effects at 6 month		
Treated	0.080***	0.306***
	(0.018)	(0.073)
Mean control	0.14	0.49
Observations	1558	1558
1 mo=2 mo (*p-val*)	0.28	
2 mos=6 mos (*p-val*)	0.44	0.09
1 mo=6 mos (*p-val*)	0.08	
1 mo=2 mos=6 mos (*p-val*)	0.18	

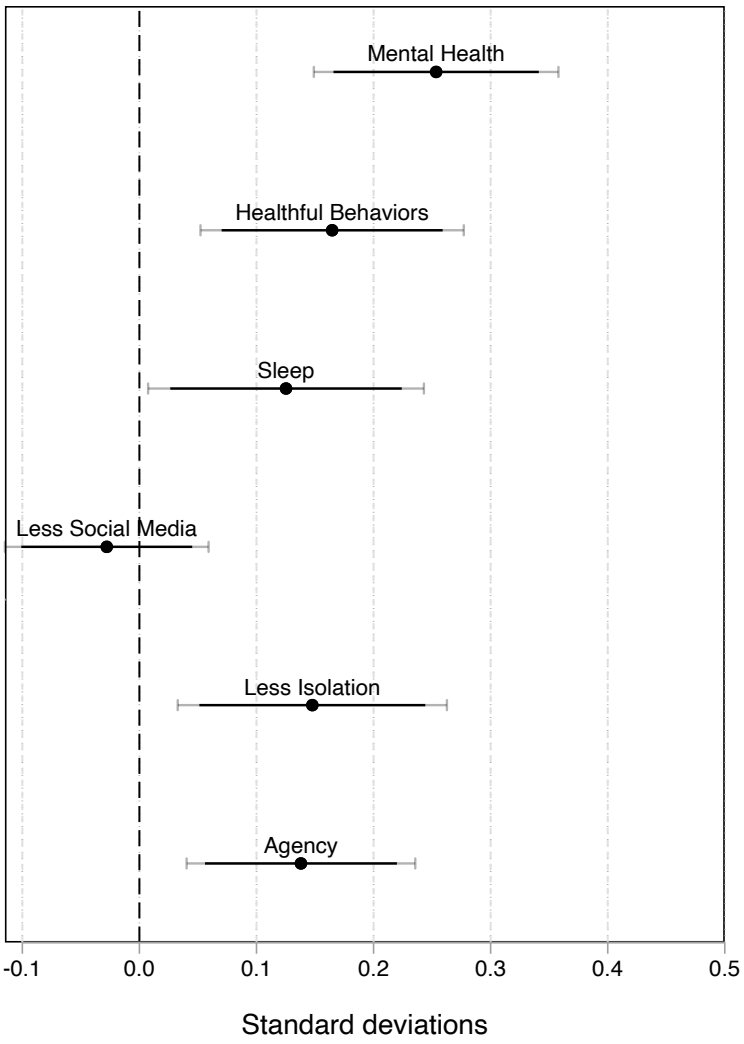
Notes: We restrict the sample to participants for whom we have data at one, two, and six months, excluding participants who do not respond to at least one of the three follow-up surveys. The first outcome, *Any Visit Psychologist*, is an indicator variable equal to 1 if the respondent reports at least one visit to a psychologist or psychotherapist in the previous month. The second outcome, *Number of Visits*, reports the unconditional psychotherapy visits in the previous month. Regressions control for strata, baseline dependent variable when available, and lasso-selected baseline covariates. Robust standard errors. * $p < .10$, ** $p < .05$, *** $p < .01$.

Table N.2: Mediation of Treatment Effects on Mental Health by Psychotherapy Use

	Natural Indirect Effect (NIE)	Natural Direct Effect (NDE)	Total Effect (TE) (NIE+NDE)	Proportion Mediated (NIE/TE)	Observations
1 month	0.005	0.306***	0.311***	0.017	1,850
	(0.004)	(0.037)	(0.038)	(0.012)	
2 months	0.016**	0.301***	0.318***	0.052**	1,824
	(0.007)	(0.041)	(0.041)	(0.022)	
6 months	0.008	0.234***	0.242***	0.033	1,599
	(0.007)	(0.043)	(0.043)	(0.029)	

Notes: This table reports estimates from a causal mediation analysis with a binary treatment. The outcome is a standardized mental health index measured at 1, 2, and 6 months. The mediator is a binary indicator for whether the respondent attended at least one psychotherapy session in the previous month. All models condition on baseline covariates and strata as specified in the main specifications. The outcome equation is estimated using linear regression. The mediator equation is estimated using a linear probability model to ensure comparability with the main specifications; results are robust to using a probit mediator model. The outcome equation includes a treatment–mediator interaction. The proportion mediated is defined as NIE/TE, where $\text{TE} = \text{NIE} + \text{NDE}$. Identification relies on cross-world independence (sequential ignorability) and common support assumptions conditional on covariates. Robust standard errors in parentheses.
$^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$

Figure N.1: Treatment Effects for Individuals Who Did Not Attend Psychotherapy



Notes: This figure reports the differences in mental health outcomes between treatment and control groups, restricting the sample to respondents who did not attend psychotherapy during the study period. Thick and thin bars indicate 90% and 95% confidence intervals. Because psychotherapy attendance is a post-treatment outcome, this analysis conditions on an endogenous variable and should not be interpreted causally. Standard errors clustered by individual.

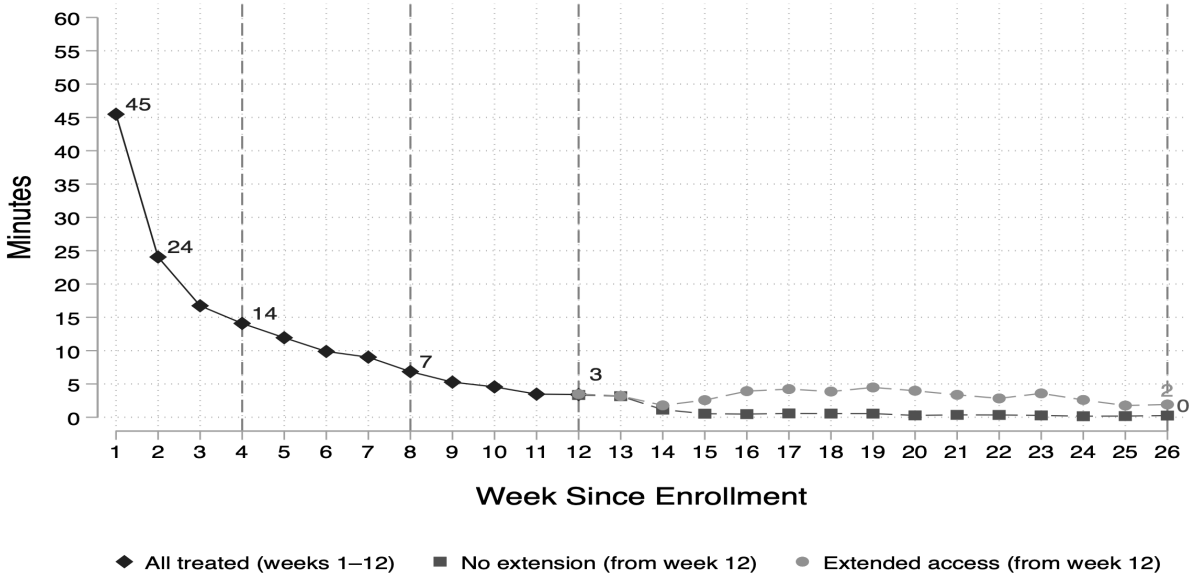
O Supplemental material - Not for Publication

Table O.1: Deviations from the Pre-Analysis Plan

Pre-Specified Approach	Deviation	Rationale & Exhibits
If baseline imbalances, present weighted estimates; otherwise use all available observations.	For weekly affect outcomes, the paper reports Lee bounds due to differential weekly-response attrition.	We applied IPW and the results were largely unchanged. Thus, we provide Lee bounds in the paper because they are more conservative. See figure 5 and table F.3.
Not pre-specified.	We created a "Labor Market Index" at 2 and 6 months using absenteeism, an employment indicator, and hours worked.	This follow from finding that app access reduced absenteeism (pre-specified), which led us to explore additional labor-market outcomes. Results are reported in Table 3.
Measure "missed work" conditional on employment ("if employed").	"Missed work" is measured unconditionally (asked of all respondents).	We cannot condition on employment because the likelihood of being employed is affected by treatment. Results are reported in table 2.
Not pre-specified.	At 2 and 6 months, we measured self-reported use of behavioral tools taught by the app.	This supports the conjecture that app use leads to sustained mental health benefits by teaching tools consistent with CBT. Results are reported in table J.1.
Not pre-specified.	We perform mediation analyses to estimate which share of the mental health ITT effects are mediated by (i) healthful behavior and sleep, and (ii) psychotherapy.	This helps assess how app use leads to sustained mental health impacts. Results are reported in tables J.2 and N.2.
Use a IV-LATE specification in which we instrument the indicator "used the app in the relevant time period for at least 20 minutes per week" with the indicator for random assignment to the treatment group.	Use a IV-LATE specification in which we instrument the indicator "having downloaded the app in the relevant time period" with the indicator for random assignment to the treatment group.	The exclusion restriction in the pre-specified specification is unlikely to hold, given that there appear to be positive treatment effects on mental health even for low-engagement users. We report the revised IV-LATE estimates in table G.2.

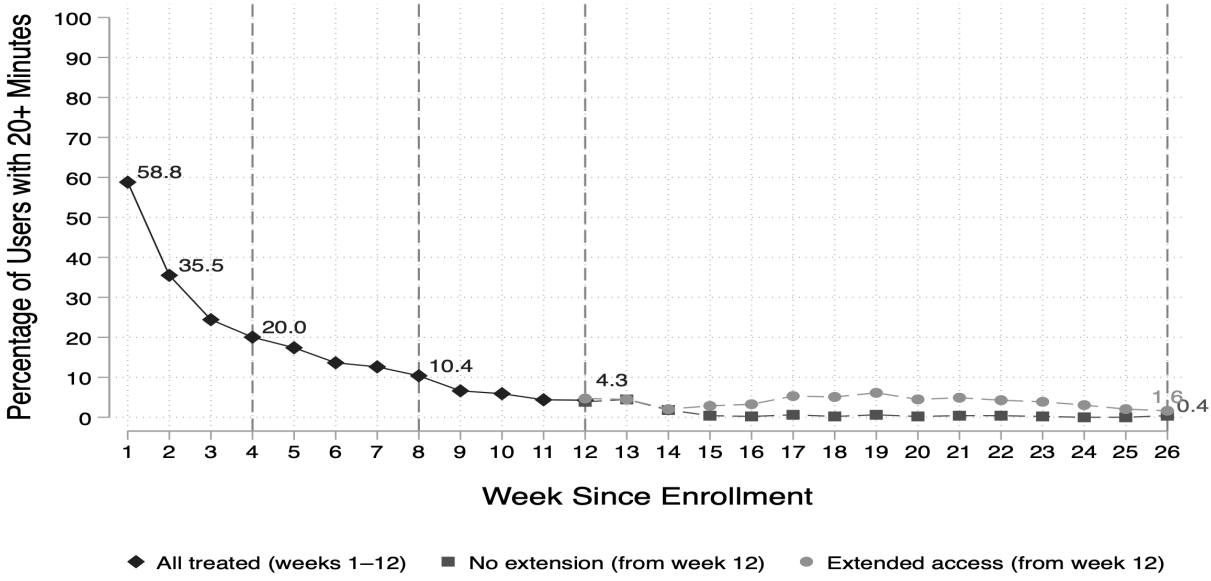
Note: The table lists and explains all deviations from our analysis plan. The analysis plan and its addendum are available through entry AEACTR-0015877 in the AEA RCT Registry.

Figure O.1: App Use in Minutes per Week



Notes: The figure shows the average minutes per week of app use. Vertical dashed lines indicate weeks 4, 8, and 26 since enrollment. Markers represent weekly values, and labels are displayed for weeks 1, 2, 4, 8, 12, and 26.

Figure O.2: Proportion of People Using the App at least 20 Minutes per Week



Notes: The figure shows the weekly share of users with at least 20 minutes of app use. Percentages are calculated as the proportion of enrolled users with total weekly app use of at least 20 minutes. Vertical dashed lines indicate weeks 4, 8, and 26 since enrollment. Markers represent weekly values, and labels are displayed for weeks 1, 2, 4, 8, 12, and 26.

Table O.2: Treatment Effects on Index Outcomes by Baseline Psychotherapy Use

	Mental Health (1)	Healthful Behaviors (2)	Sleep (3)	Less Social Media (4)	Less Isolation (5)	Agency (6)	Visit Psychologist (7)
Panel A. Pooled							
Treated	0.286***	0.214***	0.181***	0.022	0.210***	0.199***	0.057***
	(0.030)	(0.034)	(0.033)	(0.033)	(0.034)	(0.028)	(0.009)
Treated × Visited psych.	-0.009	0.099*	0.029	0.118*	-0.036	-0.050	-0.016
	(0.054)	(0.060)	(0.061)	(0.062)	(0.065)	(0.050)	(0.027)
Observations	5274	5274	5274	5274	5274	5274	5273
Panel B. At 1 month							
Treated	0.313***	0.315***	0.206***	0.055	0.261***	0.191***	0.047***
	(0.045)	(0.053)	(0.053)	(0.052)	(0.054)	(0.044)	(0.014)
Treated × Visited psych.	-0.045	0.057	0.041	0.088	-0.112	-0.096	-0.049
	(0.083)	(0.097)	(0.098)	(0.096)	(0.104)	(0.079)	(0.043)
Observations	1850	1850	1850	1850	1850	1850	1850
Panel C. At 2 months							
Treated	0.304***	0.144***	0.182***	0.047	0.183***	0.237***	0.057***
	(0.050)	(0.055)	(0.056)	(0.054)	(0.056)	(0.046)	(0.015)
Treated × Visited psych.	0.023	0.141	0.001	0.124	0.038	0.020	-0.011
	(0.088)	(0.097)	(0.099)	(0.098)	(0.103)	(0.080)	(0.044)
Observations	1825	1825	1825	1825	1825	1825	1824
Panel D. At 6 months							
Treated	0.233***	0.171***	0.132**	-0.048	0.176***	0.168***	0.071***
	(0.051)	(0.061)	(0.057)	(0.058)	(0.058)	(0.048)	(0.018)
Treated × Visited psych.	0.017	0.103	0.053	0.214**	-0.040	-0.084	0.018
	(0.096)	(0.102)	(0.109)	(0.106)	(0.114)	(0.087)	(0.047)
Observations	1599	1599	1599	1599	1599	1599	1599

Notes: Each panel reports impact heterogeneity for seven index outcomes (columns): Mental Health, Healthful Behaviors, Sleep, Less Social Media, Less Isolation, Agency, and for Visit Psychologist last month. Rows report the main effect of treatment (Treated) and the interaction between treatment and a dummy variable indicating whether they had therapy in 2024 at baseline. All models include strata, lasso-selected baseline covariates, and the heterogeneity variable. Standard errors clustered at the respondent level are in parentheses for the pooled analysis, and robust standard errors for Panels B, C, and D. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table O.3: Treatment Effects on Index Outcomes by Baseline Mental Health

	Mental Health (1)	Healthful Behaviors (2)	Sleep (3)	Less Social Media (4)	Less Isolation (5)	Agency (6)	Visit Psychologist (7)
Panel A. Pooled							
Treated	0.288***	0.244***	0.189***	0.059**	0.198***	0.183***	0.052***
	(0.025)	(0.028)	(0.028)	(0.028)	(0.029)	(0.023)	(0.010)
Treated × Mental Health Index (Baseline)	-0.070***	-0.054*	-0.013	0.082***	-0.033	-0.023	-0.022**
	(0.027)	(0.028)	(0.028)	(0.028)	(0.030)	(0.025)	(0.010)
Observations	5274	5274	5274	5274	5274	5274	5273
Panel B. At 1 month							
Treated	0.305***	0.331***	0.219***	0.082*	0.226***	0.161***	0.032**
	(0.038)	(0.045)	(0.044)	(0.044)	(0.046)	(0.037)	(0.016)
Treated × Mental Health Index (Baseline)	-0.023	-0.043	-0.029	0.033	0.019	0.028	-0.019
	(0.040)	(0.046)	(0.045)	(0.045)	(0.047)	(0.038)	(0.016)
Observations	1850	1850	1850	1850	1850	1850	1850
Panel C. At 2 months							
Treated	0.312***	0.187***	0.183***	0.086*	0.194***	0.243***	0.054***
	(0.041)	(0.045)	(0.046)	(0.045)	(0.047)	(0.038)	(0.017)
Treated × Mental Health Index (Baseline)	-0.074*	-0.026	0.040	0.099**	-0.052	-0.024	-0.028*
	(0.044)	(0.045)	(0.045)	(0.047)	(0.049)	(0.040)	(0.016)
Observations	1825	1825	1825	1825	1825	1825	1824
Panel D. At 6 months							
Treated	0.239***	0.202***	0.149***	0.016	0.164***	0.143***	0.076***
	(0.043)	(0.049)	(0.049)	(0.049)	(0.050)	(0.040)	(0.018)
Treated × Mental Health Index (Baseline)	-0.112**	-0.085*	-0.052	0.103**	-0.057	-0.062	-0.014
	(0.047)	(0.051)	(0.051)	(0.050)	(0.049)	(0.042)	(0.018)
Observations	1599	1599	1599	1599	1599	1599	1599

Notes: Each panel reports double-ML estimates for seven index outcomes (columns): Mental Health, Healthful Behaviors, Sleep, Less Social Media, Less Isolation, Agency, and for Visit Psychologist last month. Rows report the main effect of treatment (Treated) and the interaction between treatment and a dummy variable indicating the mental health index at baseline. All models include strata, lasso-selected baseline covariates, and the heterogeneity variable. Standard errors clustered at the respondent level are in parentheses for the pooled analysis, and robust standard errors for Panels B, C, and D. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.