The Long-Run Effects of Psychotherapy on Depression, Beliefs, and Economic Outcomes*

Bhargav Bhat† Jonathan de Quidt Johannes Haushofer Vikram Patel
Gautam Rao Frank Schilbach Pierre-Luc Vautrey

May 9, 2022

Abstract

We revisit two clinical trials that randomized depressed adults in India (n=775) to a brief course of psychotherapy or a control condition. Four to five years later, the treatment group was 11 percentage points less likely to be depressed than the control group. The more effective intervention averted 9 months of depression on average over five years and cost only $66 per recipient. Therapy changed people's beliefs about themselves in three ways. First, it reduced their likelihood of seeing themselves as a failure or feeling bad about themselves. Second, when faced with a novel work opportunity, therapy reduced over-optimistic belief updating in response to feedback and thus reduced overconfidence. Third, it increased self-assessed levels of patience and altruism. Therapy did not increase levels of employment or consumption, possibly because of other constraints on employment in the largely female study sample.

*We dedicate this paper to our coauthor Bhargav Bhat, who passed away during the course of this work. We thank our entire field team in Goa for excellent research assistance, especially Siddhant Gokhale, Advait R. Aiyer, Gunjita Gupta, and Anahita Karandikar. We thank our research managers at the Behavioral Development Lab, Adrien Pawlik, and Erik Hausen. We thank Ralph Ignacio Lawton, Hongyuan Xia, Zhangchi Ma, Haonan Ye, Kartik Vira, and Sejal Aggarwal for research assistance with the data analysis. We thank Benedict Weobong for his early support of our project. We thank all our study participants for their time and patience. We received IRB approval from Princeton. The experiment was pre-registered on the AEA registry, number AEARCTR-0003823. We gratefully acknowledge funding from the Weiss Family Program Fund for Research in Development Economics, the Mind Brain Behavior Initiative at Harvard, the Dean's Fund at Harvard, the SHASS Research Fund at MIT, the Pershing Square Venture Fund for Research on the Foundations of Human Behavior, and Princeton University. de Quidt acknowledges financial support from Handelsbanken's Research Foundations (grant BF17-0003).

†Bhat: Sangath; de Quidt: Institute for International Economic Studies, Stockholm University, CAGE, CEPR, CESifo, ThReD (Jonathan.deQuidt@iies.su.se); Haushofer: Stockholm University (johannes.haushofer@ne.su.se); Patel: Harvard and Sangath (Vikram_Patel@hms.harvard.edu); Rao: Harvard and NBER (grao@fas.harvard.edu); Schilbach: MIT and NBER (fschilb@mit.edu) Vautrey: MIT (vautrey@mit.edu)

1 Introduction

Psychotherapy aims to treat psychological disorders by changing dysfunctional beliefs, thoughts, and behaviors. It is effective in treating common mental-health conditions such as depression and anxiety (Barth et al., 2016; Cuijpers et al., 2016). Psychotherapy is widely used in the rich world: nearly 10 percent of US adults received some form of therapy from a mental-health professional in 2019 (Terlizzi and Zablotsky, 2020). Simplified forms of psychotherapy have been developed to fill the enormous treatment gaps for mental illness in low-income countries (Patel et al., 2009, 2018). Administered by trained non-specialists at low cost, these therapies have been shown to cause sizable short-run improvements in mental health across a variety of low-income settings (Singla et al., 2017).

This paper studies the long-run effects of psychotherapy on mental health, behavioral and economic outcomes in a low-income setting. We aim to fill three gaps in the literature. First, little is known about whether the effects of psychotherapy on depression persist beyond the three to twelve months typically documented in the literature (Figure 1).[1] Second, by measuring economic beliefs and preferences using methods from experimental economics, we provide a behavioral-science view of the effects of psychotherapy (Blattman, Jamison and Sheridan, 2017; Heller et al., 2017). Third, while poor mental health is associated with lower income, it is unclear whether psychotherapy increases economic well-being in the long run, and can thus be a tool to reduce poverty (Ridley et al., 2020).

We revisit participants from two psychotherapy RCTs in India, four to five years after the original trials. In both studies, participants who had screened positive for depression—the world's leading cause of disability (Friedrich, 2017)—were randomized into receiving a brief course of psychotherapy or to a control condition. Therapy was delivered by non-specialist counselors and cost under $70 per recipient to deliver. The interventions primarily employed 'behavioral activation', teaching patients the relationship between their activities and mood, and seeking to increase mood-improving activities. We successfully locate and interview 76 percent of participants—589 out of 775—from the two trials.[2] To quantify the extent to which our findings might shift existing scientific understanding, we asked 234 experts on economics and mental health to forecast the main findings of the larger of the two trials via the Social Science Prediction Platform (DellaVigna, Pope and Vivalt, 2019).

Our first set of results concerns the long-term effects of psychotherapy on mental health. On average across the two trials, the brief courses of psychotherapy led to sustained reductions in depression. We measured symptoms of depression using the PHQ-9 questionnaire, a standard and locally-validated screening tool also used in the original trials. Pooling the two trials, the treatment group had 0.15 standard deviations (SD) lower PHQ-9 scores compared to the control group (p=0.08). Using standard diagnostic cutoffs in PHQ-9 scores, we find 8 to 11 percentage points lower incidence of mild and moderate depression in the treatment group, respectively (p<0.01 and p=0.03). Decomposing the PHQ-9 improvements, we see significantly lower frequency of specific symptoms such as feeling bad about oneself and having a poor appetite. Participants' mood also improved, by 0.17 SD (p=0.04).

The long-run improvement in mental health is entirely driven by one of the two interventions: the

[1] Exceptions are Baranov et al. (2020) and Maselko et al. (2020), which study the effects of psychotherapy to reduce *maternal* depression seven and three years after treatment respectively.

[2] Attrition and baseline characteristics are balanced across treatment and control groups for both trials.

Healthy Activity Program (HAP), which delivered 6 to 8 sessions of therapy to adults with moderately-severe depression (N=495, Patel et al., 2016). The HAP intervention reduces depression symptoms five years later by 0.23 SD (p=0.02) and incidence of both mild and moderate depression by 13 percentage points (p<0.01). This effect equals the 91st percentile of the expert predictions, implying that a vast majority of experts—including mental-health professionals—under-estimate the long-run mental-health effects (despite being informed about the short-run effects). We calculate that treated participants experienced 9 fewer months of depression (with PHQ-9≥10) over the five years since enrollment.

In contrast, the Thinking Healthy Program Peer-Delivered (THPP) intervention—delivered to pregnant women with lower (moderate) levels of depression—has no detectable long-run effects on depression (N=280). This finding is consistent with small short-run effects reported in Fuhr et al. (2019) after our data collection began.[3] The control group in the THPP trial shows high rates of spontaneous recovery, leaving little room for further reductions in depression.[4]

To study participants' views of therapy, we elicit their quantitative beliefs about the treatment effects of therapy, with incentives for accurate guesses. On average, experiencing treatment significantly increases participants' beliefs regarding treatment efficacy. This suggests that making psychotherapy more widely available might increase perceived effectiveness and demand for it from the low levels documented in the literature (Cronin, Forsstrom and Papageorge, 2020; Sapiens Lab, 2021). A caveat is that we find similar effects on perceived efficacy in both trials. Individual patients may not be able to distinguish effective from ineffective treatments, presumably because it is difficult to disentangle treatment effects from natural improvements over time.

We next turn to the effect of therapy on how people form beliefs about themselves, specifically, their beliefs about their ability and performance relative to others. Such beliefs have been shown to affect important economic decisions (Russo and Schoemaker, 1992; Malmendier and Tate, 2005), with widespread existence of both over- and under-confidence (Moore and Healy, 2008).[5] The expected effect of psychotherapy on self-confidence is ambiguous. On the one hand, therapy seeks to make patients see themselves in a more realistic light as having both strengths and weaknesses (Beck, 2020). If therapy causes patients to have a more robust self-image—as evidenced by the significant reductions in feeling bad about oneself or feeling like a failure reported in the PHQ-9 questionnaire—treated participants might have a reduced psychological need for overconfidence in a specific domain (Blanton et al., 2001; Sherman and Cohen, 2006; Kolubinski et al., 2018). These mechanisms predict the treatment group

[3]The small effects of THPP in Fuhr et al. (2019) resemble findings from a similar trial evaluating THPP in Pakistan (Maselko et al., 2020). In contrast, THP—of which THPP is a greatly simplified version delivered by peers—has been found effective in several studies, e.g. Baranov et al. (2020).

[4]As argued in Fuhr et al. (2019), the high rate of remission in the control group may be explained by lower depression severity at recruitment in the trial. THPP used an eligibility criterion of moderate depression (PHQ-9 10), compared to moderately-severe depression (PHQ-9 15) in the HAP trial. More severe depression is less likely to exhibit spontaneous remission. Perinatal depression may also be more likely than other kinds of depression to exhibit remission due to its clearly circumscribed environmental trigger (pregnancy and childbirth).

[5]Self-confidence may also directly affect psychological well-being (Köszegi, 2006), motivation (Bénabou and Tirole, 2002), and task performance (Schwardmann and Van der Weele, 2019). An influential literature has studied how people 'manage' their self-confidence, trading off the benefits of higher self-confidence against the cost of inaccurate beliefs. See e.g., Eil and Rao (2011); Zimmermann (2020); Mobius et al. (2021); Coutts (2019); Buser, Gerhards and Van Der Weele (2018); Ertac (2011). Not all studies find asymmetries in the direction of optimism and some find the opposite, see Benjamin (2019) for discussion.

to have more accurate beliefs about their specific abilities relative to others. On the other hand, an influential literature on 'depressive realism' has posited that depression is associated with more realistic, less overconfident beliefs (Alloy and Abramson, 1979), although empirical evidence for this hypothesis is mixed (Moore and Fresco, 2012; Korn et al., 2014). In this view, psychotherapy might induce overconfidence via reduced depression.

To measure self-confidence, we use a lab-experimental paradigm developed by Mobius et al. (2021). After completing a novel paid work task, participants are asked to assess their performance relative to other study participants. Specifically, we elicit their belief about the probability they performed in the top half of a randomly-formed group of ten. They are then given noisy feedback (signals) on their relative standing and, after each signal, we re-elicit their beliefs. This allows us to measure how they update their beliefs in response to positive and negative signals, relative to a fully rational (Bayesian) benchmark. The median expert predicted that control-group participants, who have high rates of depression, would update their beliefs equally to positive and negative feedback—consistent with the idea of depressive realism. They further predicted that therapy causes patients to put greater weight on positive feedback than negative feedback, thus becoming more overconfident.

Contrary to expert predictions, we find substantial over-confidence and optimistic belief updating in the control group, despite high rates of depression. Prior to receiving any feedback, control-group participants overestimate their probability of being in the top half in performance by 13 percentage points (standard error, s.e.=2 percentage points). When given feedback on their performance, they then update their beliefs optimistically and become even more overconfident: on average, they weight noisy positive signals at 73 percent (s.e.=8 percent) of the Bayesian benchmark, but essentially ignore negative signals altogether.

Before receiving any feedback, the treatment group is just as overconfident as the control group. However, they then update their beliefs less optimistically when given feedback, leading to 8 percentage points lower overconfidence by the end of the experiment compared to the control group (s.e.=3 percentage points).[6] The point estimates for these effects are strikingly similar across the two trials, despite only one trial showing reductions in depression. This suggests that psychotherapy—independently from its effects on depression—helps people have more accurate views of themselves.

We next provide evidence that psychotherapy increases people's self-assessments of their levels of patience (0.24 SD, $p<0.01$) and altruism (0.19 SD, $p<0.05$)—measured using validated survey questions from Falk et al. (2018)—while not changing self-assessed risk tolerance. In contrast, we detect no significant effects on incentivized measures of altruism, patience, and risk tolerance: a dictator game, a savings task, and choices between lotteries.[7] Combining the different types of measures, we find significant increases in overall indices of patience (0.18 SD, $p=0.03$) and altruism (0.21 SD, $p=0.01$) but no effect on risk tolerance. One interpretation of these results is that therapy affects how people

[6] Studying responses to positive and negative feedback separately, we find that this effect is driven by the treatment group being significantly less responsive to positive feedback (p=0.03), with no effect on responsiveness to negative feedback (p=0.51). Combined, this implies less asymmetry in response to positive and negative feedback (p=0.08). This could be be interpreted as an increase in 'rationality' as a Bayesian should respond equally to positive and negative signals, but is also in some sense further from the Bayesian benchmark, since the treatment group becomes even more conservative in their response to positive signals.

[7] The point estimate on dictator-game giving is positive (0.1 SD) but not significant.

interpreted their own behaviors, causing them to have less negative beliefs about themselves.[8]

Finally, we report impacts on poverty-related outcomes. We find no evidence of significant impacts on consumption or on self-reported employment, labor supply, or earnings, nor on revealed-preference measures of people's willingness to work and on their productivity in the bracelet-making task. In the case of the labor-market outcomes such as employment, reservation wages and real-stakes willingness to accept a job offer, these null effects are fairly precise. For example, we can rule out increases in employment of 5 percentage points compared to a control-group mean of 25 percent. The absence of long-run impacts on labor-market outcomes and consumption may be explained by the fact that the sample is predominantly female, and women face numerous barriers to working outside the household in this context (Fletcher, Pande and Moore, 2017).[9]

This paper contributes to several literatures. First, it expands the fledgling literature on the long-run impacts of psychotherapy on depression. Numerous studies have demonstrated the short-run efficacy of inexpensive psychological interventions in improving mental health in the developing world (Singla et al., 2017). However, with the exception of Baranov et al. (2020), little work exists on the long-run impacts of such interventions or indeed of other psychotherapy interventions to treat depression in rich countries (Steinert et al., 2014). We find substantial long-run effects of therapy on depression, exceeding the quantitative predictions of experts who had been presented with the short-run effects. This suggests that the findings of Baranov et al. (2020) generalize beyond their specific context and population of pregnant women and extend to a simpler form of therapy delivered by counselors outside of the overburdened public health system. The intervention was remarkably cost-effective: about $7 per month of depression averted.

Second, our paper takes a step towards providing a behavioral-science view of psychotherapy, building on work by Blattman, Jamison and Sheridan (2017) and Heller et al. (2017).[10] Our key contribution relative to these papers is the use of lab-experimental methods to study effects of psychotherapy on beliefs, a central aspect of depression and therapy. We find that therapy durably reduced highly negative beliefs about oneself: treated participants were substantially less likely to report feeling like a failure or feeling bad about themselves or about letting others down. They also had higher self-perceptions of their own patience and altruism. Yet therapy did not simply make their beliefs about themselves uniformly more positive. They also developed more accurate—less overconfident—beliefs about themselves in a novel work domain. This pattern of findings is broadly consistent with the theory of CBT (Beck, 2020)—of which behavioral activation is a key element—and arguably contrary to the theory of depression realism, for which we find little support.

Third, we contribute to the literature on the causal effects of therapy on economic outcomes more broadly (Ridley et al., 2020). Lund et al. (2021) review RCTs of psychotherapy for depression in the

[8] Alternatively, self-assessments may reflect a broader set of real-world behaviors than those captured by the experimental measures.

[9] We do detect significant improvements in self-reported sleep quality and duration, but no impacts on measures of female empowerment, experienced intimate partner violence, loneliness, or locus of control.

[10] Blattman, Jamison and Sheridan (2017) show that cognitive-behavioral therapy (CBT) targeted at increasing self-control and reducing impulsive behavior successfully increases patience and reduces violent behavior. Heller et al. (2017) show that an intervention with many elements of CBT reduces crime and dropout among disadvantaged youth in Chicago, partly by causing recipients to act less on automatic thoughts.

developing world and identify short-term reductions in self-reported days missed at work. Also in the short-run, Barker et al. (2021) find that a CBT intervention improves mental health, cognition, and self-perceived economic status among poor (not necessarily depressed) individuals in Ghana. In the longer-run, Baranov et al. (2020) show effects of therapy on female empowerment and investments in children in a sample of depressed pregnant women in Pakistan. More broadly, Bossuroy et al. (2022) show that adding psychosocial support—life-skills training and community sensitization around aspirations and social norms—to a multi-faceted intervention including cash transfers, coaching and entrepreneurship training further improves economic and psychological outcomes among extremely poor households in Niger. We detect no long-run impacts of psychotherapy for depression on consumption and work-related outcomes, suggesting that improving mental health is not enough to generate sustained economic gains in this population. It is possible that therapy would generate greater economic benefits as part of a multi-faceted intervention, as in Bossuroy et al. (2022).[11]

Finally, we provide novel evidence of people's perceptions of the efficacy of psychotherapy. When asked about reasons for not seeking mental healthcare, respondents in ten countries commonly cited lack of confidence in treatment as a reason (Sapiens Lab, 2021). Our incentivized belief measures indeed show that HAP trial control-group participants underestimate the efficacy of the treatment, which points to possible demand-side barriers to the adoption of effective treatments. We also show that experiencing therapy durably increases beliefs about its efficacy, indicating that demand for therapy might increase in a society as the population comes to have more experience with it.

2 Background and Study Design

2.1 Study background

Depression, one of the most common mental disorders, has a life-time prevalence of about 20 percent (Kessler and Wang, 2009). Estimates for India show that about 3.3 percent of the population suffers from depression at a given point in time (Sagar et al., 2020). More recent evidence suggests stark increases in mental disorders since the onset of the Covid-19 pandemic in India and globally (Verma and Mishra, 2020; Santomauro et al., 2021). While depression affects all segments of society, it is particularly pronounced among the poor in any given setting (Ridley et al., 2020). Depression is also about 70% more prevalent among women than men (Albert, 2015).

Despite the high prevalence of depression and the effectiveness of therapy shown in numerous trials (e.g. Cuijpers et al., 2010, 2015), professional treatment for common mental disorders in the form of medication or psychotherapy is not available in many parts of the world. A key reason for such treatment gaps is the low supply of trained psychiatrists. A promising solution is the development of inexpensive treatments to be delivered by non-specialist counsellors (Patel et al., 2009, 2011). Such interventions have been found to be effective in a range of low-resource settings worldwide (Singla et al., 2017), including India (Patel et al., 2010).

[11] Haushofer, Mudida and Shapiro (2020) find no effects on either mental health or economic well-being of a therapy intervention one year after treatment. However, that study did not establish short-term effects on mental health.

2.2 Description of trials

Located in Goa, a state on the west coast of India, the non-profit research organization Sangath seeks to make effective mental-health services more widely accessible by developing and testing non-specialist mental healthcare interventions.[12] We follow up on two RCTs of psychotherapy for depression implemented by Sangath in partnership with academic researchers between 2013 and 2016.

Healthy Activity Programme (HAP). The first trial was designed to estimate the effects of the 'Healthy Activity Program' (HAP) psychotherapy intervention on depression 3 and 12 months post enrollment (Patel et al., 2017; Weobong et al., 2017). The sample of 495 participants aged 18 to 65 was recruited between October 2013 and July 2015 at ten primary health centers (PHCs) in Goa. Potential participants were screened for depression using the Patient Health Questionnaire (PHQ-9), a nine-item depression screening tool.

Adults visiting the health centers who screened positive for at least moderately-severe depression—as measured by a PHQ-9 score of at least 15—were offered participation in the trial. About 2 percent of those screened were found to be eligible for the trial.[13] These participants were generally not seeking treatment for depression and were instead visiting the PHC for other medical conditions. Those who agreed to participate in the trial—about two-thirds of those eligible—were then randomized to one of two conditions:

- *Control group: EUC.* The control group received 'enhanced usual care' (EUC), which entailed informing both the participant and the physician at the PHC of the positive depression screening result. The physician was also provided with adapted treatment guidelines for depression from the WHO's Mental Health Gap Action Programme (WHO, 2010). Given the scarcity of trained psychiatrists and the workload of physicians in the PHCs, this condition in practice entailed little active treatment of depression.

- *Treatment group: EUC+HAP.* The treatment group received EUC plus HAP, a psychological treatment based on behavioral activation. The HAP intervention consisted of 6 to 8 weekly sessions of 30 to 40 minutes each, delivered individually at participants' homes or at the local PHC. Counselors were members of the local community, recruited through newspaper advertisements and word of mouth, screened through an interview procedure, and trained by Sangath. The central aspect of the treatment was to encourage participants to schedule and engage in pleasurable activities of their choice. Depression often involves withdrawing from such activities, which in turn can feed back into low mood. Behavioral activation is theorized to operate by breaking these negative cycles of inactivity and depression. Counselors also educated participants about mental health, taught strategies to avoid rumination, and encouraged them to take steps to solve the problems they faced in their lives.[14]

About 70 percent of HAP participants completed the full course of therapy by attending all the assigned sessions. Intent-to-treat analyses revealed that the HAP intervention increased remission from

[12] Sangath was co-founded by coauthor Vikram Patel.

[13] 8 percent of those screened had at least moderate depression (a PHQ-9 score of at least 10).

[14] In case of specific needs, counselors also taught strategies to improve interpersonal communication skills, improve sleep and train relaxation.

depression (having a PHQ-9 score below 10) by 25 percentage points after 3 months (Patel et al., 2017) and by 16 percentage points after 12 months (Weobong et al., 2017). These findings were published before data collection for our follow up began.

Thinking Healthy Programme Peer-Delivered (THPP). The second RCT was designed to measure the impacts of the Thinking Healthy Program Peer-delivered (THPP) on depression among pregnant women (Fuhr et al., 2019). THPP is a simplified version of a psychological intervention (THP) for treating perinatal depression that has been found to be effective in similar settings and is recommended by the WHO (Rahman et al., 2008, 2013; WHO, 2015; Baranov et al., 2020). While the original THP trials employed a full-fledged cognitive behavioral therapy (CBT) intervention, THPP was a simpler intervention focused on behavioral activation, as in the HAP trial described above. THPP was designed to be delivered by peer counselors, instead of community health workers as in previous trials.

The trial recruited 280 women with perinatal depression between October 2014 and June 2016. Study participants were recruited from two antenatal clinics and two primary health centers in the north district of Goa. Adult women were eligible if they were in their second or third trimester of pregnancy and screened for depression based on a PHQ-9 score of at least 10. Five percent of pregnant women exceeded this screening threshold and were thus eligible to participate in the study. Given the lower screening threshold compared to the HAP trial (10 vs. 15), THPP participants exhibited on average lower depression severity than HAP participants.

- *Control Group: EUC.* The control group again received enhanced usual care (EUC). This included standard care from the gynaecologist as well as information to both patients and gynaecologists that the participant had screened positive for depression. Gynaecologists were also provided with adapted treatment guidelines for perinatal depression (WHO, 2010).

- *Treatment Group: EUC+THPP.* In addition to EUC, participants in the treatment group received the Thinking Healthy Programme (Peer-delivered). The counsellors were middle-aged women with children, who had shown an interest in supporting other women in the community, selected for their good communication skills. The intervention entailed 6 to 14 individual sessions over 7 to 12 months for about 30 to 45 minutes each. The sessions were divided into four phases, beginning in the second or third trimester of pregnancy and ending six months after childbirth. As with the HAP intervention, the THPP intervention focused on behavioral activation.

Seventy-two percent of THPP participants completed the course of therapy. Fuhr et al. (2019) report that the intervention led to moderate improvement in mental health six months after childbirth. Remission from depression—defined in Fuhr et al. (2019) as a PHQ-9 score below 5—was increased significantly by 11 percentage points 6 months after childbirth, although average PHQ-9 scores were not significantly different between treatment and control groups. These results were not yet unblinded when we began our data collection for this study.

Training of counsellors and cost of treatment. In both trials, counsellors received classroom-based training that focused on intervention content and relationship-building skills, followed by a an

internship involving additional training and group supervision (Singla et al., 2014; Sikander et al., 2015). Given their brief nature and mode of delivery, the interventions were inexpensive, with an estimated costs (relative to EUC) of $66 per person for HAP (Patel et al., 2016) and $72 for THPP (Fuhr et al., 2019).

2.3 Recruitment, balance, and incentives

This paper reports data based on follow-up visits conducted on average 5 years after enrollment in the HAP trial and 4 years after enrollment in the THPP trial. Figures A.1 and A.2 report CONSORT diagrams for the two trials. Since the trials had not been designed for long-run follow-ups, this entailed significant difficulty in locating study participants. We attempted to contact all participants through multiple phone calls or home visits using contact information from the original trials. We asked all original study participants whom we were able to locate whether they were interested in completing a follow-up study. All interested participants were offered the choice of completing the follow-up study activities at at a local study office or in their homes (or nearby location, such as a local temple). Activities took place over two sessions, approximately one week apart. Appendix B details the division of activities across sessions.

Data collection started in December 2018 and ended in March 2020 due to Covid-19 restrictions in India. Until then, we were able to follow up with 589 out of 773 participants (76.2%) across the two trials (Table 1). This follow-up rate is somewhat higher than in Baranov et al. (2020), which might be due the fact that their follow-up study in Pakistan was about seven years after the treatment was delivered. Participation in the control group were slightly higher (78%) compared to the treatment group, though this difference is not statistically significant. Table A.1 shows that follow-up rates were higher for the HAP trial (79%) than for the THPP trial (69%).

Table 1 shows demographics and baseline mental health across treatment and control groups for the pooled study sample. The vast majority of study participants were female (88 percent) and married (81 percent), with an average of just over six years of education. About a third of the sample was employed at enrollment. Both at baseline and in our follow-up sample, we find no statistically significant differences across treatment and control groups in their *baseline* characteristics. An F-test for balance in baseline characteristics cannot be rejected either at enrollment (p=0.91) or in the follow-up sample (p=0.73). This implies that the initial randomization was successful and that attrition from baseline to our follow-up was not differential across groups.

Appendix Table A.1 shows the same analysis separately by trial. As expected given the different screening criteria, PHQ-9 scores at the time of enrollment are higher in the HAP sample (17.8) than in the THPP trial (12.8). For each trial, we again find no evidence of imbalances in baseline characteristics—including baseline PHQ-9 scores—either at baseline or at follow-up.

Our data collection involved a number of experimental tasks to measure beliefs about treatment effects, belief-updating behavior, economic preferences, and labor supply. Apart from a few hypothetical measures, these tasks were financially incentivized: within each task, one choice was randomly selected for payment and the rewards incorporated into the participant's earnings for that session.[15]

[15]Earnings were rounded up to Rs. 400 if less than that amount, and participants also received reimbursement of their

3 Empirical Framework

3.1 Description of empirical framework

Most of our empirical analyses estimate treatment effects on outcomes measured at the participant level using variants of a standard OLS regression framework:

$$y_i = \beta T_i + g(\mathbf{X}_i) + \varepsilon_i, \tag{1}$$

where y_i is the relevant outcome for participant i. T_i is an indicator variable capturing the treatment that participant i was assigned to, and $g(\mathbf{X}_i)$ is a function of a vector of control variables selected by a double machine learning (DML) procedure, described below. β is the key coefficient of interest, capturing the impact of each treatment on the outcome of interest. For all outcomes, we show results for pooling the two trials (our main specification) as well as results for each of the two trials separately.

Following our pre-registration, we report both estimates without controls and using the "double machine learning" (DML) approach from Chernozhukov et al. (2016).[16] We include as potential controls all baseline variables that were collected, a subset of which are shown in Table 1 (categorical variables converted to dummies), plus surveyor fixed effects. In the pooled specification, for variables that were not measured in both trials, we impute the mean value of this variable for participants for whom it is missing, and additionally include a trial fixed effect. Both specifications give very similar point estimates and precision, consistent with the absence of any meaningful imbalances as demonstrated in Section 2.3.

When reporting results for categories of outcomes for which we have multiple measures (e.g. the preference outcomes in Table 5), we also report treatment effects on an index across the measures, following the inverse covariance weighted approach to index construction recommended by Anderson (2008). If a participant is missing one or more index components, we adjust the weights proportionally on the non-missing components, so that the sum of weights is the same for all participants. After constructing the index measure, we normalize it to mean zero, standard deviation one, in the control group of the particular regression of interest.

3.2 Pre-analysis plan and multiple hypothesis testing

Our empirical analysis closely follows our pre-analysis plan.[17] We deviate from this plan in three ways. First, we specified that we would stratify the sample-splitting procedure in the DML estimation by trial. This turned out to be impractical to implement. Since our estimation uses many sample splits, we believe this is very unlikely to make a material difference. Second, we pre-specified that our primary measure of overconfidence would be initial overconfidence before the belief-updating task; later we realized that overconfidence after the task (i.e. after receiving structured feedback) is at least

transportation costs, plus a final bonus of Rs. 200, 2 packets of oil, or an umbrella, for those that completed the whole study.

[16] We use algorithm DML1. We use the Random Forest algorithm to predict outcome and treatment based on controls with a 2-folds sample splitting procedure with 100 splits and 1000 trees.

[17] We posted the pre-analysis plan to the AEA trial registry shortly after the start of our data collection activities: https://doi.org/10.1257/rct.3823-1.0

as interesting, so we report both effects and adjust for multiple comparisons. Third, the elicitation of beliefs about treatment effects was designed later, and does not appear in the pre-analysis plan.[18]

We pre-specified five families of primary outcomes (depression measured by PHQ-9, overconfidence, belief updating, hiring scheme decisions, and preferences). We adjust for multiple comparisons within each family when relevant, and report false discovery rate (FDR) q-values in the regression tables.[19] Secondary outcomes such as employment, consumption, female empowerment, intimate partner violence, and sleep are not corrected for multiple-hypotheses testing and their analysis should be interpreted as exploratory.

3.3 Expert predictions

To quantify how our results compare with current scientific understanding, we conducted surveys of experts in economics and mental health to elicit their forecasts about the treatment effects of psychotherapy on the main outcomes we study (DellaVigna, Pope and Vivalt, 2019). The survey was posted on the Social Science Prediction Platform, which maintains a mailing list of experts (primarily faculty and graduate students in economics and behavioral science). In addition, we solicited participation via emails to experts in psychology and psychiatry. Altogether, 234 experts made forecasts, of which 145 were economists and 41 were experts in mental health (usually psychiatry or psychology). Others were in fields such as public policy, biostatistics, and political science.

For simplicity and to reduce the burden on forecasters, we elicited predictions only for the HAP trial. We chose HAP since this was the larger of our two trials and had a significant short-term impact on depression, making forecasts of long-run effects more meaningful. Forecasters were informed about the design of the study, the baseline levels of depression, and impacts on PHQ-9 scores three and twelve months after enrollment. They were then asked to predict the effects found during our five-year follow-up on (i) PHQ-9 scores, (ii) initial overconfidence and asymmetry in belief-updating, (iii) indices of time, risk and social preferences, and (iv) consumption and employment. The expert predictions are shown in Figure 8 and in Appendix Table A.15, and discussed when presenting results.

4 Impacts on Depression

4.1 Treatment effects on depression

Our main measure of depression is the participant's score on the Patient Health Questionnaire-9 (PHQ-9), a screening tool which asks patients about the frequency of experiencing nine symptoms of depression over the past two weeks (Kroenke, Spitzer and Williams, 2002). This measure is widely used in clinical research, validated in India, and designed to be administered by lay people (Patel et al., 2008;

[18] There are two other minor points of note. First, we specified that our primary analysis would follow an intent-to-treat approach, but that we would report instrumental variables estimates if the "first stage" was sufficiently strong. Our estimates are statistically significant but the F-statistic is below conventional thresholds, so we do not report IV estimates. Second, we pre-specified that we would drop binary outcomes if there was significant bunching (90% or more with the same outcome). This applies to one of our outcomes, which sought to measure default effects.

[19] While results regarding multiple families of outcomes (e.g. overconfidence and belief updating) are presented in the same table for ease of exposition, FDR corrections are done within family as described here.

Manea, Gilbody and McMillan, 2012; Indu et al., 2018). It has also been shown to adequately capture improvements in depression due to clinical interventions, thus mitigating concerns about experimenter demand effects (Löwe et al., 2006; McMillan, Gilbody and Richards, 2010). We use average PHQ-9 scores as a measure of severity of depression and standard thresholds of PHQ-9 scores below 10 and 5 to categorize moderate and mild depression, respectively, following earlier phases of the two trials (Patel et al., 2017; Weobong et al., 2017; Fuhr et al., 2019).[20]

Pooling the two trials, we find evidence of long-run reductions in depression in the treatment group (Table 2, col 2). Specifically, the treatment reduced participants' average PHQ-9 score by 0.85 points on a base of just below 8 (p=0.08), a 0.15 SD improvement. Treatment increases remission from depression by 8 and 11 percentage points, for the the mild and moderate depression screening thresholds, respectively (p=0.03 and p=0.01). Considering each sub-component of the PHQ-9 index, treatment significantly reduces the frequency with which participants report feeling bad about themselves—like a failure, letting down themselves or their families—and of experiencing poor appetite or overeating (Table A.2). Point estimates also suggest reductions in feeling down or hopeless, having poor sleep, and feeling tired and low energy. Consistent with the effects on depression, we find a 0.39 point (or 0.17 SD) increase in a mood score—average self-reported happiness over three days—across the two trials (p=0.04).

The treatment effect on depression is entirely driven by the HAP intervention (Table 2, col 4). HAP reduced PHQ-9 scores five years later by 1.37 points, a 0.23 SD reduction (p=0.02), and increased remission from both mild and moderate depression by 13 percentage points (p<0.01). These treatment effects are concentrated among participants who—five years after enrollment—would have had mild or moderate depression in the absence of treatment. Figure 2 Panel A shows a clear difference in the distributions of HAP treatment and control groups for endline PHQ-9 scores of 12 and below, while the two distributions are nearly indistinguishable for higher PHQ-9 scores. In contrast, we find no significant effect of the THPP treatment on depression (Table 2, col 6), and the distributions of PHQ-9 scores appear identical across the THPP treatment and control groups (Figure 2 Panel B).[21] We also find no systematic evidence of treatment effects varying by baseline characteristics.[22]

Figure 3 Panel B shows the full trajectory of depression over the course of the study, including the previous waves of these trials. As documented in previous work (Patel et al., 2017; Weobong et al., 2017), HAP reduced depression three months and one year after the intervention. Remarkably, the treatment effects one year after the treatment (17 percentage point prevalence reduction) were largely maintained in our follow-up study another four years later. The trajectory of depression resembles the

[20] We report estimates both without controls and from the DML specification. Both specifications give very similar conclusions, we report the slightly more conservative DML estimates in the text.

[21] Due to limited power when comparing across trials, we cannot reject that the THPP and HAP trials had equal treatment effects except for one marginally significant estimate (Table 2, column 7).

[22] We explore heterogeneous treatment effects in two ways. First, as pre-registered, we implement the Chernozhukov et al. (2018) machine learning approach, which searches for combinations of variables that jointly predict heterogeneity in treatment effects. Figure A.4 presents Sorted Group Average Treatment Effects (GATES) for quartiles of the Machine Learning-based proxy predictors. While there appears to be some heterogeneity—the difference between the largest and smallest estimate is quantitatively large—it is noisy and unpredictable. Second, we run regressions interacting the treatment indicators with key baseline variables, and find that only participant's age at baseline predicts treatment effects on depression (Table A.3).

one found in Baranov et al. (2020), though remission rates for the control group in the HAP trial are lower, likely due to their higher severity of depression at enrollment.

The relatively small effects of THPP we report are consistent with the short-run effects reported after our data collection began: Fuhr et al. (2019) find a marginally significant reduction of 1.5 PHQ-9 points three months after the intervention, but the treatment effect is less pronounced and no longer statistically significant six months after the intervention. The control group showed high remission rates, which might explain why THPP did not cause large treatment effects. For instance, 3 months after recruitment, 51 percent of untreated participants have a PHQ-9 score below 5 (the threshold for mild depression) and 77 percent a PHQ-9 score below 10 (the threshold for moderate depression). The muted treatment effects and high remission rates in the control group mirror the findings by Maselko et al. (2020) in a similar study of a THPP intervention in Pakistan.

4.2 Discussion

Arguably the most policy-relevant result in this study is the sustained reduction in depression due to the HAP intervention. This enhances the already favorable cost-effectiveness estimate for HAP, as illustrated in Figure A.7 Panel A. Based on the short-run estimates, the HAP intervention averted 0.4 depression months—i.e. a participant with a PHQ-9 score above 10 in a given month—at 3 months and 1.9 depression months at 1 year.[23] The long-run effects increase this estimate to 9 depression months averted at 5 years, more than quadrupling the overall reduction of months of depression experienced compared to the short-run estimates alone. Given the costs of the treatment of $\$66$ per participant, this implies a cost of at most $\$7.33$ per depression month saved, conservatively assuming no further effects beyond five years and not including the short-run earnings gains and averted health expenditures documented in Patel et al. (2017) and Weobong et al. (2017).

These long-run effects of HAP significantly exceeded expert forecasts. Experts were informed about the short-run effects of HAP on PHQ-9 scores, and asked to predict the effects five years after enrollment, as reported in Figure 8 and Table A.15. The median expert predicted a reduction in PHQ-9 scores of 0.08 SD, with over 90% of experts underestimating the actual estimated effect of 0.23 SD. The median prediction among economists (-0.07 SD) was similar to that of mental health experts (-0.1 SD). Our findings imply that psychotherapy for depression may be even more effective—through having more persistent effects—than experts believe.

The under-estimation of the long-run effects of therapy in our context might be explained by the paucity of similar long-run studies. Figure 1 plots our estimates compared to those from the literature, focusing on non-specialist psychotherapy trials in low- and middle-income countries. A vast majority of studies measure impacts at 12 months or less after enrollment, showing sizable short-run effects on average, with substantial variation across trials. Only Baranov et al. (2020) measures effects at a longer time-horizon than this study (seven years). Like our HAP study, they find significantly lower depression prevalence and severity in the treatment group, with comparable effect sizes (0.18 to 0.22 SD). Our findings complement theirs by studying a general population of adults as opposed to their

[23] These calculations involve linear interpolation of effects between the 3 and 12 month effects, and similarly between the 12 and 60 month effects.

sample of pregnant women. Our findings from the THPP trial instead differ from theirs and resemble another THPP evaluation in Pakistan (Maselko et al., 2020). The difference may be explained by THPP employing a much simpler intervention, delivered by peers.

Why do the HAP and THPP trials have different long-run (and short-run) effects? A striking difference between the two samples is the much higher rate of spontaneous remission from depression in the THPP trial. One reason could be that the HAP trial recruited a sample with higher baseline severity of depression, which tends to involve less spontaneous remission without treatment. HAP trial participants also reported having been depressed for much longer at enrollment than THPP trial participants (43 weeks versus 11 weeks, Appendix Table A.1).[24] Another possibility is that depression in the sample of pregnant women in the THPP trial has a more circumscribed environmental trigger (pregnancy and childbirth), explaining the high rates of remission. In either case, we note that the lack of effects in the THPP trial arguably reduces any concerns that treated participants feel differential social pressure to report improvements in symptoms of depression.[25]

Finally, it is worth discussing *why* therapy had such persistent effects in the HAP trial. As described below, we do not find evidence of improvements in economic well-being, and thus improved material circumstances cannot explain the persistent psychological benefits. Instead, a likely explanation is that participants learned the principles or tools of behavioral activation and employed them to manage future challenges to their mental health. Figure A.6 reports the results of a mediation analysis, which finds that the strongest mediator of the long-run effect on depression is—not surprisingly—the short-run effect on depression. The next strongest mediator is a measure of behavioral activation—the extent to which participants were active and engaged in enjoyable activities. These are precisely the principles and tools taught to participants during the course of the therapy intervention.

5 Perceived Treatment Efficacy

Many people who might benefit from therapy do not seek it out (Cronin, Forsstrom and Papageorge, 2020). A lack of familiarity with therapy and skepticism about its efficacy are commonly stated in surveys (Sapiens Lab, 2021). Especially in settings such as India, where psychotherapy was unavailable until recently, learning by experience (and subsequent diffusion through social networks) may foster demand for therapy.

To test this idea, we elicited study participants' beliefs about the short- and long-run effects of therapy on depression. During our long-run follow-up surveys, each participant was asked about the treatment effects in their own trial for three time horizons: at three or six months, at one year, and at four to five years after the completion of the treatment. We elicited participants' beliefs about the probability of remission from depression—as measured by having a PHQ-9 score below 10—for

[24] However, Figure A.5 shows no evidence of a treatment effect even among those THPP participants with PHQ-9 $>$ 15, as in the HAP trial. Even for this sub-sample, spontaneous remission in the control group remains more common in the THPP trial than in HAP.

[25] More generally, PHQ-9 is widely used as an outcome measure in clinical research, where it has been validated against more in-depth structured diagnostic interviews and has been found to adequately capture improvements in mental health (Löwe et al., 2006; McMillan, Gilbody and Richards, 2010; Beard et al., 2016).

both the treatment and control groups, with the difference being the perceived treatment effect.[26] We incentivized the belief measures for accuracy by comparing their answers to our causal estimates.[27] Figure 4 and Table 3 show the resulting estimates.

Our first result is that control-group participants tend to under-estimate the treatment effects of the highly-effective HAP intervention while over-estimating the effects of the largely ineffective THPP intervention. The HAP control group underestimates the HAP treatment effects at all time horizons, by between 5 and 15 percentage points, a statistically significant difference for the two shorter-run time horizons (Figure 4 Panel A). In contrast, the THPP control group *over*estimates the THPP treatment effects at all time horizons (Figure 4 Panel B). Thus, we cannot expect inexperienced patients—as in the control group—to identify effective treatments without additional guidance. Consistent with this, Table A.3 shows that ex-ante expectations of the efficacy of treatment—measured in the baseline survey using an unincentivized Likert-scale question—do not predict individual treatment effects.

Second, experiencing the treatment increases perceived effectiveness of therapy in the long-run. For both trials, the treatment group's beliefs (dark blue) about the medium and long-run treatment effects are significantly more positive (by 7 to 11 percentage points) than the control group's beliefs (light blue). We find no differences in beliefs about the short-run impacts of therapy, which may reflect difficulty recalling how quickly treatment led to improvements. This latter finding also arguably provides evidence against differential social pressure to state positive beliefs among the treatment group, since we might expect such demand effects to upward-bias responses for all time horizons.

What are the takeaways from these results? First, a common result in both trials is that experiencing psychotherapy makes people more optimistic about its effectiveness. If such changes in beliefs spread in social networks, they might increase demand for treatment. Second, we cannot expect people to identify the most effective treatments by themselves. Indeed, even experience with therapy may not solve this problem, as it causes participants to become more optimistic even about a relatively ineffective treatment. Participants appear to find it difficult to disentangle true treatment effects from the improvements in their mental health that would occur even in the absence of any treatment. This is a challenge patients face in judging healthcare efficacy in many contexts, and suggests a potential role for information provision and regulation of mental health treatments to be explored in future work.

6 Self-confidence and Belief Updating

In the previous section, we provided evidence that therapy durably improves mental health. This included notable changes in patients' self-perceptions: treated participants were significantly less likely to perceive themselves as a failure or to feel they had let themselves or their families down. This

[26] Specifically, we described the program to them, detailing the intervention received by the treatment group and the control group (for HAP participants we described the HAP intervention, for THPP participants we described the THPP intervention). We then asked, for each horizon and for the treatment and control group separately, what the participant believed would be the remission rate in that group. Specifically, we asked, out of 10 randomly selected members of the group, how many would have had their depression "reduced to healthy levels." We then compute the participant's belief about the treatment effect as the difference between these two remission rates. Incentives used a quadratic scoring rule. If selected for payment, the question paid Rs. 75 minus 0.75 times their squared error.

[27] We added the beliefs about treatment effects module a short while after the surveys were launched, so we do not have this measure for all participants. For the same reason, these beliefs are not part of our pre-registration.

kind of self-perception is inherently subjective and broad, and the findings imply that therapy results in a more robust self-image. In this section, we turn to rigorously measuring self-perceptions in an economic domain with an objective measure of truth. In particular, we study self-confidence: people's beliefs about how well they perform at a novel work task relative to others. We find that therapy causes such beliefs about oneself to become more accurate by a□ecting how people's beliefs change in response to informative feedback.

6.1 Measuring confidence and belief updating

We measure self-confidence and belief updating using an experimental paradigm extensively applied in previous work with Western, Educated, Industrialized, Rich, and Democratic (WEIRD) samples (e.g. Eil and Rao, 2011; Ertac, 2011; Buser, Gerhards and Van Der Weele, 2018; Coutts, 2019; Zimmermann, 2020; Mobius et al., 2021). We closely follow the design of Mobius et al. (2021), adapting it for low-numeracy participants by developing detailed instructions augmented with examples and employing a novel belief elicitation method using physical aids in the spirit of Delavande, Giné and McKenzie (2011).

Participants complete an incentivized work task before reporting their beliefs about whether they scored in the top half of a group of participants. They then update these beliefs following noisy but informative feedback. We use this paradigm to study whether the treatments a□ected participants' confidence in their own performance, and the nature of their belief updating. Of particular interest is whether treatment participants put disproportionate weight on positive compared to negative signals, and thus become more overconfidence over time. Such belief-updating patterns have been interpreted as evidence of 'optimism bias' in an attempt to protect one's ego (Mobius et al., 2021).

Work task. Participants are asked to create bracelets using string and beads. During a practice phase, participants learned the quality criteria: a precise number of beads, a knot, and the extra string being cut o□. After this practice phase, participants make bracelets for ten minutes and are paid Rs. 5 for each completed bracelet, subject to quality standards. This task was chosen as it requires dexterity, skill, and concentration, and is thus potentially seen as ego-relevant. It is also relevant to their real economic lives, as creating beaded bracelets for sale is similar to the types of jobs that many of them could pursue in their everyday lives. To further enhance this e□ect, we told participants in advance that, depending on their performance, they may be o□ered a real bracelet-making job after the experiment (see below). As we describe in in Section 8, we find no evidence of di□erences in objectively measured performance across the two treatment groups.

Eliciting beliefs. Upon completing the task, participants are asked to assess their performance relative to nine randomly selected previous study participants.[28] Specifically, we elicited their belief about the probability of being in the top half of a group of 10 participants (including themselves). Participants express their beliefs by allocating water between two containers, represent the two states of the world (top half/bottom half). Thus we use visual aids representing probabilities—as recom-

[28]For the first few participants we sampled from pilot participants, who were not part of the THPP and HAP trials. Once we had sufficiently many comparators, we switched to sampling from the actual trial population. We sampled comparators of the same gender as the participant, to avoid imbalances due to gender differences in performance.

mended by, e.g., Delavande, Giné and McKenzie (2011)—while allowing for continuous subjective beliefs. Beliefs are incentivized using a truncated log scoring rule. Participants could directly see how an allocation of water translated into monetary payo□s in each state of the world, via a labeled scale on the containers.[29]

Noisy signals. After reporting their priors, participants receive five rounds of noisy but informative signals, and can update their reported belief after each signal. Signals are independently generated messages indicating whether the participant scored in the top half of the group ("high" signal, H) or not ("low" signal, L). The message is accurate with probability $\frac{2}{3}$ and inaccurate otherwise.[30] After the bracelet-making task was finished, we explained the belief elicitation, and elicited priors ("Initial"). Then, we explained how signals would work, and gave the opportunity to adjust the initial response ("Adjust"). Then, we revealed each signal one by one, and elicited posteriors following each one, giving us a total of seven elicited beliefs per participant. Participants are incentivized to truthfully report their beliefs at every round, since one of their (prior or posterior) beliefs is randomly chosen to 'count' for the payment of accuracy incentives.

Measuring confidence. We create a measure of confidence by comparing participants' subjective beliefs to objective measures of their relative performance. Using the actual distribution of task performance in our sample, we derive a full-information benchmark for each participant, i.e., their true probability of being in the upper half of a randomly-sampled group of 10. We define initial overconfidence as the di□erence between participants' subjective vs. true probability of being in the top half. After each signal, we update the Bayesian benchmark using Bayes rule, and compute the path of overconfidence as the di□erence between reported posteriors and the updated benchmark.

Estimating belief updating parameters. Closely following Mobius et al. (2021), we use the design to study deviations from Bayesian information processing using a simple linear regression framework. Writing down Bayes rule in logit form yields

$$\ln\left(\frac{\mu_{i,t+1}}{1-\mu_{i,t+1}}\right) = \ln\left(\frac{\mu_{i,t}}{1-\mu_{i,t}}\right) + \ln\left[\frac{\Pr(s_{i,t}|i \text{ in upper half})}{\Pr(s_{i,t}|i \text{ in lower half})}\right], \tag{2}$$

where $\mu_{i,t+1}$ is the Bayesian posterior for participant i being in the upper half given prior $\mu_{i,t}$ and signal $s_{i,t}$. Adding three parametric degrees of freedom leads to a simple structural model:

$$\ln\left(\frac{\mu_{i,t+1}}{1-\mu_{i,t+1}}\right) = \delta \ln\left(\frac{\mu_{i,t}}{1-\mu_{i,t}}\right) + \beta_H \cdot \mathbb{1}(s_{i,t} = H) \cdot \ln\left[\frac{\Pr(H|\text{upper half})}{\Pr(H|\text{lower half})}\right] + \beta_L \cdot \mathbb{1}(s_{i,t} = L) \cdot \ln\left[\frac{\Pr(L|\text{upper half})}{\Pr(L|\text{lower half})}\right],$$

where δ captures the weight on the prior, and β_H and β_L measure the weight on positive and negative

[29] We implemented a truncated log scoring rule, i.e. leaving a container empty corresponds to assigning a very small floor probability ϵ to that state of the world. When a participant puts a fraction p_i of the water in the container corresponding to the true state of the world, incentives are calculated as constant $\cdot(\log(p_i + \epsilon) - \log(\epsilon))$ which can be represented on a visual scale starting at 0 on each container. As long as subjective beliefs assign a probability of at least ϵ to each state of the world, this truncated log scoring rule is strictly proper, i.e. it is incentive compatible for a risk-neutral expected-utility maximizer to report beliefs truthfully (Selten, 1998). While other mechanisms hold additional desirable properties in theory, they require sophisticated computations that are challenging to explain to low-numeracy participants, and complicated mechanisms can backfire (Danz, Vesterlund and Wilson, 2020).

[30] Signals are explained to participants using the image of a die that determines whether the signal is truthful (if the die rolls a 1, 2, 3, or a 4) or not (if the die rolls a 5 or a 6). Actual realizations were generated using the survey software (SurveyCTO), so the enumerator is blind to the true state of the world.

signals, respectively. The models nests Bayesian updating with $\delta = \beta_H = \beta_L = 1$ and can be estimated using OLS and the specific likelihood ratios in our experiment:

$$\ln\left(\frac{\mu_{i,t+1}}{1-\mu_{i,t+1}}\right) = \delta \ln\left(\frac{\mu_{i,t}}{1-\mu_{i,t}}\right) + \beta_H \cdot \ln(2) \cdot \mathbb{1}(s_{i,t} = H) - \beta_L \cdot \ln(2) \cdot \mathbb{1}(s_{i,t} = L) + \epsilon_{it} \quad (3)$$

We focus on the coefficients β_H (response to good news), β_L (response to bad news), and their difference $\beta_H - \beta_L$ which captures asymmetric belief updating. We also consider "conservatism": the value of β obtained by imposing $\beta_H = \beta_L = \bar{\beta}$, i.e., forcing symmetric responses to good and bad news.

Sample restrictions. As pre-registered, our main analysis drops observations with a degenerate prior or posterior, i.e. $\mu_{i,t}, \mu_{i,t+1} \in (0,1)$, since the log likelihood ratios in the regression equation are not defined for degenerate beliefs. In robustness checks, following Mobius et al. (2021), we sequentially drop observations who violate basic rules of Bayesian updating: (i) participants who never update their beliefs in response to feedback; (ii) individual observations without an update; (iii) individual observations that update in the wrong direction (i.e., opposite to the signal), and (iv) participants that *ever* update in the wrong direction.[31]

Job application decision. After participants have completed the beliefs task, they are offered a job application decision: They can either take an outside option of Rs. 300 for sure (about a day's wage), or choose to apply for a job that offers Rs. 3,000 for making 1,000 bracelets over the course of one month. While the job is lucrative for participants with capacity to take on additional work, it is only offered to participants whose performance is in the upper half of the sample.[32] Anyone who apply but is in the lower half of their group is not hired and receives Rs. 0. Thus, the more likely a participant believes that they are in the upper half, the more attractive the opportunity becomes. We discuss impacts on the application decision alongside other labor market outcomes in Section 8.

6.2 Overconfidence and belief updating in the control group

Before discussing the treatment effect of therapy on confidence and belief updating, we first study the beliefs of participants who did *not* receive therapy. This is interesting because, consistent with the theory of depressive realism, some scholars have argued that depression is related to an absence of optimistic belief-updating (Korn et al., 2014). However, evidence in this regard is scarce, and more generally such optimistic biases in belief-updating have been studied exclusively in WEIRD populations (Benjamin, 2019).

Prior to receiving any signal, control-group participants exhibit significant overconfidence relative to the Bayesian benchmark, as illustrated by the light-blue line in Figure 5 Panel A, and the first row of Table 4. On average, participants' priors overestimate their true probability of being in the top half by around 13 percentage points ($p<0.001$). Initial overconfidence is entirely driven by HAP

[31] In practice, surveyors would place each beaker on an electronic scale after each round of feedback to capture the reported beliefs. This leads to very small variations in beliefs even when the participant did nothing, due to measurement error. For the purpose of defining non-updates, we include updates that are below 0.004 in magnitude, which corresponds to 4 milliliters of water, within the measurement error of this weighing procedure.

[32] The average participant made around 5 bracelets in the allotted 10 minutes. Under the conservative assumption of no learning by doing, that would imply around 33 hours' work to make 1,000 bracelets, for which the compensation approximates 10 days' wages.

trial participants, who overestimate their performance by 22 percentage points (p<0.001). In contrast, THPP trial participants are closer to the truth and in fact slightly under-confident. Overconfidence in the control group increases further after receiving noisy but on average truthful signals. After receiving all five signals, control participants overestimated their performance by 16 percentage points (p<0.001), a relative increase of 23 percent. This increase in overconfidence is suggestive evidence of optimistic updating: since signals are informative on average, a Bayesian should trend toward accuracy. The increase in confidence is larger in THPP, where the average participant has moved from slightly under- to slightly overconfident by the end of the task.

Estimating the parameters from equation 3, we find clear evidence of optimistic belief updating in the control group (Table 4 Panel B). The estimate of $\beta_H = 0.73$ for the full sample shows that participants on average update reasonably close to the Bayesian benchmark ($\beta_H = 1$) in response to positive signals (though we can reject $\beta_H = 1$). In contrast, the estimate of $\beta_L = -0.09$ is close to zero, and we cannot reject that on average the control group does not update at all in response to negative signals. The stark difference between these estimates implies that belief updating is highly asymmetric and optimistic (Panel C). We estimate $\beta_H - \beta_L = 0.82$, implying a strikingly strong pattern of optimistic belief updating (p<0.001), much more pronounced than in previous studies with non-depressed, high-income samples (Mobius et al., 2021; Buser, Gerhards and Van Der Weele, 2018).[33]

6.3 Effects of psychotherapy on belief updating

Prior to receiving any signal, we find no evidence of significant differences in overconfidence between treated individuals (dark blue) and control individuals (light blue) (Figure 5 Panel A). The treatment group is slightly less overconfident initially than the control group, though this difference in confidence is not statistically significant (Table 4 Panel A). However, unlike the control group, the treated group's beliefs trend *downward* over the course of the experiment, becoming significantly less overconfident by the end of the task. After receiving all signals, the treated group is 8 percentage points less overconfident than the control group (p=0.02, q=0.05). The treatment effect on overconfidence is similar across the two trials (-0.07 vs. -0.09), though only the estimate for HAP is (marginally) significant.

The estimated belief updating parameters confirm the widening gap in overconfidence between treatment and control participants (Table 4 Panel B and Figure 5 Panel B). Across the two trials, psychotherapy reduces β_H by about a third (0.21 units, p=0.03, q=0.07), while β_L slightly increases (by 0.06, not significant). As a result, we find that optimistic belief updating ($\beta_H - \beta_L$), decreased by 0.27 units (p=0.08, q=0.16), a reduction of about a third compared to the control group mean. Average responsiveness to signals ('conservatism', $\bar{\beta}$) reduces slightly by -0.06 SD (p=0.30,q=0.40). In other words, psychotherapy appears to have made participants treat feedback on their performance substantially more evenhandedly, and be slightly less sensitive to feedback in general. As a result,

[33]Since good and bad news are on average equally likely, it also means that participants are *on average* quite unresponsive to information. Forcing the model to treat positive and negative signals equivalently, we find the average response to signals ("conservatism") is β = 0.29. While there was some difference in initial overconfidence between samples, our estimates of the belief-updating parameters are quantitatively very similar between samples: for HAP we estimate β_H = 0.73 and β_L = 0.13, while for THPP we estimate β_H = 0.67 and β_L = 0.03. This translates into very similar estimates of asymmetry, of 0.85 and 0.69 respectively (both highly significant, $p < 0.001$).

the treatment group exhibits lower levels of overconfidence at the end of the belief updating exercise. Finally, just as control group parameters were similar between trials, the treatment effects on belief updating are quantitatively strikingly similar between trials. For instance, the treatment effect on β_H is -0.22 in the HAP sample (p<0.09, q=0.18) and -0.20 in the THPP sample (p=0.15, q=0.31).

Our main findings of reduced belief updating based on good news and reduced overoptimism in response to treatment are robust to tighter sample restrictions in two main ways (Appendix Table A.7, Figure 6).[34] First, closely following Mobius et al. (2021), we re-estimate equation 3 dropping participants and/or observations who did not update at all. Excluding participants who *never* update at all (Panel B) increases the magnitude of the treatment effect on β_H to -0.28 and that on asymmetry to -0.38. Additionally dropping all instances where a participant did not update relative to their previous guess (Panel C), the magnitude of the effects increases further to $\beta_H = -0.36$ and $\beta_H - \beta_L = -0.50$. Second, excluding participants who update in the wrong direction (Panel D) yields similar estimates of β_H and $\beta_H - \beta_L$ compared to the baseline specification (Panel A). Finally, Panel E drops *all* participants with at least one irrational update—including non-updated and updates in in the wrong direction. In this highly restricted (and thus lower-powered) specification, we find qualitatively similar results, though the estimates become statistically insignificant.[35]

6.4 Discussion

How do these findings relate to theories of psychotherapy and depression? First, our results provide little support for the theory of depressive realism. This influential hypothesis proposes that people with depression are 'sadder but wiser', based on (mixed) cross-sectional evidence that healthy individuals are more overoptimistic than depressed individuals (Alloy and Abramson, 1979; Moore and Fresco, 2012). Closest to our paper, Korn et al. (2014) find a significant cross-sectional correlation of depression symptoms with less optimistic belief updating in a sample of 37 adults. We find only a small, non-significant correlation in the same direction, and the causal effect of therapy—our main contribution—goes in the opposite direction of that predicted by the hypothesis.[36]

Instead, our findings are broadly consistent with the underlying theory and proposed mechanisms of forms of psychotherapy—including CBT, of which behavioral activation is a component—which seek to make patients see themselves in a more realistic light as having both strengths and weaknesses (Beck, 2020). Therapy reduced highly negative self-perceptions in a broad and subjective domain by reducing feelings of being a failure, as described in Section 4. This more robust self-image might have

[34] Appendix Table A.6 explores what types of updating behaviors are contributing to our findings. We measure the frequency of degenerate priors (4%) or posteriors (8%); of non-updates (13% of participants never update, 44% of individual observations are coded as non-updates); of wrong-signed updates (18% of observations). 68% of participants have at least one 'irrational' update (degenerate prior or posterior, non-update or wrong-signed update).

[35] We also measure comprehension and find no meaningful differences in comprehension across treatment groups (Table A.6 Panel G). Participants answered 19 comprehension questions during the introduction of the beliefs task, and answered 15 correctly on average.

[36] Appendix Table A.8 investigates if healthy individuals (those in remission) are more overoptimistic than depressed individuals (those not in remission). To do this, we re-estimate equation (3) separately for the currently depressed, and those in remission (having a PHQ-9 score below 10). We use only control participants for this exercise, since in the treatment group current depression status is confounded with treatment. In line with depressive realism we find that remission is *positively* (but not significantly) associated with overconfidence and asymmetric updating. Therefore, our cross-sectional data are weakly consistent with the hypothesis, while treatment effects go in the opposite direction.

reduced participants' psychological need for overconfidence in the specific novel economic domain we studied (Blanton et al., 2001; Sherman and Cohen, 2006; Kolubinski et al., 2018).

Our findings have substantial potential to shift existing views in the scientific community. The expert surveys we conducted reveal belief in the hypothesis of depressive realism. The median expert predicted that the control group—which displays high rates of depression—would respond symmetrically to negative and positive signals ($\beta_L = \beta_H$).[37] In contrast, we found much stronger belief-updating in response to positive than negative signals in the control group ($\beta_L/\beta_H = -0.13$). The median expert forecast lies outside the 99% confidence interval for this estimate, and every single forecast lies outside the 90% confidence interval.[38] Experts also incorrectly expected treatment to increase initial overconfidence—before receiving any signals—and to increase optimistic updating, presumably through the channel of reduced depression and thus less depressive realism.[39]

Intriguingly, we find similar treatment effects in both trials. This suggests that depression is unlikely to be the key mediator of the impacts on self-confidence and belief updating. A formal mediation analysis presented in Figure A.6 confirms this. Therapy may thus change these economic beliefs about oneself through channels beyond mental health. This finding merits future investigation.

7 Preferences

We next turn to the long-run impacts of therapy on three important economic preferences: patience, altruism, and risk tolerance. These preferences drive behaviors across many important economic domains, such as consumption, investment, and public-goods provision. Existing research has examined the correlation between depression and these preferences. We contribute to this literature by adding evidence on the *causal* effect of psychotherapy for depression.

For each type of preference, we collected real-stakes experimental measures as well as secondary validated survey measures taken from the Global Preference Survey (GPS) of Falk et al. (2016). We also distinguish between specific choices participants make (such as how to split an experimental budget between themselves and another recipient) and survey questions in which participants are asked to report a broader self-perception (e.g. "How willing are you to give to good causes without expecting anything in return?"). We report the effects on each measure and also combine measures of a given preference into a standardized index following Anderson (2008).[40]

Altruism. A number of studies have found that depression negatively correlates with altruistic

[37] For simplicity, we did not separately elicit forecasts about β_H and β_L but instead asked experts to predict their ratio (β_L/β_H = 1), which is an intuitive measure of asymmetry.

[38] A caveat to this analysis is that the expert forecast survey did not allow for negative values of β_L/β_H = 1. However, none of the experts chose the smallest permitted value, suggesting that this is unlikely to have affected the results.

[39] Specifically, experts forecasted a change in the ratio β_L/β_H = −0.22, while we estimated β_L/β_H = +0.06. All forecasts lay outside the 90% confidence intervals. Expert forecasts for treatment effects on initial overconfidence also went in the wrong direction (p=0.01).

[40] The construction of each variable is described in more detail in Appendix Table A.5. Comprehension was high across tasks as my measured by a set of comprehension questions for each task. We additionally attempted to measure susceptibility to "default effects" with a task in which participants received one good and had the option to switch to another at a later date. This outcome showed little variation (over 95% of participants stuck with the default). We therefore drop it in accordance with our pre-analysis plan which specifies dropping binary outcomes if over 90% take the same action.

behavior (Alarcón and Forbes, 2017). Different explanations have been proposed, including that depression takes the pleasure out of altruistic behavior or that depressed people are more focused on their own needs. Experts predicted a modest causal effect of the HAP intervention on altruism, with a median forecast of a 0.1 SD increase in the altruism index.

We collect two measures of altruism. The first is an incentivized dictator game, in which participants choose how much of Rs. 50 to send to another (unknown) participant in the experiment (keeping the remainder for themselves). The secondary measure is a self-assessment survey question from the GPS. This asks, on a scale of 1–10, "How willing are you to give to good causes without expecting anything in return?". The index variable combines these two outcomes.

Table 5 Panel A shows a 0.21 SD effect of treatment on the index of altruism (p=0.01, q=0.04). The effect in the HAP trial is similar to the overall effect (0.25 SD), which is somewhat larger than the median expert forecast of 0.1 SD (42 percent of experts lie outside the 90% CI). Considering the two underlying measures separately, we find a significant increase in the self-assessment of altruism and a non-significant but economically meaningful increase in giving in the dictator game. In the dictator game, the mean giving in the control group was around Rs. 17 out of Rs. 50, in line with behavior in dictator games in many other settings. The treatment group on average gave an additional Rs. 1 (or around 0.1 SD) more, consistent with greater altruism (p=0.25). The treatment group also reported a 0.19 SD greater willingness to do good without expecting anything in return (p<0.05).

Patience. Theoretically, the effect of depression on patience is ambiguous. For example, the anhedonia often experienced in depression may reduce the pleasure of immediate consumption, thus making people appear more patient when trading off immediate and future consumption (Lempert and Pizzagalli, 2010). Alternatively, depression might make it difficult to attend to the future, reducing patience (Keller et al., 2019). The expert forecasts indicate a modest belief that the HAP intervention would increase patience, with a median forecast of a 0.12 SD increase in the patience index variable.

Our incentivized measure of patience is a "saving a note" task. We gave each participant a Rs. 100 bank note at the first session, and told them that if they could show the exact same note (with matching serial number) at the second meeting, they would receive a Rs. 30 bonus. Returning the note is a measure of participants' patience and/or their ability to resist temptations.[41] As secondary unincentivized measures, we asked two self-assessment survey questions and a discount-parameter elicitation using hypothetical money-earlier-or-later decisions. The survey questions asked, on a scale of 1–10, to what extent the participant was willing "to give up something that is beneficial for you today in order to benefit more from that in the future," and willing "to complete tasks at the earliest, and not leave them for later/postpone them". The discounting elicitation uses a "ladder" design that, through a sequence of choices between hypothetical sooner or later monetary amounts, finds a narrow range for the participant's indifference point. We elicited discounting between money today or in 12 months' time, and between 12 and 24 months' time.[42]

[41] We chose this task since implementing future-dated monetary payments over traditional time-frames used in money earlier or later designs was difficult to implement in this context. As with other monetary discounting tasks, behavior in the saving-a-note task could also capture time-variation in liquidity constraints (Cohen et al., 2020).

[42] From these choices, we compute present and future discount factors δ_a and δ_b (under a linear utility approximation), as well as a present bias parameter $\beta = {}^{\delta_b}/_{\delta_a}$. For the 12 versus 24 month question, we increased all monetary amounts by 20 percent relative to the today versus 12 month question. The staircase design does not allow for inconsistent

Table 5 Panel B shows a 0.18 SD treatment effect on the patience index—indicating higher patience—which combines the above measures (p=0.03, q=0.05). The effect in the HAP trial is nearly identical at 0.17 SD, and is close to the median expert forecast of 0.12 SD. While the point estimates for all the patience measures are positive, the impact on the index is largely driven by the two self-assessments. In particular, participants are 0.24 SD more likely to state that they are willing to give something up for future benefits (p<0.01). In contrast, we find small and insignificant impacts on whether participants return the bank notes, and on the discounting parameters inferred from the money-earlier-or-later questions. While 79 percent of control group participants returned the bank note, the treatment increased this fraction by 2 percentage points (not significant). In the hypothetical discounting task, participants discount the future heavily: δ_a and δ_b are both around 0.6. They are slightly future-biased on average, with mean β equal to 1.08.[43] The treatment effects on the discount factors are small, but we cannot rule out moderate increases in patience of the order of 0.2 SD.

Risk and loss tolerance. Some scholars have argued that depression reduces risk tolerance (see Kamstra, Kramer and Levi (2003) for a discussion). Proposed mechanisms include that depression reduces "sensation-seeking," which is in turn associated with risk-taking. In addition, depression is associated with pessimism, which might make risks less attractive. Recent correlational evidence, however, casts doubt on this hypothesis (Cobb-Clark, Dahmann and Kettlewell, 2020). Experts did not anticipate meaningful effects of the HAP intervention on risk preferences, with a median forecast of a 0.05 SD increase in the index of risk tolerance.

Our two incentivized measures use lottery choice lists to elicit risk and loss tolerance.[44] The risk tolerance task elicits the monetary amount Y that makes the participant indifferent between receiving Rs. 100 for sure versus a 50-50 lottery over winning Rs. Y/Rs. 200. The smaller is Y, the more risk the participant is willing to tolerate (and $Y = 0$ corresponds to the risk-neutral case). Similarly, the loss tolerance task elicits the monetary amount Z at which the participant is willing to accept a 50-50 lottery over gaining Rs. 100/losing Rs. Z, with losses to be deducted from the participant's show-up fee. In this case the *larger* is Z the more loss-tolerant the participant is (and $Z = 100$ corresponds to risk neutrality with no loss aversion). Our unincentivized measure is taken from the GPS and asks "In general, how willing you are to choose uncertain outcomes in real life?", answered on a 1–10 scale. Similarly-phrased questions have been used in many other settings and have been shown to predict real-world risky decisions well across contexts (Dohmen et al., 2011).

choice (defined as a participant that rejects a smaller amount but accepts a larger amount). Due to an error in the survey implementation, three monetary amounts in the 12/24 month instrument were calibrated incorrectly, permitting inconsistent responses. We find that 8 percent of participants are inconsistent, and code their values of δ_b and β as missing.

[43] These estimates are comparable in magnitude to those from Bauer, Chytilová and Morduch (2012)'s study of self-help group members in Karnataka, India. They estimate mean three-month discount rates of 0.244 between the present and three months' time, and 0.193 between 12 and 15 months' time. These correspond to annual discount factors $\delta_a = (1/1.244)^4 = 0.42$ and $\delta_b = (1/1.193)^4 = 0.49$. However, their estimates indicate present bias on average whereas we find modest future bias.

[44] The choice lists begin with the most favorable option and work toward the least favorable. We allowed multiple switching on these choice lists, and take the first switch point as the participant's choice. If participants made inconsistent choices this was pointed out to them and they were given an opportunity to reconsider. We estimate the indifference point as the midpoint of the last-accepted and first-rejected option. If the participant rejected (accepted) all options, we take the first (last) value as their indifference point.

We find small and non-significant effects on all three measures as well as a small (0.05 SD) increase in the overall index of risk tolerance (Table 5 Panel C). Psychotherapy increases the index measure by a small and statistically insignificant 0.03 SD, very similar to the median expert prediction of 0.05 SD. In the risk tolerance task, the control group on average chooses $Y=53$. This amount is about Rs. 2 lower in the treatment group, indicating slightly higher risk tolerance (not significant). In the loss tolerance task, the control group chooses on average $Z=73$, and this amount is about Rs. 1 higher in the treatment group, indicating slightly higher loss tolerance (not significant). Consistent with these choices, treated individuals on average assess their willingness to choose uncertain outcomes by 0.05 SD higher (not significant).

7.1 Discussion

Our findings add to the nascent literature estimating the causal effects of psychotherapy on economic preferences. Closest to our work is Blattman, Jamison and Sheridan (2017), which shows that CBT designed to reduce impulsive behavior successfully increased patience and reduced violent behavior in a population of criminally-engaged young men Liberia. Their intervention differs substantially from those we study, since it did not target depression and did not improve mental health. Also related is Angelucci and Bennett (2021), who show mixed evidence of pharmacological treatment of depression affecting risk tolerance.[45] Therapy and pharmacological treatments for depression may, however, operate through different mechanisms.

The positive effects of psychotherapy on the aggregate index measures of patience and altruism are broadly in line with expert predictions. However, it is important to note that these effects are driven more by changes in self-assessments than by changes in actual experimental choices. One interpretation of these findings is that therapy did not affect the extent to which people actually behaved patiently or altruistically, but changed how they *interpreted* their own behavior, in line with a more robust self-image. While patience and altruism are typically viewed as normatively desirable, the valence of risk tolerance is less clear, which might explain the smaller effects observed for this outcome.

However, we cannot rule out moderate effects on real behaviors in line with the changed self-assessments, especially in the case of altruism. Self-evaluations may also reflect real behavior across a broader set of domains and a longer time-frame than captured by the experimental tasks. An alternative interpretation is thus that psychotherapy changed not just self-perceptions, but also behaviors, in the direction of greater patience and altruism.

As with the self-confidence outcomes, the pattern of results suggest that therapy affected patience and altruism through channels other than current levels of depression. The effects were similar in both trials, despite only HAP showing an impact on depression.[46] Moreover, a correlational analysis reported in Appendix Table A.9 shows weak associations of current depression status with each of

[45] Specifically, Angelucci and Bennett (2021) find that being prescribed antidepressants does not affect a lottery-based measure of risk tolerance or self-assessed risk attitudes, similar to our findings. However, they do find evidence of reductions in self-reported risky behaviors such willingness to ride a motorcycle without a helmet.

[46] The most notable difference is that we do see suggestive evidence of a treatment effect in the Saving-a-Note task in the THPP trial. Here, only 71 percent of untreated participants saved the note, increasing by 12 percentage points among the treated and significant at the 10 percent level.

these preferences.[47] A mediation analysis reported in Figure A.6 confirms that changes in current PHQ-9 scores do not explain the effects. Intriguingly, a measure of behavioral activation—being active and engaged in pleasurable activities—is the strongest mediator of the effects on patience and especially altruism, explaining a majority of each effect. This suggests that behavioral activation—the focus of the therapy interventions we study–both improves mental health and independently affects self-perceptions.

8 Work, Consumption, and Other Outcomes

We find no evidence of significant impacts of the treatments on work-related outcomes, including survey-based measures of labor supply and employment (Table 6, Panel A) and revealed-preference measures of willingness to take on paid work and productivity (Panel B). In line with these results, we also find no impacts on consumption (Panel C). Finally, we find some evidence of improved sleep due to the treatment but no impacts on other exploratory outcomes such as female empowerment, IPV, or loneliness (Panel D).[48]

Labor supply and employment. We detect no significant impacts on any of the survey measures of labor supply and employment (Table 6 Panel A), including whether people were engaged in paid work, how many hours of paid work they did in the past week, or how much they earned in the past month. Similarly, we find no significant difference in the fraction of people who say they are available to take on an employment opportunity. Among the unemployed, we find no difference in the number of job search hours per week. These results are consistent with Baranov et al. (2020), who also find no effects of psychotherapy for depression on employment in a sample of women in Pakistan. A caveat is that, due to limited power, we cannot rule out modest improvements (or reductions). For instance, the 95% confidence interval for effects on employment rates extends from -7 to +5 percentage points, compared to a control-group mean of 25%.

Table 6 Panel B shows no evidence of impacts on revealed-preference measures of productivity and willingness to take on paid work (i.e. labor supply at the margin). First, we find no evidence of impacts on the number of bracelets made as part of the work task used to measure confidence and belief updating (see Section 6). Since this incentivized task involves the kind of manual skills often required in work available to our study participants, this outcome arguably provides a relevant measure of work productivity. Second, we find no impacts of the treatment on incentivized reservation wages—the minimum wage participants were willing to work at—when asked to make a large, fixed number of bracelets from home (using as much time as needed), regardless of their performance in the task. Finally, we also find no impacts on the take-up of the performance-based hiring scheme. Recall that signing up for this hiring scheme was potentially lucrative for participants who performed above

[47] While patience and risk tolerance indices are weakly negatively associated with current depression, altruism is instead weakly positively associated. The coefficients for patience and altruism are less than half as large in magnitude as the our estimated treatment effects, which further goes to suggest that the treatment effects are not well-explained by changes in contemporaneous depression.

[48] We attempted to measure schooling attainment of school age and adult children, and reading and writing ability of children. We find no evidence of meaningful effects, but this is not particularly surprising because of high dispersion in children's ages in the sample: many were too young to be in school, or too old to be meaningfully affected.

the median in the sample. Given the treatment effects on final confidence (Table 4 Panel A), one may have expected treatment to reduce take-up of the hiring scheme. It did not.[49]

Expenditures. Given the lack of impacts on labor supply, earnings, and productivity, it is perhaps not surprising that we observe little impact on expenditures and consumption (Table 6 Panel C). We find no significant impacts on overall consumption or on individual categories such as food or durable goods. We can rule out increases in monthly expenditures greater than 10 percent of control-group levels. Improved mental health also did not translate into significant reductions in other health expenditures, unlike in the short-run study (Patel et al., 2017; Weobong et al., 2017).

Female empowerment and IPV. We also find no significant treatment effects on female empowerment and intimate partner violence (Table 6 Panel D). An index of female empowerment variables shows a positive point estimate of 0.06 SD across the entire sample, but this is not statistically significant, and few of the constituent variables show sizable effects (Appendix Table A.11). The estimated effects on female empowerment are less precise and smaller than in Baranov et al. (2020), though we cannot reject the effects found in their study. Similarly, while we find some suggestive evidence of reduced intimate partner violence (IPV) in the treatment group, these estimates are not statistically significant. Across the entire sample (restricted to female participants), the corresponding index decreases by 0.12 SD, driven by a significant 67 percent reduction in reported instances of forced sex, from 9 percent to 3 percent of women (Appendix Table A.12).

Sleep, loneliness, and locus of control. We find a statistically significant positive effect of 0.2 SD on an index of three self-reported sleep-related variables: the number of hours slept in the preceding night, sleep quality, and the number of hours spent in bed but not asleep (entering the index negatively). The positive effect on the index is driven by a 16-minute increase in sleep duration per night, and a 0.20 SD improvement in sleep quality. This provides rare evidence of an intervention capable of improving sleep quality in developing-country settings (Bessone et al., 2021; Rao et al., 2021). In contrast, we find no evidence of impacts on loneliness or locus of control.

8.1 Discussion

The above findings contribute to the literature studying the causal impacts of therapy on work-related outcomes. Previous work has found evidence of short-run impacts of therapy on brief survey measures of labor supply, particularly self-reported days of work missed (Patel et al., 2017; Lund et al., 2010). We contribute to this evidence base by studying longer-run impacts and by using revealed-preference measures of productivity and labor supply in addition to survey measures of employment and earnings. Our findings suggest that therapy does not have long-run impacts on work-related outcomes in this context. While our statistical power is limited, point estimates are small across outcomes and we can rule out moderate-sized positive effects. We interpret our findings as showing both a lack of real-world labor supply responses—which could be attributed to low labor demand or to

[49] One interpretation of this result is that people's labor supply decisions are constrained by other factors such as social norms or childcare, thus limiting any potential impacts of therapy on choices in the hiring scheme. Alternatively, people might optimistically update their beliefs to feel good about themselves but then revert to their priors in higher-stakes choices, as discussed in Mobius et al. (2021). We find some evidence in support of this hypothesis (Appendix Table A.10).

constraints on women's work outside the home—but also no effects on willingness to take on flexible work from home.

The null effect on employment contrasts with the forecasts of experts. The median expert predicted a 0.12 SD increase in employment. The true empirical estimate of -0.01 SD was equal to the 13th percentile of the expert forecasts. The median expert predicted more modest effects ($+0.06$ SD) on consumption, which was close to the true empirical estimate of 0.05 SD. Ultimately, even a very effective psychotherapy intervention on its own was not sufficient to unleash greater labor supply and earnings, and help recipients escape poverty.

9 Conclusion

This study presents the long-run effects of two trials of psychotherapy for depression. Most directly relevant for policy, we document that the mental-health benefits of the Healthy Activity Program—an inexpensive, scalable therapy delivered by a non-specialist—are remarkably persistent. Therapy increased remission from depression by 13 percentage points five years after the brief intervention was delivered. Accounting for these long-run impacts further increases estimates of the (already-high) cost effectiveness of such therapies. We calculate that being offered the HAP treatment averted 9.1 months of depression on average over five years, at a cost of at most $7.33 per month of depression averted. This result reinforces calls for expanding access to psychotherapy in India and similar contexts worldwide (Singla et al., 2017).

As the supply of therapy grows in the developing world, studying potential demand-side barriers becomes increasingly important. Low perceptions of treatment efficacy could be a barrier to seeking treatment, possibly due to lack of familiarity. We document that receiving therapy increases people's beliefs regarding the efficacy of therapy. Future work may explore whether such changes in beliefs are transmitted in social networks and lead to increased demand for therapy among peers. Our results also suggest that regulation and/or information interventions that help people identify effective treatments could be beneficial.

This paper also takes a first step towards providing a behavioral-science view of psychotherapy, documenting its long-run effects on beliefs and preferences. We find that therapy durably changes recipients' beliefs about themselves. It reduces extremely negative self-perceptions which are a common symptom of depression. It also increases participants' self-assessed levels of patience and altruism. However, therapy does not simply cause patients to see themselves more positively across the board. It also causes them to see themselves more *accurately* in certain domains. When faced with a novel work opportunity, therapy reduced overconfidence by making people react to feedback more evenhandedly. This finding is arguably consistent with the theory and goals of CBT (Beck, 2020). Of course, our study only captures certain aspects of beliefs and economic preferences, and omits many other aspects of how people perceive themselves and the world, which future research should explore.

Finally, this paper reports that even a highly-effective psychotherapy intervention which durably reduced depression did not translate into increases in employment and earnings among low-income adults in India. At least in this context, therapy did not prove to be a tool for long-term reductions in

poverty. One explanation could be that other constraints such as social norms depress women's work outside the household in the study context (Fletcher, Pande and Moore, 2017). Pairing psychotherapy with other interventions might be necessary to unlock the economic benefits of improved mental health (Bossuroy et al., 2022). Yet the remarkably persistent effects of psychotherapy on mental health, together with the short-run reductions in health expenditures documented in previous work (Patel et al., 2017; Weobong et al., 2017), already make a strong policy case for such psychotherapy interventions in low-income contexts.

References

Alarcón, Gabriela, and Erika E. Forbes. 2017. "Prosocial Behavior and Depression: a Case for Developmental Gender Di□erences." Child and Developmental Psychiatry, 4(2): 117–127.

Albert, Paul R. 2015. "Why is depression more prevalent in women?" Journal of psychiatry & neuroscience: JPN, 40(4): 219.

Alloy, L.B., and L.Y. Abramson. 1979. "Judgment of contingency in depressed and nondepressed students: Sadder but wiser." Journal of Experimental Psychology, 108(4): 441–485.

Anderson, Michael L. 2008. "Multiple Inference and Gender Di□erences in the E□ects 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, Manuela, and Daniel Bennett. 2021. "The Economic Impact of Depression Treatment in India." IZA DP No. 14393.

Baranov, Victoria, Sonia Bhalotra, Pietro Biroli, and Joanna Maselko. 2020. "Maternal Depression, Women's Empowerment, and Parental Investment: Evidence from a Randomized Controlled Trial." American Economic Review, 110(3): 824–59.

Barker, Nathan, Gharad T Bryan, Dean Karlan, Angela Ofori-Atta, and Christopher R Udry. 2021. "Mental Health Therapy as a Core Strategy for Increasing Human Capital: Evidence from Ghana." National Bureau of Economic Research.

Barth, Jürgen, Thomas Munder, Heike Gerger, Eveline Nüesch, Sven Trelle, Hansjörg Znoj, Peter Jüni, and Pim Cuijpers. 2016. "Comparative e□cacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis." Focus, 14(2): 229–243.

Bauer, Michal, Julie Chytilová, and Jonathan Morduch. 2012. "Behavioral Foundations of Microcredit: Experimental and Survey Evidence from Rural India." American Economic Review, 102(2): 1118–1139.

Beard, C, KJ Hsu, LS Rifkin, AB Busch, and T Björgvinsson. 2016. "Validation of the PHQ-9 in a psychiatric sample." Journal of A□ective Disorders, 193: 267–273.

Beck, Judith S. 2020. Cognitive behavior therapy: Basics and beyond. Guilford Publications.

Bénabou, Roland, and Jean Tirole. 2002. "Self-confidence and personal motivation." The quarterly journal of economics, 117(3): 871–915.

Benjamin, Daniel J. 2019. "Errors in probabilistic reasoning and judgment biases." Handbook of Behavioral Economics: Applications and Foundations 1, 2: 69–186.

Bessone, Pedro, Gautam Rao, Frank Schilbach, Heather Schofield, and Mattie Toma. 2021. "The economic consequences of increasing sleep among the urban poor." The Quarterly Journal of Economics, 136(3): 1887–1941.

Blanton, Hart, Brett W. Pelham, Tracy DeHart, and Mauricio Carvallo. 2001. "Overconfidence as Dissonance Reduction." Journal of Experimental Social Psychology, 37(5): 373–385.

Blattman, Christopher, Julian C Jamison, and Margaret Sheridan. 2017. "Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia." Am. Econ. Rev., 107(4): 1165–1206.

Bossuroy, Thomas, Markus Goldstein, Dean Karlan, Harounan Kazianga, William Pariente, Patrick Premand, Catherine Thomas, Christopher Udry, Julia Vaillant, and Kelsey Wright. 2022. "Tackling psychosocial and capital constraints to alleviate poverty." Nature.

Buser, Thomas, Leonie Gerhards, and Joël Van Der Weele. 2018. "Responsiveness to feedback as a personal trait." Journal of Risk and Uncertainty, 56(2): 165–192.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. "Double/debiased/neyman machine learning of treatment e□ects." American Economic Review, 107(5): 261–65.

Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2016. "Double/debiased machine learning for treatment and causal parameters." arXiv preprint arXiv:1608.00060.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Ivan Fernandez-Val. 2018. "Generic machine learning inference on heterogenous treatment e□ects in randomized experiments." National Bureau of Economic Research.

Cobb-Clark, Deborah A., Sarah C. Dahmann, and Nathan Kettlewell. 2020. "Depression, Risk Preferences and Risk-taking Behavior." Journal of Human Resources, 0419–10183R1.

Cohen, Jonathan, Keith Marzilli Ericson, David Laibson, and John Myles White. 2020. "Measuring time preferences." Journal of Economic Literature, 58(2): 299–347.

Coutts, Alexander. 2019. "Good news and bad news are still news: Experimental evidence on belief updating." Experimental Economics, 22(2): 369–395.

Cronin, Christopher J, Matthew P Forsstrom, and Nicholas W Papageorge. 2020. "What good are treatment e□ects without treatment? Mental health and the reluctance to use talk therapy." National Bureau of Economic Research.

Cuijpers, Pim, Erica Weitz, Eirini Karyotaki, Judy Garber, and Gerhard Andersson. 2015. "The e□ects of psychological treatment of maternal depression on children and parental functioning: a meta-analysis." Eur. Child Adolesc. Psychiatry, 24(2): 237–245.

Cuijpers, Pim, Filip Smit, Ernst Bohlmeijer, Steven Hollon, and Gerhard Andersson. 2010. "E□cacy of cognitive-behavioural therapy and other psychological treatments for adult depression: meta-analytic study of publication bias." The British Journal of Psychiatry, 196: 173–178.

Cuijpers, Pim, Ioana A Cristea, Eirini Karyotaki, Mirjam Reijnders, and Marcus JH Huibers. 2016. "How e□ective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence." World psychiatry, 15(3): 245–258.

Danz, David, Lise Vesterlund, and Alistair J Wilson. 2020. "Belief elicitation: Limiting truth telling with information on incentives." National Bureau of Economic Research.

Delavande, Adeline, Xavier Giné, and David McKenzie. 2011. "Measuring subjective expectations in developing countries: A critical review and new evidence." Journal of development economics, 94(2): 151–163.

DellaVigna, Stefano, Devin Pope, and Eva Vivalt. 2019. "Predict science to improve science." Science, 366(6464): 428–429.

Dohmen, Thomas, Armin Falk, David Hu□man, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." Journal of the European Economic Association, 9(3): 522–550.

Eil, David, and Justin M Rao. 2011. "The Good News-Bad News E□ect: Asymmetric Processing of Objective Information about Yourself." American Economic Journal: Microeconomics, 3(2): 114–138.

Ertac, Seda. 2011. "Does self-relevance a□ect information processing? Experimental evidence on the response to performance and non-performance feedback." Journal of Economic Behavior & Organization, 80(3): 532–545.

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Hu□man, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." The Quarterly Journal of Economics, 133(4): 1645–1692.

Falk, Armin, Anke Becker, Thomas Dohmen, David Hu□man, and Uwe Sunde. 2016. "The preference survey module: A validated instrument for measuring risk, time, and social preferences." IZA Discussion Paper No. 9674.

Fletcher, Erin, Rohini Pande, and Charity Maria Troyer Moore. 2017. "Women and work in India: Descriptive evidence and a review of potential policies."

Friedrich, M.J. 2017. "Depression is the leading cause of disability around the world." JAMA, 317(15): 1517.

Fuhr, Daniela, Benedict Weobong, Anisha Lazarus, Fiona Vanobberghen, Helen A Weiss, Daisy Radha Singla, Hanani Tabana, Ejma Afonso, Aveena De Sa, Ethel D'Souza, Akankasha Joshi, Priya Korgaonkar, Revathi Krishna, Price LeShawndra, Atif Rahman, and Vikram Patel. 2019. "Delivering the Thinking Healthy Programme for perinatal depression through peers: an individually randomised controlled trial in India." Lancet Psychiatry, 6: 115–127.

Haushofer, Johannes, Robert Mudida, and Jeremy P Shapiro. 2020. "The comparative impact of cash transfers and a psychotherapy program on psychological and economic well-being." National Bureau of Economic Research.

Heller, Sara B, Anuj K Shah, Jonathan Guryan, Jens Ludwig, Sendhil Mullainathan, and Harold A Pollack. 2017. "Thinking, fast and slow? Some field experiments to reduce crime and dropout in Chicago." The Quarterly Journal of Economics, 132(1): 1–54.

Huang, Rurui, Chunli Yan, Yumei Tian, Beimei Lei, Dongqi Yang, Dan Liu, and Jun Lei. 2020. "E□ectiveness of peer support intervention on perinatal depression: A systematic review and meta-analysis." Journal of A□ective Disorders, 276(1): 788–796.

Indu, Pillaveetil Sathyadas, Thekkethayyil Viswanathan Anilkumar, Krishnapillai Vijayakumar, K.A. Kumar, P. Sankara Sarma, Saradamma Remadevi, and Chittaranjan Andrade. 2018. "Reliability and validity of PHQ-9 when administered by health workers for depression screening among women in primary care." Asian Journal of Psychiatry, 37: 10–14.

Kamstra, Mark J, Lisa A Kramer, and Maurice D Levi. 2003. "Winter Blues: A SAD Stock Market Cycle." American Economic Review, 93(1): 324–343.

Keller, Arielle S, John E Leikauf, Bailey Holt-Gosselin, Brooke R Staveland, and Leanne M Williams. 2019. "Paying attention to attention in depression." Translational psychiatry, 9(1): 1–12.

Kessler, Ronald C, and Philip S Wang. 2009. "Epidemiology of depression." In I. H. Gotlib & C. L. Hammen (Eds.), Handbook of depression. The Guilford Press., 5–22.

Kolubinski, Daniel C., Daniel Frings, Ana V. Nikčević, Jacqueline A. Lawrence, and Marcantonio M. Spada. 2018. "A systematic review and meta-analysis of CBT interventions based on the Fennell model of low self-esteem." Psychiatry Research, 267: 296–305.

Korn, C W, T Sharot, H Walter, H R Heekeren, and R J Dolan. 2014. "Depression is related to an absence of optimistically biased belief updating about future life events." Psychol. Med., 44(3): 579–592.

Köszegi, Botond. 2006. "Ego utility, overconfidence, and task choice." Journal of the European Economic Association, 4(4): 673–707.

Kroenke, Kurt, Robert L Spitzer, and Janet B W Williams. 2002. "The PHQ-9: a new depression diagnostic and severity measure." Psychiatr. Ann., 32(9): 509–515.

Lempert, Karolina M., and Diego A. Pizzagalli. 2010. "Delay discounting and future-directed thinking in anhedonic individuals." Journal of Behavior Therapy and Experimental Psychiatry, 41(3): 258–264.

Löwe, Bernd, Irini Schenkel, Caroline Carney-Doebbeling, and Claus Göbel. 2006. "Responsiveness of the PHQ-9 to psychopharmacological depression treatment." Psychosomatics, 47(1): 62–67.

Lund, Crick, Alison Breen, Alan J Flisher, Ritsuko Kakuma, Joanne Corrigall, John A Joska, Leslie Swartz, and Vikram Patel. 2010. "Poverty and common mental disorders in low and middle income countries: A systematic review." Soc. Sci. Med., 71(3): 517–528.

Lund, Crick, Kate Orkin, Marc Witte, Thandi Davies, Johannes Haushofer, Judy Bass, Paul Bolton, Sarah Murray, Laura Murray, Wietse Tol, Graham Thornicroft, and Vikram Patel. 2021. "The economic e□ects of mental health interventions in low and middle-income countries." Working Paper.

Malmendier, Ulrike, and Geo□rey Tate. 2005. "CEO overconfidence and corporate investment." The journal of finance, 60(6): 2661–2700.

Manea, Laura, Simon Gilbody, and Dean McMillan. 2012. "Optimal cut-o□ score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis." Canadian Medical Association Journal, 184(3): E191–E196.

Maselko, Joanna, Siham Sikander, Elizabeth L Turner, Lisa M Bates, Ikhlaq Ahmad, Najia Atif, Victoria Baranov, Sonia Bhalotra, Amina Bibi, Tayyaba Bibi, Samina Bilal, Pietro Biroli, Esther Chung, John A Gallis, Ashley Hagaman, Anam Jamil, Katherine LeMasters, Karen O'Donnell, Elissa Scherer, Maria Sharif, Ahmed Waqas, Ahmed

Zaidi, Sha□aq Zulfiqar, and Atif Rahman. 2020. "E□ectiveness of a peer-delivered, psychosocial intervention on maternal depression and child development at 3 years postnatal: a cluster randomised trial in Pakistan." The Lancet Psychiatry, 7(9): 775–787.

McMillan, Dean, Simon Gilbody, and David Richards. 2010. "Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods." Journal of a□ective disorders, 127(1-3): 122–129.

Mobius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. 2021. "Managing Self-Confidence: Theory and Experimental Evidence." Management Science (forthcoming).

Moore, Don A, and Paul J Healy. 2008. "The trouble with overconfidence." Psychological review, 115(2): 502.

Moore, Michael T, and David M Fresco. 2012. "Depressive realism: A meta-analytic review." Clinical psychology review, 32(6): 496–509.

Patel, Vikram, Benedict Weobong, Helen A Weiss, Arpita Anand, Bhargav Bhat, Basavraj Katti, Sona Dimidjian, Ricardo Araya, Steve D Hollon, Michael King, Lakshmi Vijayakumar, A-La Park, David McDaid, Terry Wilson, Richard Velleman, Betty R Kirkwood, and Christopher G Fairburn. 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." Lancet, 389(10065): 176–185.

Patel, Vikram, Gregory Simon, Neerja Chowdhary, Sylvia Kaaya, and Ricardo Araya. 2009. "Packages of Care for Depression in Low- and Middle-Income Countries." PLoS Med., 6(10): e1000159.

Patel, Vikram, Helen Weiss, Neerja Chowdhary, Smita Naik, Sulochana Pednekar, Sudipto Chatterjee, Mary De Silva, Bhargav Bhat, Ricardo Araya, Michael King, Gregory Simon, Helen Verdeli, and Betty Kirkwood. 2010. "E□ectiveness of an intervention led by lay counsellors for depressive and anxiety disorders in primary care in Goa, India (MANAS): a cluster randomised controlled trial." Lancet, 376(9758): 2086–2095.

Patel, Vikram, Neerja Chowdhary, Atif Rahman, and Helen Verdeli. 2011. "Improving access to psychological treatments: Lessons from developing countries." Behav. Res. Ther., 49(9): 523–528.

Patel, Vikram, Shekhar Saxena, Crick Lund, Graham Thornicroft, Florence Baingana, Paul Bolton, Dan Chisholm, Pamela Y Collins, Janice L Cooper, Julian Eaton, et al. 2018. "The Lancet Commission on global mental health and sustainable development." The Lancet, 392(10157): 1553–1598.

Patel, Vikram, Shuiyuan Xiao, Hanhui Chen, Fahmy Hanna, A T Jotheeswaran, Dan Luo, Rachana Parikh, Eesha Sharma, Shamaila Usmani, Yu Yu, Benjamin G Druss, and Shekhar Saxena. 2016. "The magnitude of and health system responses to the mental health treatment gap in adults in India and China." Lancet, 388(10063): 3074–3084.

Patel, V, R Araya, N Chowdhary, M King, B Kirkwood, S Nayak, G Simon, and H A Weiss. 2008. "Detecting common mental disorders in primary care in India: a comparison of five screening questionnaires." Psychol. Med., 38(2): 221–228.

Rahman, Atif, Abid Malik, Siham Sikander, Christopher Roberts, and Francis Creed. 2008. "Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: a cluster-randomise controlled trial." Lancet, 372: 902–909.

Rahman, Atif, Jane Fisher, Peter Bower, Stanley Luchters, Thach Tran, M Taghi Yasamy, Shekhar Saxena, and Waquas Waheed. 2013. "Interventions for common perinatal mental disorders in women in low- and middle-income countries: a systematic review and meta-analysis." Bulletin of the World Health Organization, 91: 593–601.

Rao, Gautam, Susan Redline, Frank Schilbach, Heather Schofield, and Mattie Toma. 2021. "Informing sleep policy through field experiments." Science, 374(6567): 530–533.

Ridley, Matthew, Gautam Rao, Frank Schilbach, and Vikram Patel. 2020. "Poverty, depression, and anxiety: Causal evidence and mechanisms." Science, 370(6522).

Russo, J Edward, and Paul JH Schoemaker. 1992. "Managing overconfidence." Sloan

management review, 33(2): 7–17.

Sagar, Rajesh, Rakhi Dandona, Gopalkrishna Gururaj, RS Dhaliwal, Aditya Singh, Alize Ferrari, Tarun Dua, Atreyi Ganguli, Mathew Varghese, Joy K Chakma, et al. 2020. "The burden of mental disorders across the states of India: the Global Burden of Disease Study 1990–2017." The Lancet Psychiatry, 7(2): 148–161.

Santomauro, Damian F, Ana M Mantilla Herrera, Jamileh Shadid, Peng Zheng, Charlie Ashbaugh, David M Pigott, Cristiana Abbafati, Christopher Adolph, Joanne O Amlag, Aleksandr Y Aravkin, et al. 2021. "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the COVID-19 pandemic." The Lancet, 398(10312): 1700–1712.

Sapiens Lab. 2021. "Mental Health Has Bigger Challenges Than Stigma." Rapid Report.

Schwardmann, Peter, and Joel Van der Weele. 2019. "Deception and self-deception." Nature human behaviour, 3(10): 1055–1061.

Selten, Reinhard. 1998. "Axiomatic characterization of the quadratic scoring rule." Experimental Economics, 1(1): 43–61.

Sherman, David K., and Geo□rey L. Cohen. 2006. "The Psychology of Self-defense: Self-A□rmation Theory." In Advances in Experimental Social Psychology. 183–242. Elsevier.

Sikander, Siham, Anisha Lazarus, Omer Bangash, Daniela Fuhr, Benedict Weobong, Revathi Krishna, Ikhlaq Ahmad, Helen Weiss, LeShwandra Price, Atif Rahman, and Vikram Patel. 2015. "The E□ectiveness and Cost-E□ectiveness of the Peer-Delivered Thinking Healthy Programme for Perinatal Depression in Pakistan and India: the SHARE Study Protocol for Randomised Controlled Trials." Trials, 16(534).

Singla, Daisy, Benedict Weobong, Abhijit Nadkarni, Neerja Chowdhary, Sachin Shinde, Arpita Anand, Christopher Fairburn, Sona Dimijdan, Richard Velleman, Helen Weiss, and Vikram Patel. 2014. "Improving the scalability of psychological treatments in developing countries: An evaluation of peer-led therapy quality assessment in Goa, India." 60(100): 53–59.

Singla, Daisy R, Brandon A Kohrt, Laura K Murray, Arpita Anand, Bruce F Chorpita, and Vikram Patel. 2017. "Psychological Treatments for the World: Lessons from Low- and Middle-Income Countries."

Steinert, Christiane, Mareike Hofmann, Johannes Kruse, and Falk Leichsenring. 2014. "Relapse rates after psychotherapy for depression–stable long-term e□ects? A meta-analysis." Journal of A□ective Disorders, 168: 107–118.

Terlizzi, E.P., and B. Zablotsky. 2020. "Mental health treatment among adults: United States, 2019." NCHS Data Brief, 380.

Valley, Zahir, and Lameze Abrahams. 2016. "The e□ectiveness of peer-delivered services in the management of mental health conditions: a meta-analysis of studies from low-and middle-income countries." International Journal for the Advancement of Counselling, 38: 330–344.

Verma, Shankey, and Aditi Mishra. 2020. "Depression, anxiety, and stress and socio-demographic correlates among general Indian public during COVID-19." 66(8): 756–762.

Weobong, Benedict, Helen A Weiss, David McDaid, Daisy R Singla, Steven D Hollon, Abhijit Nadkarni, A-La Park, Bhargav Bhat, Basavraj Katti, Arpita Anand, and Others. 2017. "Sustained e□ectiveness and cost-e□ectiveness of the Healthy Activity Programme, a brief psychological treatment for depression delivered by lay counsellors in primary care: 12-month follow-up of a randomised controlled trial." PLoS Med., 14(9).

WHO. 2010. "mhGAP intervention guide for mental, neurological and substance use disorders in non-specialized health settings." Technical documents, World Health Organization.

WHO. 2015. "Thinking healthy: a manual for psychosocial management of perinatal depression, WHO generic field-trial version 1.0, 2015." Technical documents, World Health Organization.

Zimmermann, Florian. 2020. "The Dynamics of Motivated Beliefs." American Economic Review, 110(2): 337–361.

10 Figures and Tables

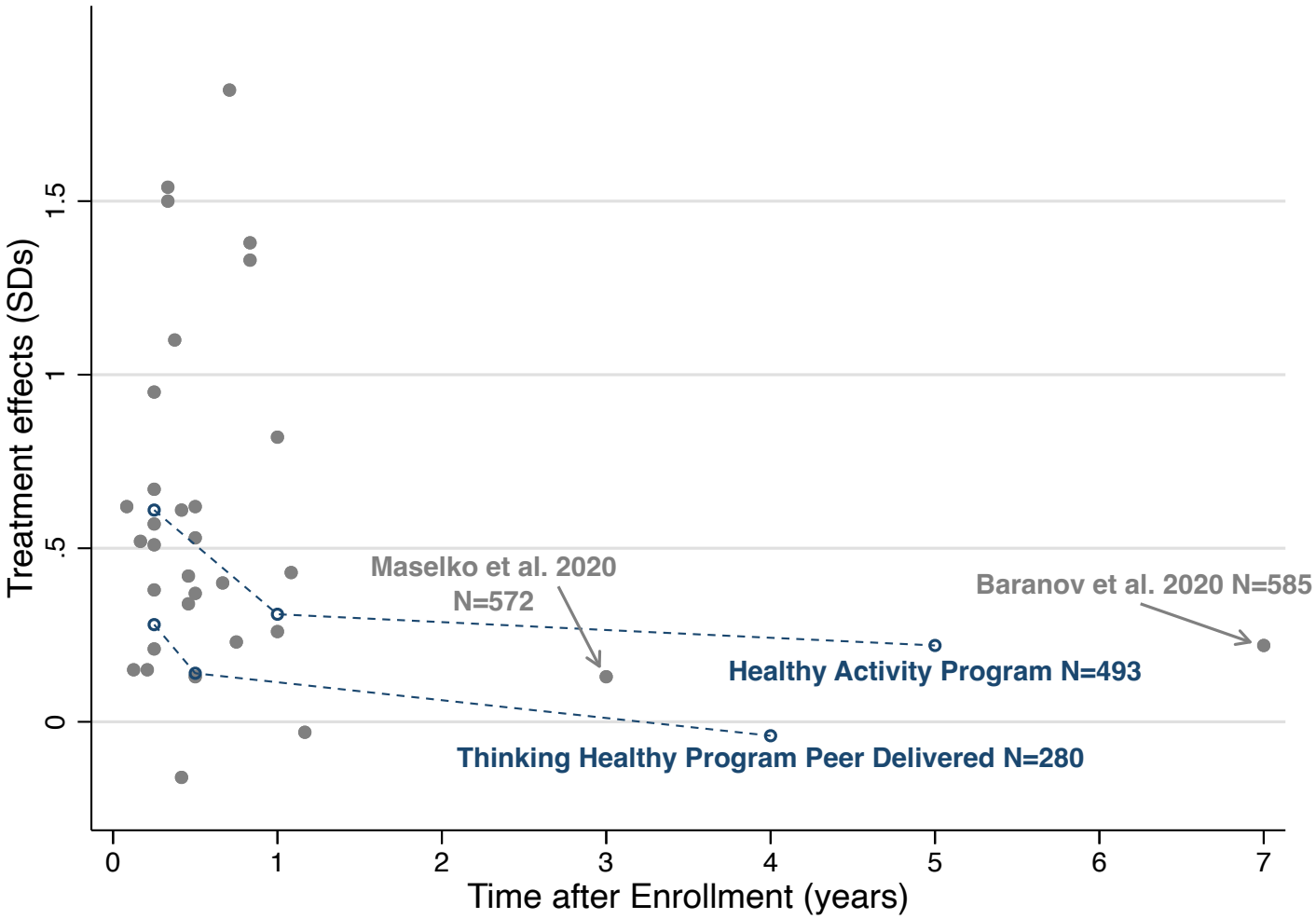


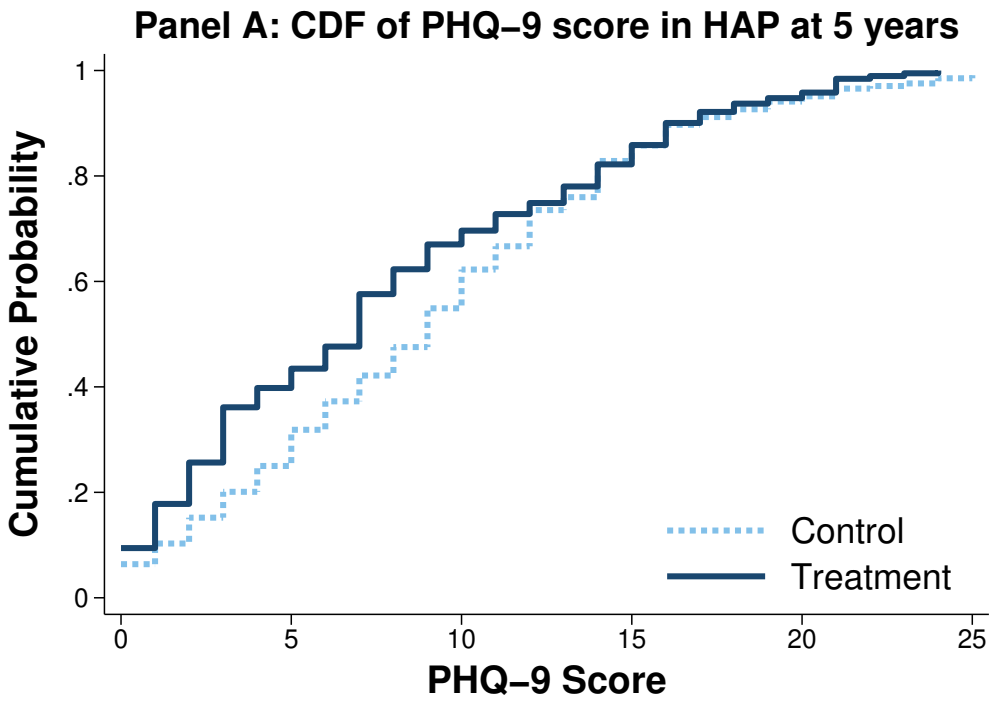
Figure 1: Treatment effects of HAP and THPP compared to previous work

Notes: This figure contextualizes our results on HAP and THPP with the results of previous RCTs studying the effects of psychological treatment on depression in low- and middle-income countries, focusing on peer-delivered interventions.

The studies are primarily taken from the literature review in Singla et al. (2017), as well as meta-analyses of peer-delivered and non-specialist interventions by Huang et al. (2020) and by Valley and Abrahams (2016).

We added further studies that were conducted after the publication of those reviews. These RCTs in low- and middle-income countries were sourced using searches of Google Scholar and PubMed, using the search terms ("depression" OR "anxiety") AND ("peer-delivered" OR "peer" OR "non-specialist" OR "interpersonal" OR "volunteer") AND ("trial" OR "randomized" OR "random" OR "controlled" OR "experiment") AND ("international" OR "LMIC" OR "low- and middle-income" OR "middle-income"), and further supplemented with the METAPSY database (`https://evidencebasedpsychotherapies.shinyapps.io/metapsy/`).

Details of the papers included can be found in Appendix Table A.16. Treatment effects are converted to standardized units to adjust for the different depression metrics used across studies.



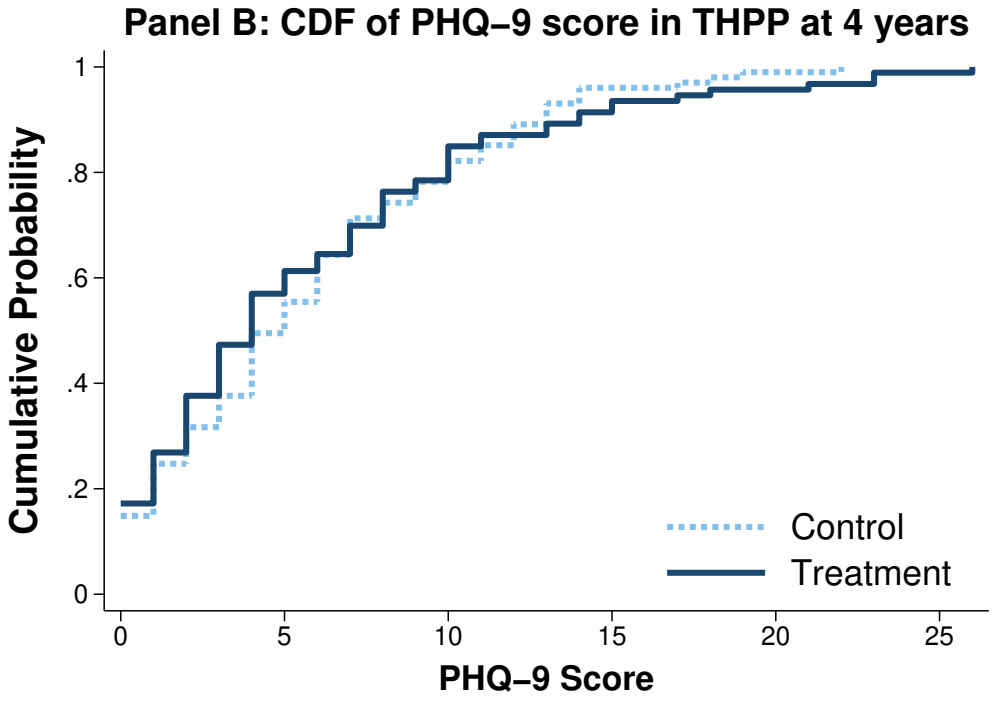


Figure 2: E□ects on distribution of depression symptom severity

Notes: This figure shows cdfs of PHQ-9 scores at the endline follow-up survey for HAP 5 years after treatment (Panel A) and for THPP 4 years after treatment (Panel B). Based on Kroenke, Spitzer and Williams (2002), the PHQ-9 screening thresholds are as follows: a score of 5 and above indicates at least mild depression, a score of 10 and above indicates at least moderate depression, and a score of 15 and above indicates at least moderately severe depression.

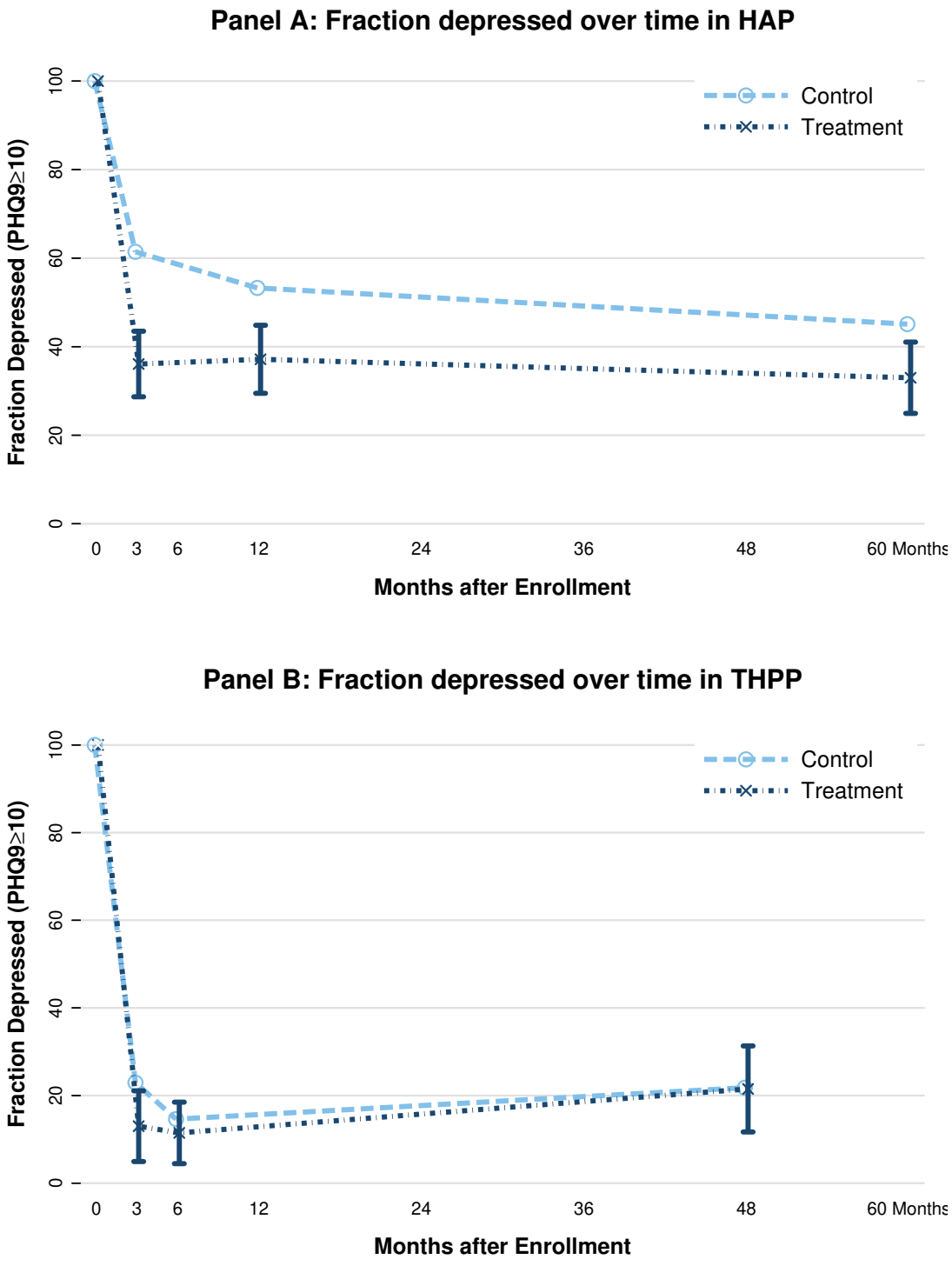


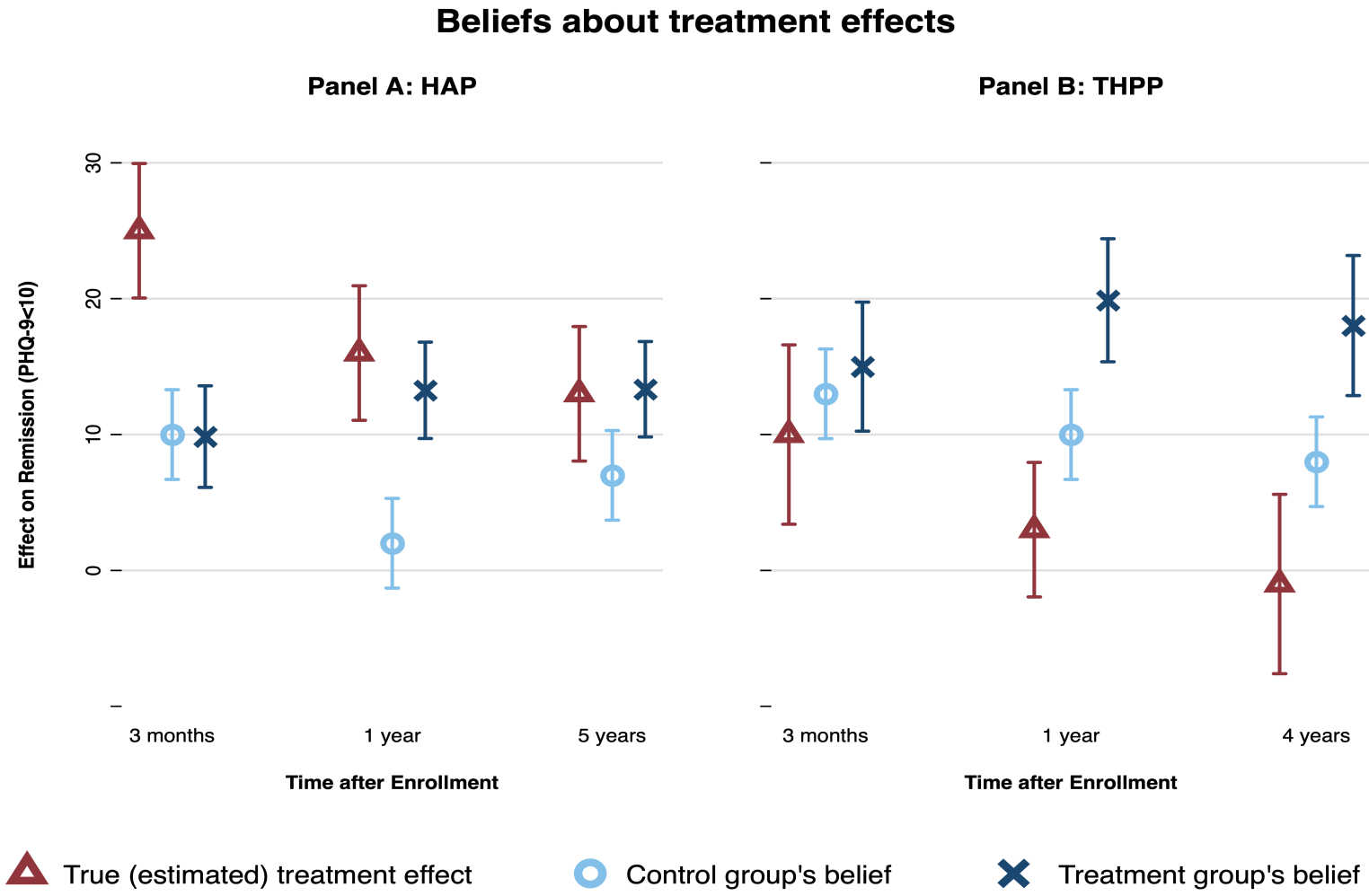
Figure 3: E□ects on depression over time

Notes: This figure shows the share of people who are depressed (PHQ-9 10) in the treatment and control groups at baseline and at three points in time during the study's follow-up. The HAP trial is shown in Panel A, and the THPP trial is shown in Panel B.

Each graph shows four data points for the respective treatment and control groups: (i) at baseline and during follow-up visits; (ii) 3 months after the intervention; (iii) 12 months (HAP) or 6 months (THPP) after the intervention; and (iv) 60 months (HAP) or 48 months (THPP) after the intervention.

90% Confidence intervals around the treatment group means are using standard errors of regressions of PHQ-9 depression levels on treatment.

Depression prevalence at baseline is 100%, since only individuals with PHQ-9 10 were eligible to be included in each study.



Notes: This figure shows control and treatment groups' mean beliefs about treatment effects on depression at each time horizon, alongside estimated true effects. Panel A shows the results of HAP participant predictions of the HAP intervention's effects on remission, and Panel B shows the same predictions among THPP respondents for the THPP intervention. Remission is defined as having a PHQ-9 score below 10.

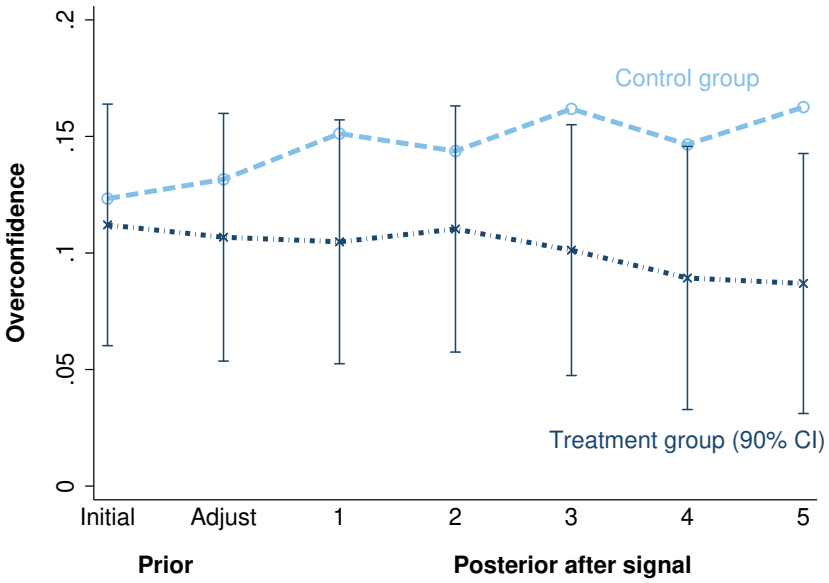
Point estimates and 90% confidence intervals for the true effects use the double machine learning approach of Chernozhukov et al. (2016).

90% confidence intervals on average beliefs are calculated from conventional standard errors of the mean.

Each participant is asked, for each time horizon and for the control group and treatment group separately, out of 10 randomly selected members of that group, how many would have had their depression "reduced to healthy levels." The difference between these beliefs is their implied belief about the treatment effect on remission from depression. Participants were incentivized for accuracy.

Figure 4: Beliefs about treatment effects

Panel A. Overconfidence over the course of the experiment



Panel B. Belief updating parameters

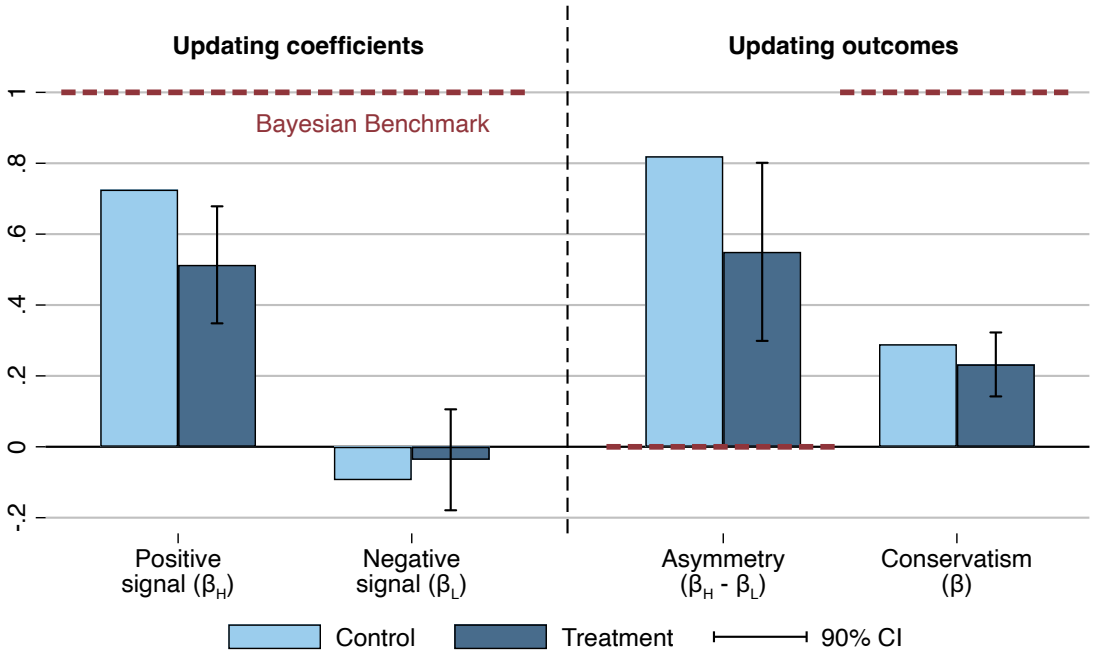


Figure 5: E□ects of psychotherapy on overconfidence and belief updating

Notes: This figure shows the impact of the treatments on belief updating, pooled across the two RCTs (split-sample estimates are reported in Table 4). Belief updating is studied in a task where individuals report their estimated probability that they are in the top half of performance on a work task, and then update their beliefs in response to informative but noisy signals (see Section 6).

Panel A reports individuals' overconfidence over the course of the experiment, defined as the difference between their reported probability of being in the top half and a full-information benchmark probability. The full-information benchmark is calculated for each individual as their true probability of being in the top half of a randomly selected group of 10 given their own score, any signals they have received up to that point, and the empirical distribution of scores. 90% confidence intervals were computed using standard errors from DML-estimated regression coefficients.

Panel B reports the coefficients from belief updating regressions (equation (3)). β_H and β_L measure responsiveness to positive and negative signals, respectively. A Bayesian is characterized by $\beta_H = \beta_L = 1$. $\beta_H \quad \beta_L$ is the difference between responsiveness to positive and negative signals, and equals 0 for a Bayesian. β measures responsiveness forcing the response to positive and negative signals to be identical, which equals 1 for a Bayesian updater. 90% confidence intervals are shown for the difference between control and treatment groups. Standard errors are clustered at the individual level.

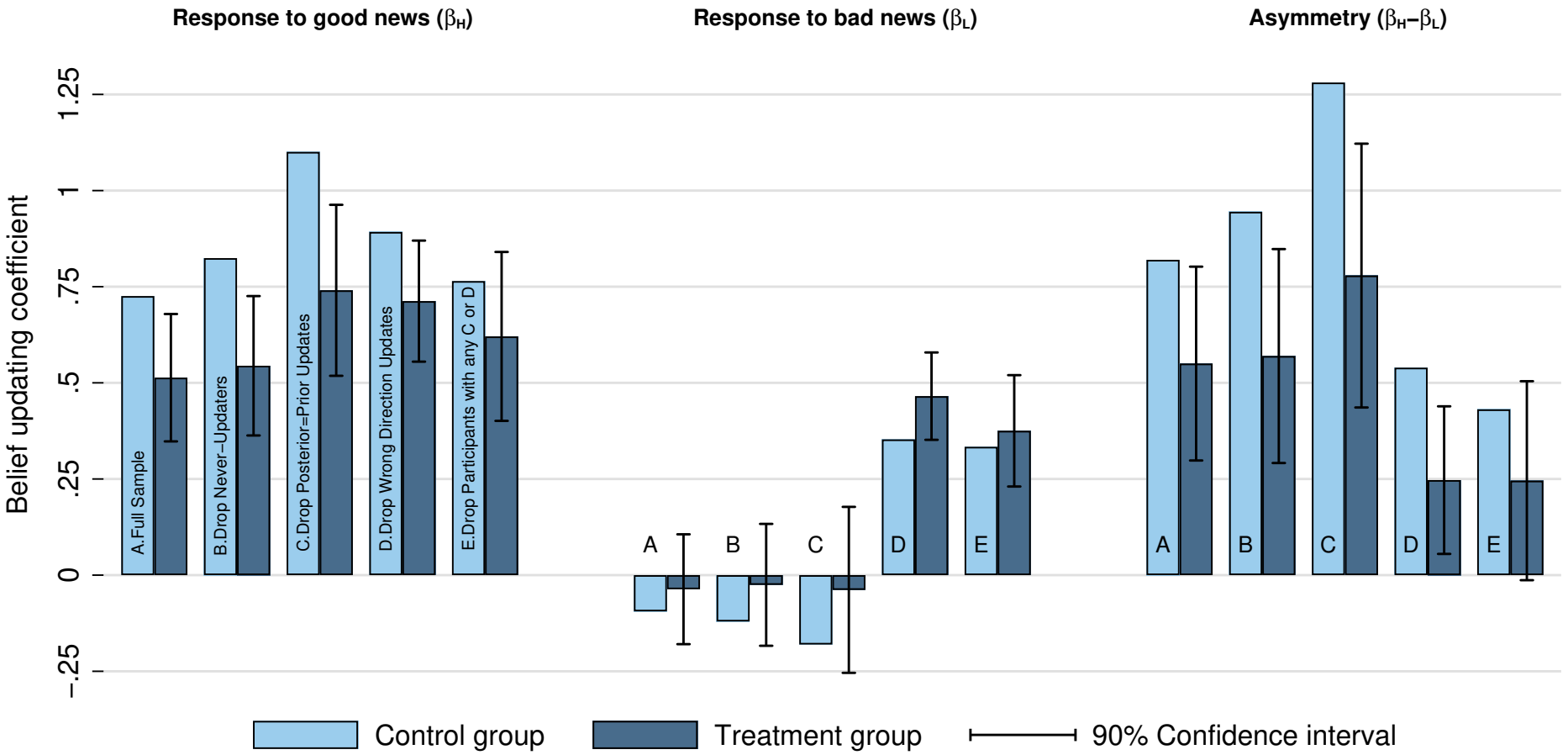


Figure 6: Belief updating coe□cients using di□erent sub-samples

Notes: This figure summarizes robustness of treatment effects on the belief updating outcomes from Panel B of Figure 5, using different subsamples. The estimated coefficients for the control group are shown in light blue and for the treatment group in dark blue. The full regression estimates are shown in Table A.7.

Coefficients come from belief updating regressions (equation (3)). β_H and β_L measure responsiveness to positive and negative signals, respectively. A Bayesian is characterized by $\beta_H = \beta_L = 1$. $\beta_H \quad \beta_L$ is the difference between responsiveness to positive and negative signals, and equals 0 for a Bayesian. 90% confidence intervals are shown for the difference between control and treatment groups. Standard errors are clustered at the individual level as these regressions include multiple observations for each individual.

Sample A utilizes the primary analysis sample as shown in Figure 5 Panel B (full sample except observations with degenerate beliefs, since the likelihood ratios are not defined in this case). Sample B drops individuals who never update their beliefs. Sample C drops individual observations where the participant did not update (posterior equals prior). Sample D drops observations with wrong-signed updates (negative updates following good news or positive updates following bad news). Sample E drops individuals who ever do not update, update in the wrong direction, or report a degenerate belief (0 or 1).

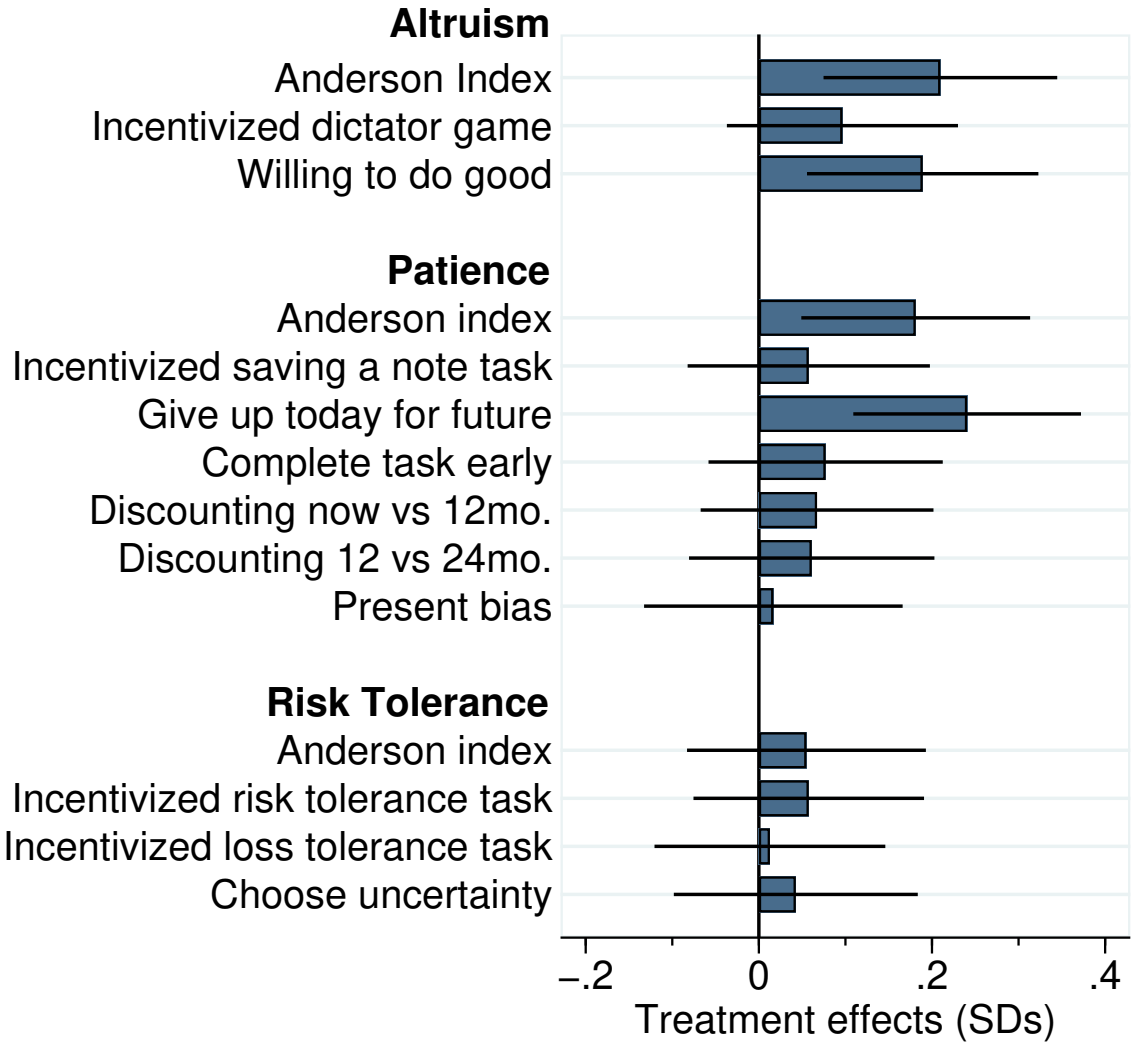


Figure 7: Standardized treatment effects on altruism, patience, and risk tolerance

Notes: This figure summarises effects of treatment on preference outcomes in the full sample, plus indices of the various sub-components. Split-sample estimates presented in Table 5.

All outcomes are standardized to mean zero, standard deviation one in the control group. Treatment effects are estimated using the double machine learning approach of Chernozhukov et al. (2016). Error bars show 90% confidence intervals.

The first panel reports measures of altruism. "Dictator game" is the amount of money out of Rs. 50 that the participant chose to send to another participant instead of keeping for themselves. "Willing to do good without expecting return" is a self-evaluation measure of altruism.

The second panel reports measures of patience. "Saving a Note Task" is a dummy equal to one if the participant saved a Rs. 100 banknote for one week, earning a Rs. 30 return. "Willing to give up today for future" and "Willing to complete task early" are self-evaluation measures of patience. Discounting and present bias parameters are computed from choices over hypothetical sooner vs. later monetary amounts. The Anderson index is constructed excluding the present bias parameter, which is a transformation of two other components.

The final panel reports measures of risk tolerance. "Risk tolerance" and "Loss tolerance" are computed from participant's switching points in incentivized risk/loss lottery choice tasks, with positive numbers corresponding to higher risk tolerance and loss tolerance. "Willing to choose uncertain outcomes" is a self-evaluation measure of risk tolerance.

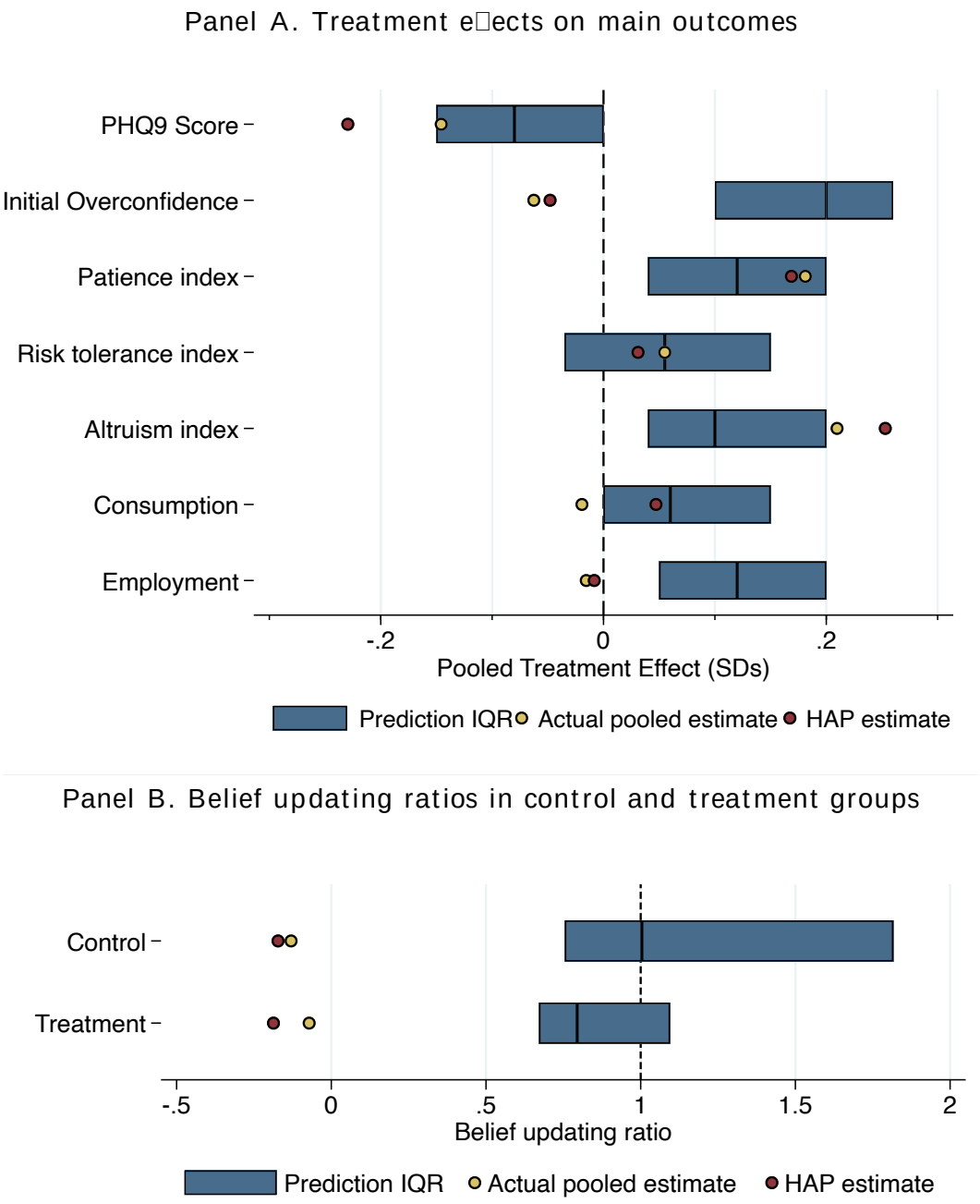


Figure 8: Survey of Experts

Notes: This figure shows the results of the survey of experts (who were only ask to forecast the results of the HAP trial), and how their predictions compare to the actual treatment effects found in the study.

Panel A shows inter-quartile ranges (in navy blue), i.e. the 25th percentile, median, and 75th percentile of experts' predictions of treatment effects for the main outcomes of the study., and dots report the actual treatment effect found in the study: yellow dots represent the pooled estimates and red dots represent estimates for HAP only. All estimates are used DML and are reported in standard deviation units.

Panel B shows experts' prediction of the ratio of responsiveness to negative and positive signals, $\frac{\beta_L}{\beta_H}$, in the control and treatment groups respectively. Experts were asked for their predictions on a scale that ranged between 0.25 and 4.0. Since the point estimate for β_L was close to 0 and negative, the true point estimates are also close to 0 and negative, and thus were outside the range of what the survey allowed experts to report.

Table 1: Attrition and balance

	Baseline (N = 773)		Follow-up (N = 589)	
	Control mean (S.D.)	Δ Treatment (S.E.)	Control mean (S.D.)	Δ Treatment (S.E.)
Completed this stage	1.00	0.00	0.78	-0.05
			(0.42)	(0.03)
Baseline PHQ-9 score	16.15	0.24	16.14	0.36
	(3.77)	(0.21)	(3.66)	(0.24)
Duration of depression at baseline (weeks)	30.87	1.57	33.25	-0.09
	(95.34)	(6.58)	(105.36)	(8.42)
Years between baseline and follow-up	-	-	4.57	0.01
			(0.73)	(0.05)
Number of children	-	-	2.16	-0.01
			(1.16)	(0.11)
Female	0.85	-0.00	0.88	-0.01
	(0.35)	(0.02)	(0.33)	(0.03)
Age	36.37	-0.22	37.64	0.21
	(12.99)	(0.72)	(12.83)	(0.82)
Married	0.80	-0.01	0.81	-0.01
	(0.40)	(0.03)	(0.39)	(0.03)
Years of education	6.51	-0.37	6.24	-0.48
	(4.52)	(0.34)	(4.38)	(0.36)
Hindu	0.78	0.01	0.79	0.03
	(0.41)	(0.03)	(0.41)	(0.03)
Expected treatment e□cacy	0.76	0.00	0.75	-0.02
	(0.43)	(0.03)	(0.43)	(0.04)
Homemaker	0.57	0.02	0.59	-0.00
	(0.50)	(0.03)	(0.49)	(0.04)
Unemployed	0.09	0.02	0.07	0.04
	(0.28)	(0.02)	(0.25)	(0.02)
Employed	0.34	-0.04	0.34	-0.04
	(0.47)	(0.03)	(0.47)	(0.04)
F-test	0.49		0.72	
p-value	0.91		0.73	

Notes: This table presents characteristics of the RCT samples, pooled across both trials. The characteristics split by trial are presented in Appendix Table A.1.

The first two columns present baseline mean characteristics in the control group (standard deviations in parentheses) and the difference between mean characteristics in the treatment and control groups (standard errors in parentheses) among all participants who took part in one of the two trials.

The last two columns present the same characteristics and differences among only those individuals who also appear in the follow-up survey. F-tests from a regression of the treatment dummy on these characteristics in each sample are presented at the bottom of the table.

'Baseline PHQ-9 score' is a standard diagnostic test for depression; it ranges from 0 to 27, with a score greater than or equal to 10 indicating at least moderate depression symptoms. The baseline score reported is the score at enrollment in the initial trial (before the interventions).

'Female' takes the value 1 for female participants and 0 for male participants. 'Expected treatment efficacy' takes the value 1 if the participant expected the treatment to be at least 'somewhat useful', and 0 if the patient expected the treatment to be 'a little useful' or less. 'Homemaker' takes the value of 1 if participant does not work for pay outside the home and is not searching for other work currently.

Table 2: Impacts of the treatments on depression

	Full Sample		HAP		THPP		HAP - THPP
	Control mean (S.D.)	Treatment Effect (S.E.)	Control mean (S.D.)	Treatment Effect (S.E.)	Control mean (S.D.)	Treatment Effect (S.E.)	Δ T.E. [p-value]
Panel A: OLS without controls							
PHQ-9 Score	7.97	−0.97**	9.10	−1.43**	5.68	−0.04	−1.40
	(5.86)	(0.48)	(5.97)	(0.61)	(4.92)	(0.78)	[0.18]
PHQ-9<5	0.33	0.12***	0.25	0.15***	0.50	0.07	0.07
	(0.47)	(0.04)	(0.43)	(0.05)	(0.50)	(0.07)	[0.38]
PHQ-9<10	0.63	0.08**	0.55	0.12**	0.78	0.00	0.12
	(0.48)	(0.04)	(0.50)	(0.05)	(0.41)	(0.06)	[0.15]
Mood Score	6.49	0.38**	6.19	0.38	7.10	0.40	−0.02
	(2.35)	(0.19)	(2.40)	(0.24)	(2.14)	(0.32)	[0.96]
Panel B: DML							
PHQ-9 Score	7.97	−0.85*	9.10	−1.37**	5.68	0.15	−1.52
	(5.86)	(0.47)	(5.97)	(0.59)	(4.92)	(0.77)	[0.13]
PHQ-9<5	0.33	0.11***	0.25	0.13***	0.50	0.07	0.06
	(0.47)	(0.04)	(0.43)	(0.05)	(0.50)	(0.07)	[0.50]
PHQ-9<10	0.63	0.08**	0.55	0.13***	0.78	−0.01	0.14*
	(0.48)	(0.04)	(0.50)	(0.05)	(0.41)	(0.06)	[0.08]
Mood Score	6.49	0.39**	6.19	0.43*	7.10	0.34	0.08
	(2.35)	(0.19)	(2.40)	(0.24)	(2.14)	(0.31)	[0.84]
N	589		395		194		

Notes: This table presents estimates of the effects of the treatments on depression as measured in the follow-up study, several years after the initial interventions.

The first two columns show the impacts for the full sample, i.e., pooling the two trials. The following columns show the impacts for the two trials separately. Odd columns report control-group means, along with standard deviations in parentheses. Even columns report treatment effects along with standard errors in parentheses. The last column reports the difference in treatment effects across the two trials, with the p-value of a test for equal treatment effects across the two trials in square brackets.

Panel A shows estimates from OLS regressions without control variables (apart from a dummy variable for the THPP trial, in column 2). Panel B reports coefficients estimated using the double machine learning approach of Chernozhukov et al. (2016), using all control variables available in the baseline data.

The PHQ-9 score is a standard diagnostic test of depression measured on a scale from 0 to 27. A score of 10 or higher indicates at least moderate depression symptoms, and a score of 5 or higher indicates at least mild depression symptoms. The Mood Score is a self-reported measure of happiness on a scale from 1 to 10, averaged across three days.

The bottom row shows the sample sizes for PHQ-9 Score related variables. Sample sizes for the Mood Score are 578 (full sample), 387 (HAP), and 191 (THPP), respectively.

Table 3: Impacts on beliefs about treatment effects

	Full Sample		HAP		THPP	
	Control Mean S.E.	Treatment Effect S.E.	Control Mean S.E.	Treatment Effect S.E.	Control Mean S.E.	Treatment Effect S.E.
Panel A: Remission after 3 months						
True Effect (N=711)		0.2*** (0.03)		0.25*** (0.04)		0.1** (0.05)
Participant Beliefs (N=450)	0.12 (0.01)	0.0 (0.02)	0.1 (0.02)	-0.01 (0.03)	0.13 (0.02)	0.01 (0.03)
p-value (control group=true effect)		[0.01]		[<0.01]		[0.52]
p-value (treatment group=true effect)		[0.02]		[<0.01]		[0.39]
Panel B: Remission after 6 months/1 year						
True Effect (N=699)		0.11*** (0.03)		0.16*** (0.05)		0.03 (0.04)
Participant Beliefs (N=450)	0.06 (0.01)	0.1*** (0.02)	0.02 (0.02)	0.11*** (0.03)	0.1 (0.02)	0.09*** (0.03)
p-value (control group=true effect)		[0.11]		[0.01]		[0.11]
p-value (treatment group=true effect)		[0.16]		[0.61]		[<0.01]
Panel C: Remission after 4/5 years						
True Effect (N=589)		0.08** (0.04)		0.13** (0.05)		-0.01 (0.06)
Participant Beliefs (N=450)	0.07 (0.01)	0.08*** (0.02)	0.07 (0.02)	0.07** (0.03)	0.08 (0.02)	0.09** (0.04)
p-value (control group=true effect)		[0.8]		[0.23]		[0.05]
p-value (treatment group=true effect)		[0.04]		[0.95]		[<0.01]

Notes: This table reports estimates of participants' beliefs of the treatment effects on depression. We report estimated treatment effects on true remission from depression (PHQ-9 < 10), and participants' beliefs of these treatment effects. Estimation uses the double machine learning approach of Chernozhukov et al. (2016).

Participants are asked, separately for the treatment group and the control group, out of 10 randomly selected individuals, how many would have had their depression "reduced to healthy levels." We use their answers to construct their implied belief about the treatment effect on remission from depression. HAP participants are asked about the effects of HAP; THPP participants are asked about the effects of THPP.

For each time horizon, we report (i) true (estimated) treatment effects; (ii) the control group's average beliefs of these treatment effects; and (iii) the treatment effect on these beliefs, i.e. the difference between treatment and control group's beliefs of treatment effects; and (iv) p-values corresponding to tests of whether mean beliefs of treatment and control participants, respectively, equal the "true" estimated effect.

Appendix Table A.3 expands this table by presenting the raw beliefs about outcomes in the control and treatment groups.

Table 4: Impacts on confidence and belief updating

	Full Sample		HAP		THPP	
	Control mean (S.E.)	Treatment Effect (S.E.) [p-values] {q-values}	Control mean (S.E.)	Treatment Effect (S.E.) [p-values] {q-values}	Control mean (S.E.)	Treatment Effect (S.E.) [p-values] {q-values}
Panel A. Overconfidence						
Initial Overconfidence	0.13	-0.03	0.22	-0.02	-0.04	-0.04
	(0.02)	(0.03)	(0.03)	(0.04)	(0.04)	(0.05)
		[0.42] {0.43}		[0.64] {0.64}		[0.50] {0.50}
Final Overconfidence	0.16	-0.08**	0.23	-0.07*	0.02	-0.09
	(0.02)	(0.03)	(0.03)	(0.04)	(0.04)	(0.06)
		[0.02] {0.05}		[0.11] {0.22}		[0.13] {0.25}
N		576		385		191
Panel B. Belief-updating coefficients						
Response to Good News (β_H)	0.73	-0.21**	0.73	-0.22*	0.67	-0.20
	(0.08)	(0.10)	(0.10)	(0.13)	(0.10)	(0.14)
		[0.03] {0.07}		[0.09] {0.18}		[0.15] {0.31}
Response to Bad News (β_L)	-0.09	0.06	-0.13	0.03	-0.03	0.11
	(0.06)	(0.09)	(0.08)	(0.11)	(0.07)	(0.12)
		[0.51] {0.51}		[0.78] {0.78}		[0.35] {0.35}
Panel C. Asymmetry and conservatism						
Asymmetry ($\beta_H - \beta_L$)	0.82	-0.27*	0.85	-0.25	0.69	-0.32
	(0.12)	(0.15)	(0.15)	(0.19)	(0.14)	(0.21)
		[0.08] {0.16}		[0.20] {0.29}		[0.12] {0.25}
Conservatism ($\bar{\beta}$)	0.29	-0.06	0.27	-0.07	0.29	-0.02
	(0.04)	(0.05)	(0.05)	(0.07)	(0.06)	(0.08)
		[0.30] {0.40}		[0.29] {0.29}		[0.80] {0.80}
N		2620		1715		905

Notes: This table reports treatment effects on self-confidence and belief updating. Table A.5 presents more details on the construction of each outcome.

Panel A uses the double machine learning approach of Chernozhukov et al. (2016). Panels B and C use the belief-updating regression specification (3). "Control mean" and "Treatment Effect" columns report standard errors in parentheses.

Stars refer to unadjusted two-sided p-values at thresholds 0.1 and 0.05 respectively. Unadjusted p-values are reported in square brackets and False Discovery Rate-adjusted q-values in curly brackets. Multiple testing adjustments correct across outcomes within each panel separately.

Panel A reports effects on overconfidence. 'Initial Overconfidence' equals the participant's initial belief about their probability of being in the upper half of performance in their group of ten people, minus the full-information benchmark, computed assuming groups drawn from the population performance distribution. 'Final Overconfidence' equals the participant's belief after observing five signals, minus the Bayesian posterior given the full-information benchmark and the observed signals.

Panel B reports belief updating coefficients, which measure the change in the posterior likelihood ratio in response to signals, relative to the Bayesian benchmark. β_H measures the response to good news, and β_L the response to bad news. Bayesian updating implies $\beta_H = 1$, and $\beta_L = 1$.

Panel C reports transformations of the belief updating coefficients. 'Asymmetry' measures the difference between the response to good and bad news. Bayesian updating implies $\beta_H \quad \beta_L = 0$. 'Conservatism' measures the average responsiveness to news, forcing the coefficients on good and bad news to be identical. Bayesian updating implies $\beta = 1$.

Table 5: Patience, risk tolerance, and altruism

	Full Sample		HAP		THPP	
	Control mean (S.D.)	Treatment Effect (S.E.) [p-values] {q-values}	Control mean (S.D.)	Treatment Effect (S.E.) [p-values] {q-values}	Control mean (S.D.)	Treatment Effect (S.E.) [p-values] {q-values}
Panel A: Altruism						
Anderson index	0.00	0.21**	0.00	0.25**	0.00	0.16
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.14)
		[0.01] {0.04}		[0.02] {0.05}		[0.27] {0.40}
Amount given away in dictator game	0.00	0.10	0.00	0.13	0.00	0.05
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.14)
		[0.25] {0.69}		[0.20] {0.72}		[0.72] {0.93}
Willing to do good without expecting return	0.00	0.19**	0.00	0.20*	0.00	0.19
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.13)
N		562		379		183
Panel B: Patience						
Anderson index	0.00	0.18**	0.00	0.17*	0.00	0.27*
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.15)
		[0.03] {0.05}		[0.08] {0.12}		[0.09] {0.27}
Note saved in saving a note task	0.00	0.06	0.00	-0.05	0.00	0.26*
	(1.00)	(0.08)	(1.00)	(0.11)	(1.00)	(0.14)
		[0.52] {0.69}		[0.66] {0.72}		[0.08] {0.30}
Willing to give up today for future	0.00	0.24***	0.00	0.33***	0.00	0.07
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.14)
Willing to complete task early (not delay)	0.00	0.08	0.00	0.10	0.00	0.04
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.15)
Discounting δ_a weight: today vs 12 months	0.00	0.07	0.00	0.09	0.00	0.02
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.15)
Discounting δ_b weight: 12 months vs 24 months	0.00	0.06	0.00	0.06	0.00	0.05
	(1.00)	(0.09)	(1.00)	(0.10)	(1.00)	(0.16)
Present bias β: δ_a/δ_b	0.00	0.02	0.00	0.02	0.00	0.07
	(1.00)	(0.09)	(1.00)	(0.11)	(1.00)	(0.17)
N		562		379		183
Panel C: Risk tolerance						
Anderson index	0.00	0.05	0.00	0.03	0.00	0.11
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.14)
		[0.52] {0.52}		[0.78] {0.79}		[0.45] {0.45}
Risk tolerance	0.00	0.06	0.00	0.08	0.00	0.01
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.14)
		[0.48] {0.69}		[0.42] {0.72}		[0.93] {0.93}
Loss tolerance	0.00	0.01	0.00	-0.04	0.00	0.11
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.14)
		[0.90] {0.90}		[0.72] {0.72}		[0.46] {0.92}
Willing to choose uncertain outcomes	0.00	0.04	0.00	0.00	0.00	0.13
	(1.00)	(0.09)	(1.00)	(0.10)	(1.00)	(0.15)
N		562		379		183

Notes: This table reports treatment effects on preference measures. See Table A.5 for more detail on the construction of each outcome.

All outcomes are standardized to mean zero, standard deviation 1 in the control group. All estimates use the double machine learning approach of Chernozhukov et al. (2016). "Control mean" columns report standard deviations in parentheses. "Treatment Effect" columns report standard errors in parentheses.

Stars refer to unadjusted two-sided p-values at thresholds 0.1, 0.05, and 0.01, respectively. Unadjusted p-values are reported in square brackets and False Discovery Rate-adjusted q-values in curly brackets. Multiple testing corrections correct across two distinct sets of outcomes. First, the three index outcomes: patience, risk tolerance, and altruism. Second, the set of revealed-preference outcomes (which are sub-components of the index measures) "Saving a Note," "Risk Aversion," "Loss Aversion," "Dictator Game," since these are the primary outcomes specified in our pre-analysis plan. Each panel reports an inverse covariance weighted index measure over the sub-components reported in that panel (Anderson, 2008).

Panel A reports measures of altruism. "Dictator game" is computed from the amount of money out of Rs. 50 that the participant chose to send to another participant instead of keeping for themselves. "Willing to do good without expecting return" is a self-evaluation measure of altruism.

Panel B reports measures of patience. "Note saved in saving a Note Task" records whether the participant saved a Rs. 100 banknote for one week, earning a Rs. 30 return. "Willing to give up today for future" and "Willing to complete task early" are self-evaluation measures of patience. Discounting and present bias parameters are computed from choices over hypothetical sooner vs. later monetary amounts. The Anderson index is constructed excluding the present bias parameter, which is a transformation of two other components.

Panel C reports measures of risk tolerance. "Risk tolerance" and "Loss tolerance" are based on the participant's switching point in incentivized risk/loss lottery choice tasks, aligned so that positive numbers correspond to higher risk tolerance and higher loss tolerance. "Willing to choose uncertain outcomes" is a self-evaluation measure of risk tolerance.

Table 6: Impacts on employment, productivity, and expenditures

	Full Sample		HAP		THPP	
	Control mean (S.D.)	Treatment E□ect (S.E.)	Control mean (S.D.)	Treatment E□ect (S.E.)	Control mean (S.D.)	Treatment E□ect (S.E.)
Panel A: Survey-based measures of labor supply and employment						
Engaged in work in the last week	0.25	-0.01	0.27	-0.01	0.21	-0.03
	(0.43)	(0.03)	(0.44)	(0.04)	(0.41)	(0.06)
Work hours in the last week	5.14	0.67	6.31	-0.35	2.76	2.03
	(13.91)	(1.20)	(15.71)	(1.61)	(8.80)	(1.68)
Earnings in the past month (PPP)	83.43	3.78	98.09	-6.96	53.63	27.73
	(230.13)	(16.50)	(263.82)	(21.80)	(134.86)	(22.90)
Available to take on an employment opportunity	0.82	0.04	0.78	0.07*	0.91	0.01
	(0.39)	(0.03)	(0.42)	(0.04)	(0.29)	(0.05)
Job search hours per week (among unemployed)	0.78	0.14	0.73	0.32	0.85	-0.18
	(2.77)	(0.27)	(2.63)	(0.36)	(3.02)	(0.39)
N	557		375		182	
Panel B: Revealed-preference measures of labor supply and productivity						
Reservation wage (PPP to make 1000 bracelets)	64.17	3.14	65.26	5.57	62.01	-3.09
	(46.21)	(3.70)	(47.55)	(4.57)	(43.57)	(6.23)
Applied for ability-based contract	0.63	0.01	0.62	0.02	0.64	0.01
	(0.48)	(0.04)	(0.49)	(0.05)	(0.48)	(0.07)
Bracelets made in ten minutes	5.04	-0.01	4.68	-0.08	5.76	0.10
	(1.52)	(0.11)	(1.42)	(0.13)	(1.44)	(0.19)
N	576		385		191	
Panel C: Expenditures						
Total monthly expenditure (PPP)	1013.27	-17.21	977.41	43.52	1086.50	-119.20
	(819.22)	(69.40)	(866.24)	(94.80)	(712.35)	(86.10)
Food	325.39	-17.54	288.20	-7.57	401.32	-37.69
	(202.44)	(15.80)	(186.97)	(18.30)	(212.40)	(29.70)
Durable goods	153.34	24.52	172.07	49.59	115.08	-21.33
	(364.84)	(36.20)	(418.48)	(52.80)	(214.60)	(26.20)
Medical	168.16	-9.90	165.94	12.59	172.69	-46.30
	(320.25)	(33.50)	(328.47)	(45.70)	(304.43)	(35.80)
Other	366.39	-11.41	351.20	-9.39	397.42	-13.86
	(303.96)	(25.60)	(298.44)	(32.70)	(314.25)	(40.70)
N	558		376		182	
Panel D: Indices of other outcomes						
Female empowerment	0.00	0.06	0.00	0.19	0.00	-0.13
	(1.00)	(0.09)	(1.00)	(0.13)	(1.00)	(0.14)
Intimate partner violence (IPV)	0.00	-0.12	0.00	-0.13	0.00	-0.11
	(1.00)	(0.08)	(1.00)	(0.12)	(1.00)	(0.10)
Sleep	0.00	0.20**	0.00	0.20*	0.00	0.25*
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.15)
Loneliness	0.00	-0.08	0.00	-0.13	-0.00	-0.03
	(1.00)	(0.09)	(1.00)	(0.11)	(1.00)	(0.15)
Locus of control	0.00	0.04	0.00	-0.02	0.00	0.19
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.14)
N	566		378		188	

Notes: This table reports treatment effects on labor-market outcomes (Panels A and B), household consumption (Panel C), and indices of other outcomes. All estimates use the double machine learning approach of Chernozhukov et al. (2016).

Panel A shows survey measures of labor market outcomes. 'Engaged in work in the last week' counts those unemployed due to disability (N=29) as not working. 'Work hours in the last week' asked of everyone, and counts those unemployed as 0. 'Earnings in the past month' asked of everyone, those who are unemployed but were employed in the last three months have their monthly earnings averaged over the three months. 'Available to take employment opportunity' is the proportion of individuals available to take up wage paying employment if an opportunity they were interested in were to arise in the next 4 months. 'Job search hours per week' are conditional on unemployment, with sample sizes of N=393 (pooled sample), N=250 (HAP), and N=143 (THPP).

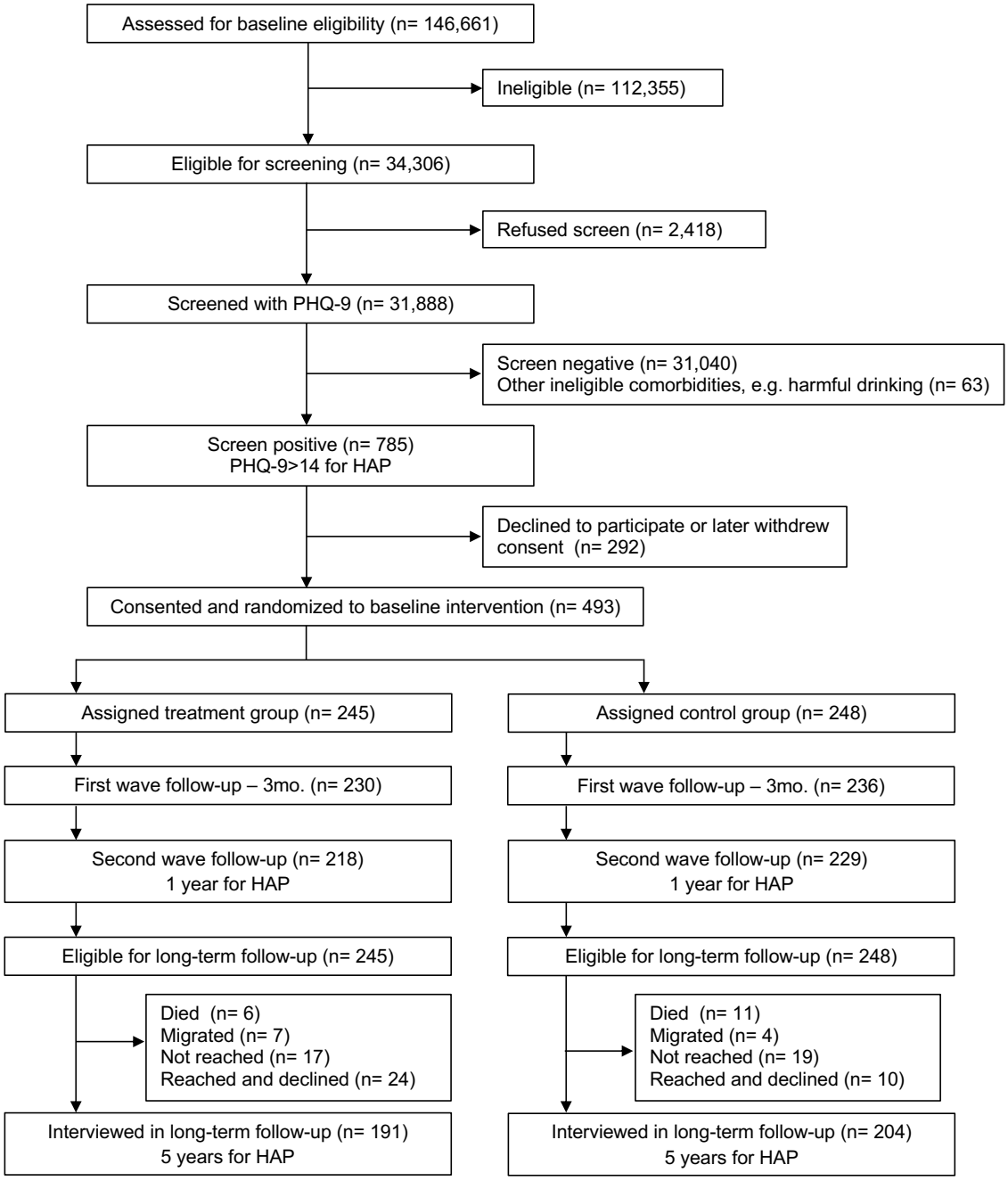
Panel B shows revealed-preference measures of work-related outcomes: (i) participant's reservation wage for the employment contract that is offered to everyone regardless of performance (price to make 1000 bracelets on own time); (ii) whether a participant chose to apply for the "ability-based" employment contract (pre-specified as our primary labor market outcome). (iii) number of bracelets made in 10 minutes, a measure of productivity;

Panel C shows survey measures of household expenditures. Monthly consumption measurement is assessed using conversion to USD at purchasing power parity, at a rate of 21.107 rupees/dollar, from the OECD 2019 USD/IND PPP Index. Medical expenditure includes both inpatient and outpatient medical expenses. Other consumption is calculated as residual total consumption after subtracting food, medical expenditure, and durable goods, and includes items like phone costs, fuel, power, tuition, transportation, ceremonies, toiletries, and taxes.

Panel D shows indices of other outcomes, which are shown in more detail in the appendix: Female Empowerment (Table A.11); Intimate Partner Violence (Table A.12); and Sleep, Loneliness, and Locus of Control (Table A.13).

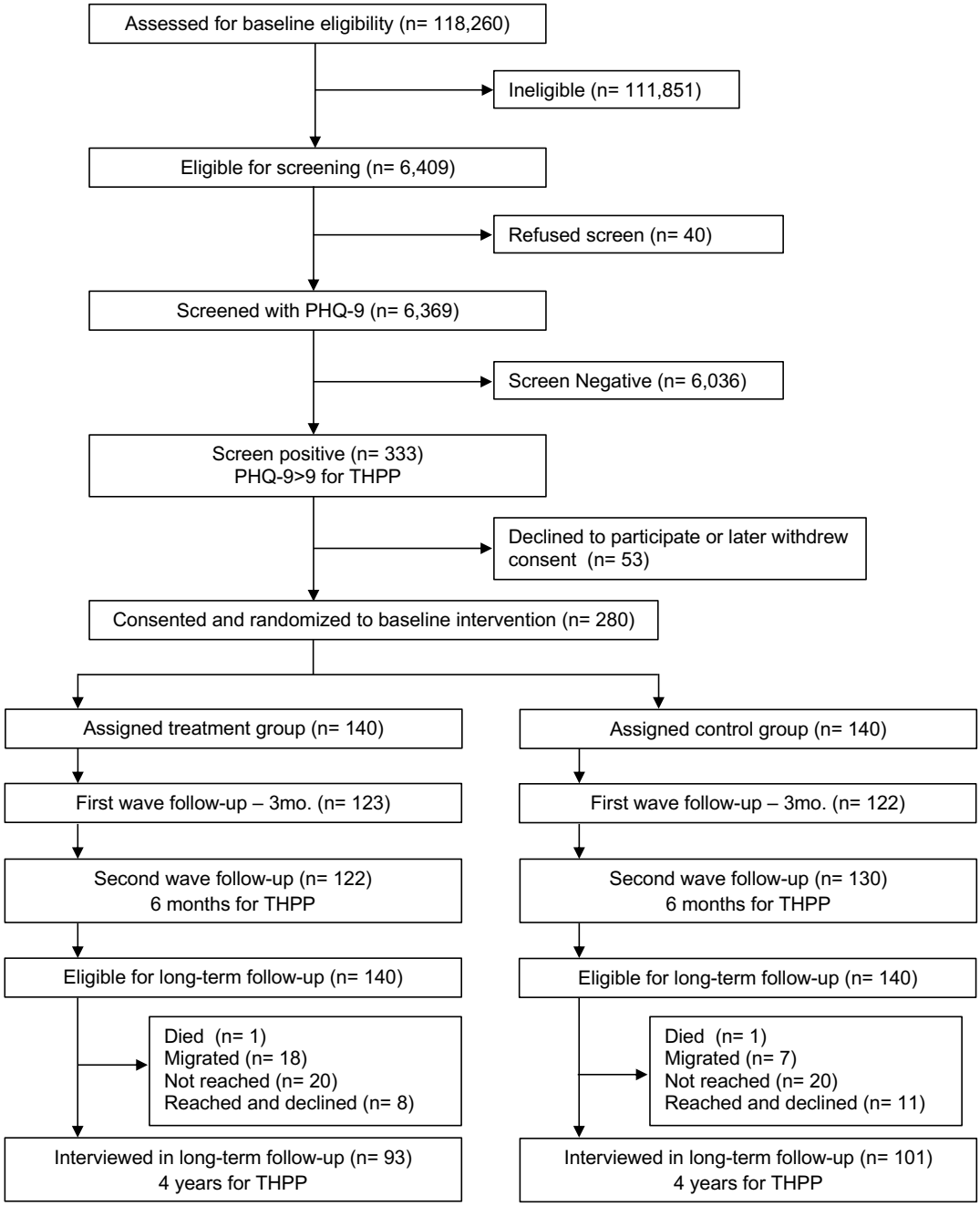
A Appendix A: Supplementary Figures and Tables

Figure A.1: Trial Flow Diagram - HAP



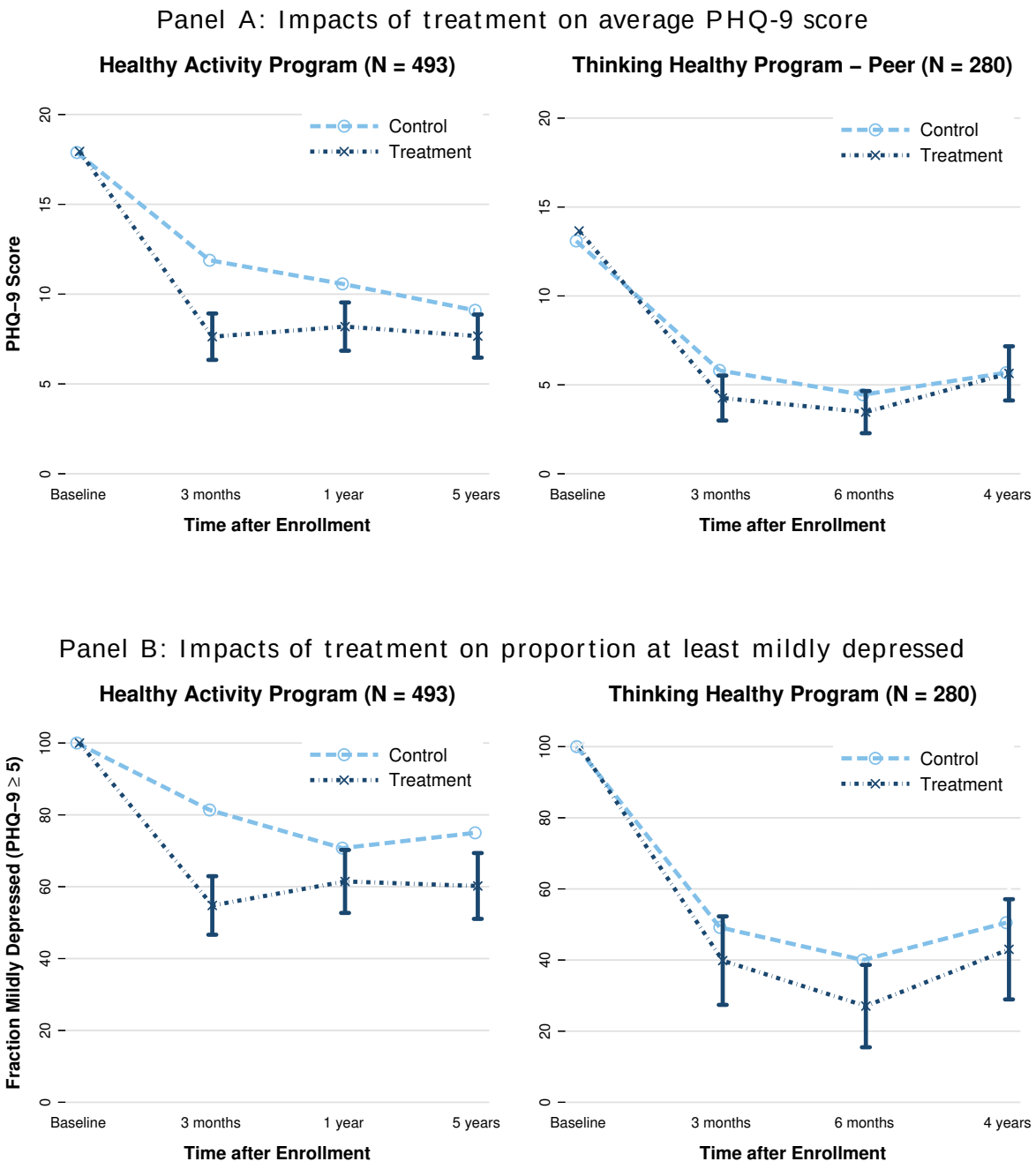
Notes: This table displays the sample sizes and the trial flow chart for the HAP baseline as well as the follow-up studies. Full trial flow charts for the first waves are available in Patel et al. (2017); Weobong et al. (2017).

Figure A.2: Trial Flow Diagram - THPP



Notes: This table displays the sample sizes and the trial flow chart for the THPP baseline as well as the follow-up studies. Full trial flow charts for the first waves are available in Fuhr et al. (2019).

Figure A.3: E□ects of treatment on PHQ-9 scores and alternate depression cuto□s



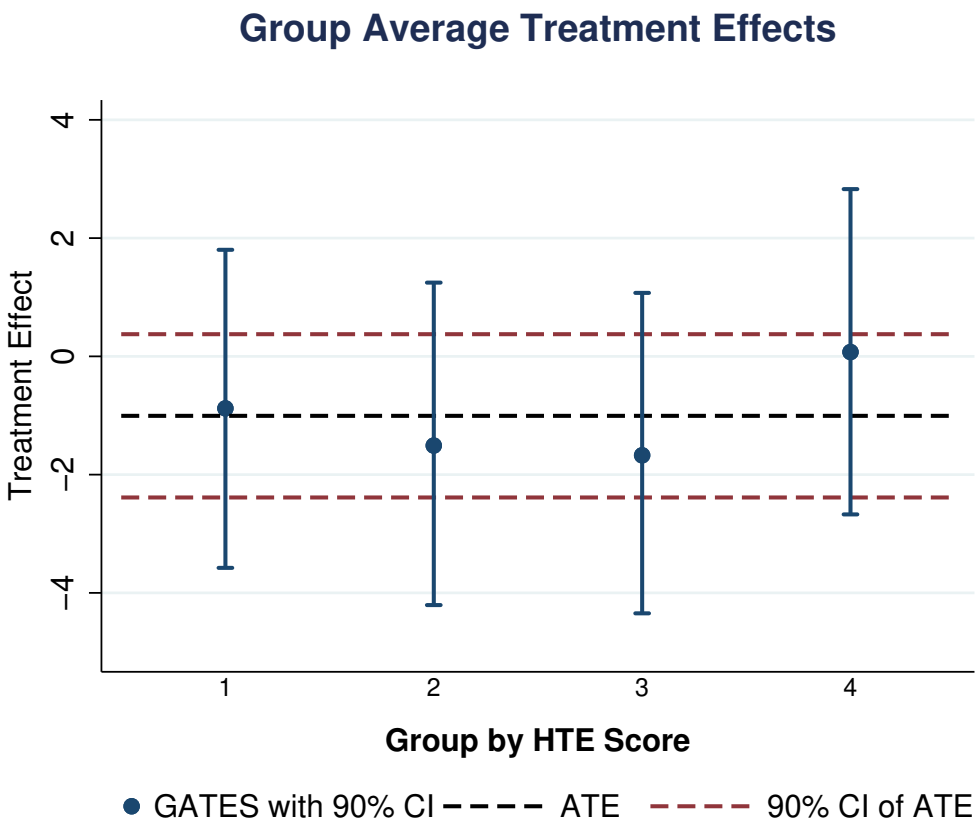
Notes: This figures shows the treatment effects on depression over time, using parallel measures of depression to those in Figure 3.

90% confidence intervals around the treatment group means are reported, and are estimated using OLS regression on the outcome with a dummy variable for the treatment covariate.

Each sub-graph shows four data points, baseline, 3 months post-intervention, 1 year (HAP) or 6 months (THPP) post intervention, and 5 years (HAP) or 4 years (THPP) post intervention. PHQ-9 scores take values from 0-21, where higher scores indicate more symptoms and severity. A score of 5 is used to indicate mild depression.

Panel A shows average PHQ-9 scores in the two groups. Panel B shows the fraction of people who are at least mildly depressed, i.e. with a PHQ-9 score of at least 5.

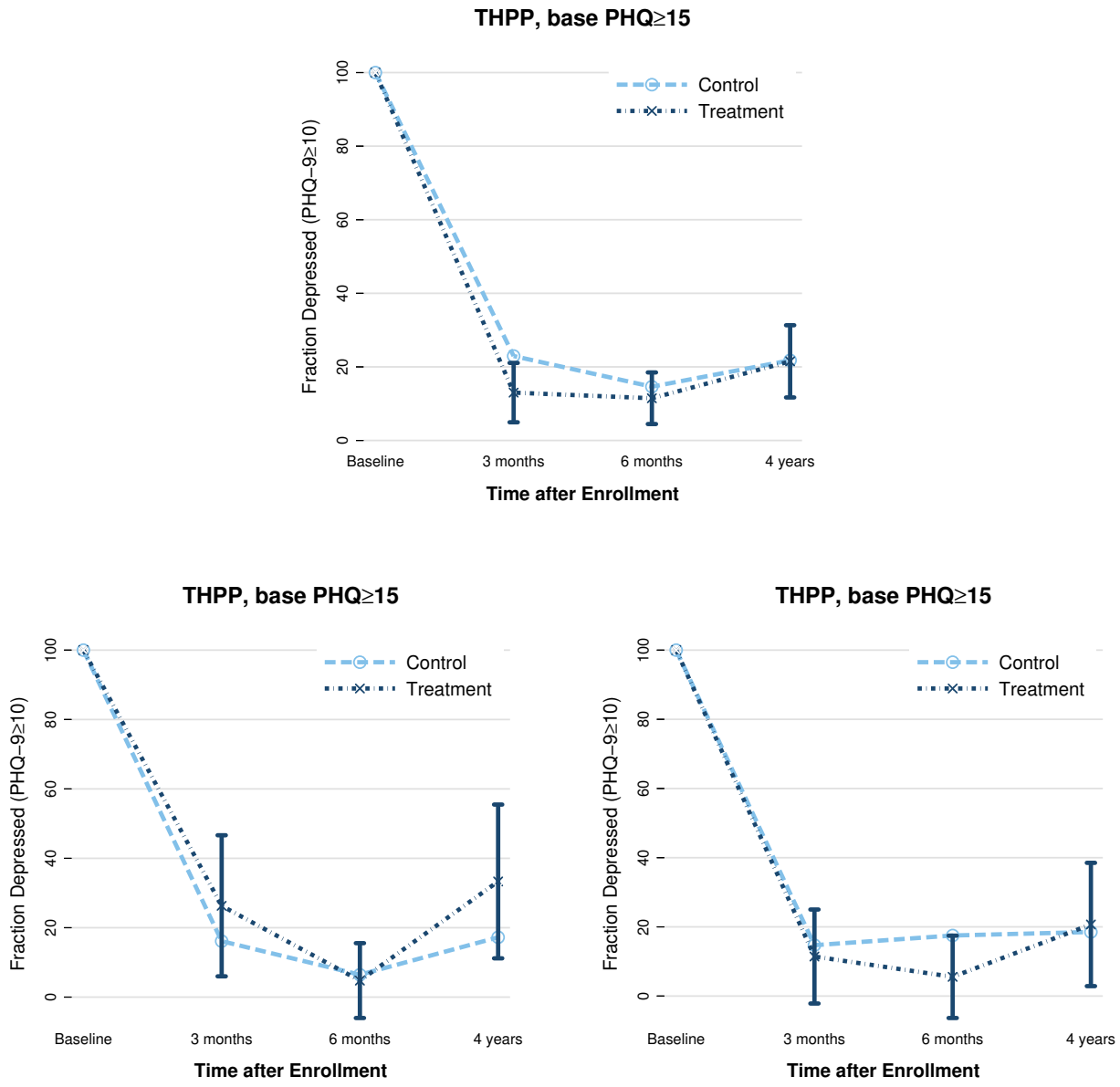
Figure A.4: Machine learning-based heterogeneous treatment effects



	Full Sample		
	Top Quartile	Bottom Quartile	Difference (p-value)
4/5 Year PHQ-9	0.07	-1.67	-1.74 (.40)

Notes: This figure and table presents heterogeneity in treatment effects across 4 groups identified using the machine learning approach of Chernozhukov et al. (2017). Effects on all quartiles are shown in the figure, and the table presents tests for differences between the top and bottom quartile's effects on the final PHQ-9 score.

Figure A.5: Treatment e□ects on depression for di□erent sub-samples of THPP

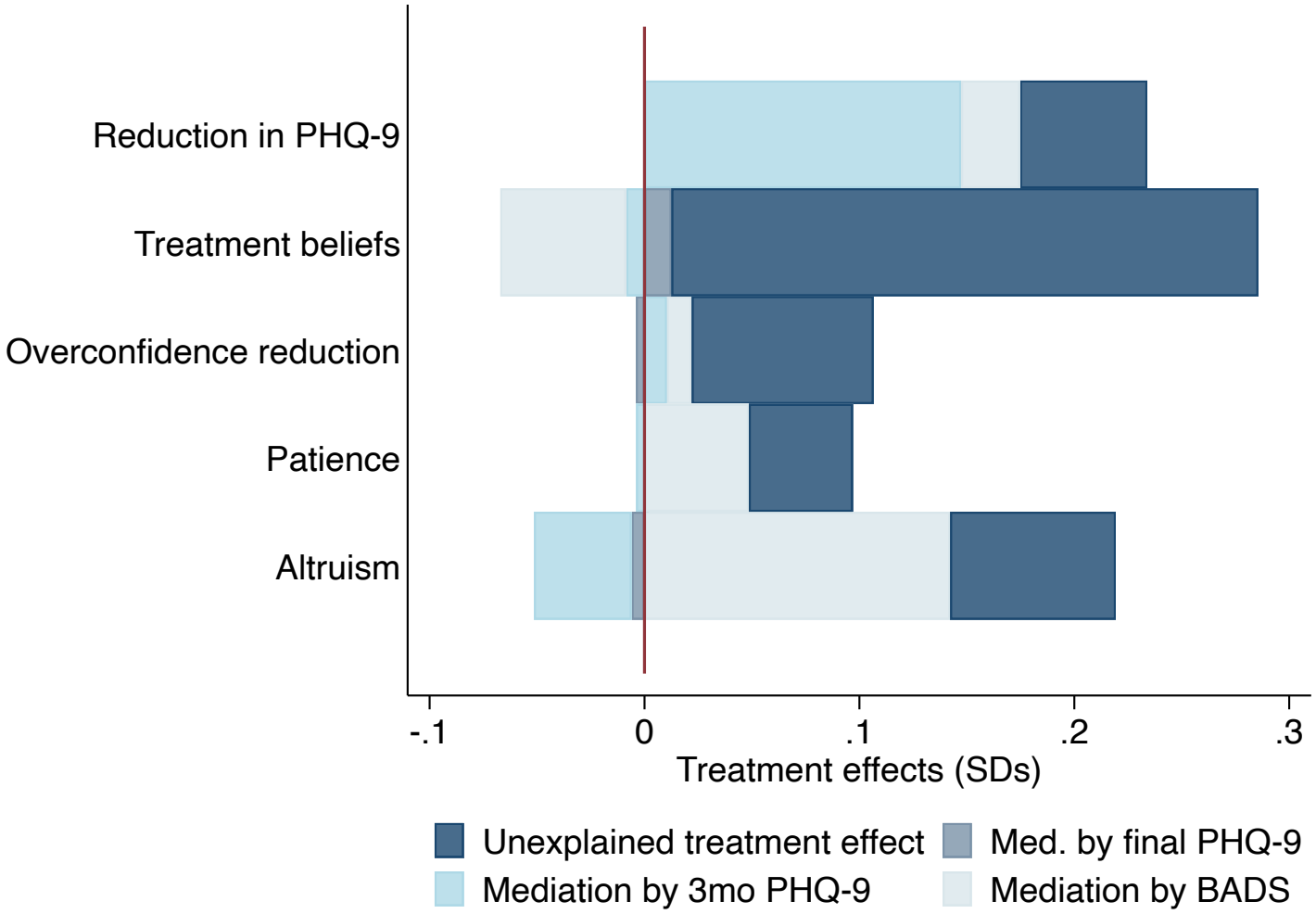


Notes: This figure shows the treatment effect trajectories on the fraction depressed in the treatment and control groups in the THPP trial, with different sub-samples taken to make the THPP sample more similar to the HAP sample.

The top panel presents the fraction of individuals depressed (PHQ-9 score 10) over time for the control and treatment group.

The bottom panels report the THPP results restricting the sample to be more similar to HAP. On the left, we restrict the sample to participants who were older than age 30 at baseline. On the right, we restrict the sample to participants with higher baseline depression (PHQ-9 15) and age (age>30).

Figure A.6: Treatment effect mediation



Notes: This figure shows the results from a mediation analysis of the treatment effects on key outcomes.

Bars show the treatment effect sizes in terms of standard deviations of each outcome, estimated with OLS. Sub-components show proportion of each treatment effect explained by 'mediating' treatment effects through mediator variables, estimated using the following equations:

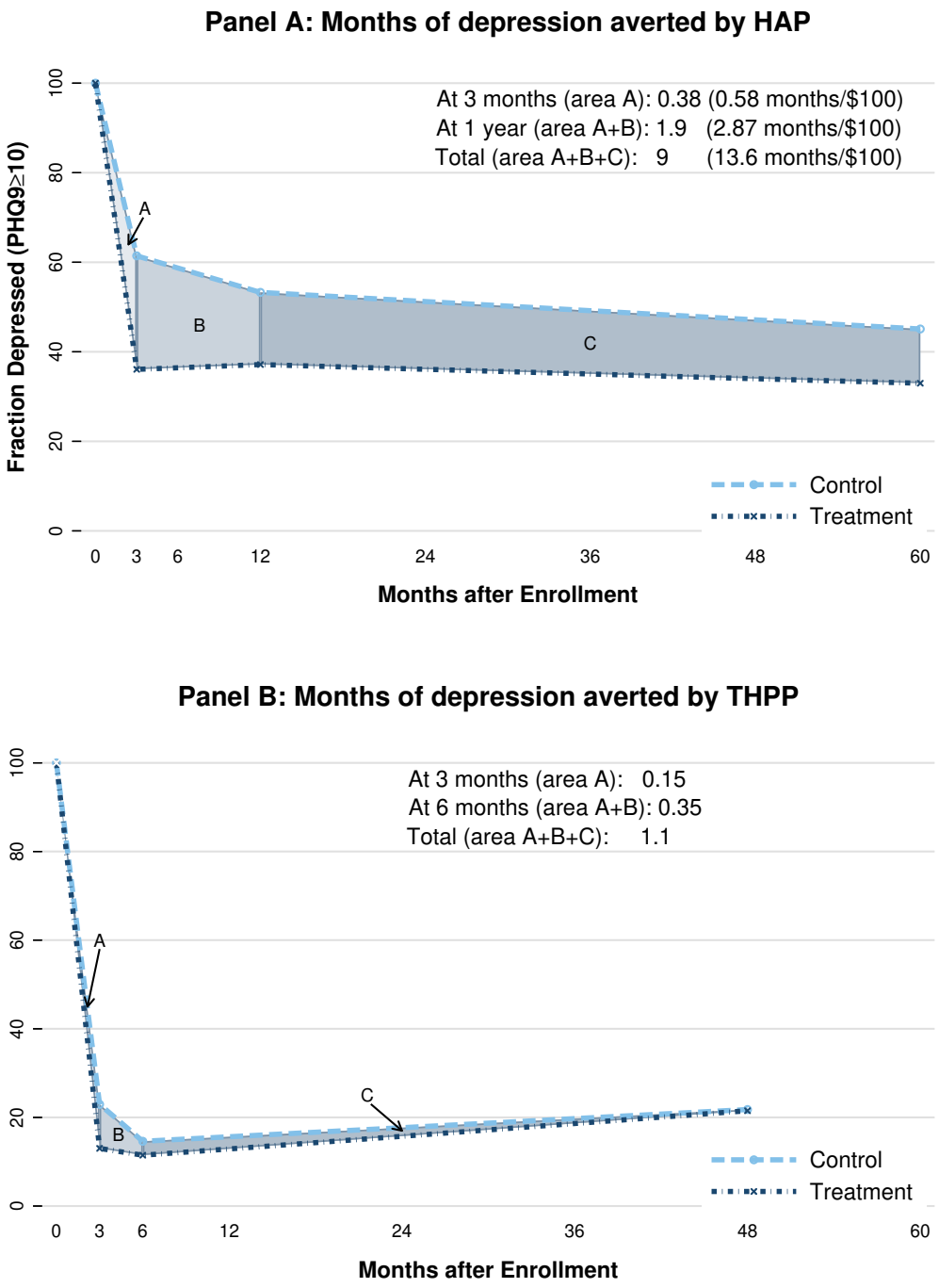
- Outcome $= \alpha_0 + \alpha_1 Treatment + \alpha_2 Mediator + \varepsilon_\alpha$
- Outcome $= \beta_0 + \beta_1 Treatment + \varepsilon_\beta$
- Mediator $= \gamma_0 + \gamma_1 Treatment + \varepsilon_\gamma$
- And the following: Outcome $= \beta_0 + \beta_1 Treatment + \varepsilon_\beta = (\alpha_0 + \alpha_2\gamma_0) + (\alpha_1 + \alpha_2\gamma_1) Treatment + (\epsilon_\alpha + \alpha_2\epsilon_\gamma)$
- Total Effect represents β_1, and mediation effects represent $\alpha_2\gamma_1$, the effect of treatment 'through' the mediator.

Thus, these equations can be used to estimate the proportion of the overall treatment effect on each outcome that can be attributed to the effect of treatment on the mediating variables, which then have impacts on the outcome of interest (which can potentially oppose the main treatment effect).

We focus on three potential mediating variables: (i) PHQ-9 scores at 3 months; (ii) PHQ-9 scores at endline; and (iii) BADS (see below). Dark blue shows the proportion of each treatment effect that is 'unexplained' by the three mediators presented here. Reduction in final PHQ-9 does not include itself in the model as a mediator.

BADS represents the 'behavioral activation' score, which assess the extent to which individuals take part in mood-improving activities, and is the total score from an index of questions including "Are you content with the amount and types of things you have done?", "Do you engage in many different activities?", "Are you an active person and have you accomplished the goals you set out to do?", "Do you spend long periods thinking over and over about your problems?", "Do you do things that were enjoyable?".

Figure A.7: Months of depression averted by the interventions



Notes: This figure shows estimates of the months of depression averted by the HAP intervention, shown in Panel A, and by the THPP intervention, shown in Panel B.

The figure shows the impacts of the two trials on remission from depression, as measured by PHQ-9 scores below 10, reproduced from Figure 3.

For each trial, we calculate the number of months of depression per participant averted by the treatment at three points in time: (i) at three months; (ii) at 6 months or 1 year; and (iii) at 4 to 5 years.

The months of depression averted is calculated as the integral of the (shaded) difference between the treatment and control curves plotted in the figures for each time interval.

Cost-effectiveness for HAP is calculated using a per-capita administration cost of $66.

Table A.1: Balance by trial

	HAP				THPP			
	Baseline (N=493)		Follow-up (N=391)		Baseline (N=280)		Follow-up (N=192)	
	Control mean (S.D.)	Δ Treatment (S.E.)	Control mean (S.D.)	Δ Treatment (S.E.)	Control mean (S.D.)	Δ Treatment (S.E.)	Control mean (S.D.)	Δ Treatment (S.E.)
Completed this stage	1.00	0.00	0.81	−0.04	1.00	0.00	0.71	−0.06
			(0.39)	(0.04)			(0.45)	(0.06)
Baseline PHQ-9 Score	17.88	0.06	17.79	0.13	13.09	0.56	12.79	0.82
	(2.85)	(0.25)	(2.80)	(0.27)	(3.22)	(0.40)	(2.82)	(0.47)
Duration of depression at baseline (weeks)	42.22	3.02	44.30	0.50	10.76	−0.99	10.93	−1.29
	(117.38)	(10.29)	(127.14)	(12.52)	(13.68)	(1.35)	(13.49)	(1.62)
Years between baseline and follow-up	-	-	4.86	0.03	-	-	3.99	−0.03
			(0.66)	(0.07)			(0.47)	(0.07)
Number of children	-	-	2.16	0.09	-	-	2.17	−0.20
			(1.30)	(0.14)			(0.85)	(0.12)
Female	0.77	0.00	0.82	−0.02	1.00	0.00	1.00	0.00
	(0.42)	(0.04)	(0.39)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)
Age	42.61	−0.19	43.46	0.47	25.31	−0.28	25.90	−0.32
	(11.97)	(1.08)	(11.51)	(1.17)	(4.74)	(0.55)	(4.82)	(0.68)
Married	0.69	−0.01	0.72	−0.01	1.00	−0.01	1.00	−0.01
	(0.46)	(0.04)	(0.45)	(0.05)	(0.00)	(0.01)	(0.00)	(0.01)
Years of education	5.96	−0.62	5.69	−0.63	7.49	0.06	7.36	−0.19
	(4.62)	(0.42)	(4.69)	(0.45)	(4.20)	(0.57)	(3.42)	(0.59)
Hindu	0.90	0.02	0.92	0.01	0.57	0.00	0.54	0.06
	(0.30)	(0.03)	(0.28)	(0.03)	(0.50)	(0.06)	(0.50)	(0.07)
Expected treatment e□cacy	0.74	−0.02	0.75	−0.04	0.79	0.03	0.77	0.03
	(0.44)	(0.04)	(0.44)	(0.05)	(0.41)	(0.05)	(0.42)	(0.06)
Homemaker	0.47	0.02	0.51	0.01	0.74	0.01	0.75	−0.03
	(0.50)	(0.05)	(0.50)	(0.05)	(0.44)	(0.05)	(0.44)	(0.06)
Unemployed	0.08	0.04	0.05	0.06	0.10	−0.01	0.10	0.00
	(0.27)	(0.03)	(0.22)	(0.03)	(0.30)	(0.03)	(0.30)	(0.04)
Employed	0.44	−0.06	0.43	−0.07	0.16	−0.01	0.15	0.02
	(0.50)	(0.04)	(0.50)	(0.05)	(0.37)	(0.04)	(0.36)	(0.05)
F test	0.66		0.78		0.55		0.74	
p-value	0.77		0.68		0.85		0.69	

Notes: This table presents characteristics of the RCT samples, stratified by trial. The characteristics of the combined RCT sample are presented in Table 1. For each trial, the first two columns present baseline mean characteristics in the control group and the difference between mean characteristics in the treatment and control groups, among all participants in a category. The last two columns present the same characteristics and differences among only those individuals who also appear in the follow-up trial. F-tests from a regression of the treatment dummy on these characteristics in each sample are also presented.

Baseline PHQ-9 score is a standard diagnostic test for depression; it ranges from 0 to 27, with a score greater than or equal to 10 indicating depression. The baseline score reported is the score at enrollment in the initial trial (before the interventions).

'Female' takes the value 1 for female participants and 0 for male participants. 'Expected treatment efficacy' takes the value 1 if the participant expected the treatment to be at least 'somewhat useful', and 0 if the patient expected the treatment to be 'a little useful' or less. 'Homemaker' takes the value of 1 if participant does not work for pay outside the home and is not search for such work currently.

Table A.2: Impacts on depression PHQ-9 sub-component

	Full Sample		HAP		THPP	
	Control mean (S.D.)	Treatment Effect (S.E.)	Control mean (S.D.)	Treatment Effect (S.E.)	Control mean (S.D.)	Treatment Effect (S.E.)
PHQ Questions						
Sleeping Difficulty	1.22	−0.14	1.40	−0.16	0.86	−0.08
	(1.15)	(0.09)	(1.15)	(0.11)	(1.06)	(0.15)
Tiredness	1.39	−0.13	1.52	−0.17	1.12	−0.04
	(1.08)	(0.09)	(1.06)	(0.11)	(1.07)	(0.16)
Poor Appetite	0.89	−0.23**	1.00	−0.35***	0.66	0.03
	(1.15)	(0.09)	(1.19)	(0.11)	(1.02)	(0.15)
Trouble Concentrating	0.84	−0.04	0.96	−0.08	0.61	0.02
	(1.12)	(0.09)	(1.17)	(0.11)	(1.00)	(0.14)
Little Interest/Pleasure	0.77	0.06	0.90	0.01	0.50	0.12
	(1.08)	(0.09)	(1.14)	(0.11)	(0.88)	(0.13)
Feeling Depressed	1.12	−0.14	1.28	−0.21*	0.79	0.00
	(1.15)	(0.09)	(1.21)	(0.12)	(0.94)	(0.13)
Feeling Bad About Oneself	1.05	−0.24***	1.25	−0.38***	0.64	0.06
	(1.19)	(0.09)	(1.22)	(0.11)	(1.01)	(0.15)
Abnormal Speech or Movement	0.35	−0.04	0.37	−0.02	0.32	−0.08
	(0.75)	(0.06)	(0.74)	(0.08)	(0.79)	(0.11)
Better off Dead/Self Harm	0.33	0.00	0.41	−0.03	0.18	0.06
	(0.76)	(0.06)	(0.82)	(0.08)	(0.59)	(0.09)
N	589		395		194	

Notes: This table presents estimates of the effects of the treatments on depression on each PHQ-9 sub-component, as measured in the follow-up study, several years after the initial interventions.

The first two columns show the impacts for the full sample, i.e., pooling the two trials. The following columns show the impacts for the two trials separately. Odd columns report control-group means, along with standard deviations in parentheses. Even columns report treatment effects along with standard errors in parentheses.

Treatment effects are estimated using the double machine learning approach of Chernozhukov et al. (2016), using all control variables available in the baseline data.

PHQ-9 score is a standard diagnostic test of depression measured on a scale from 0 to 27. Each PHQ-9 sub-component represents one of the nine questions that comprise the index, and each question is scored from 0-3 where 0 represents that a given symptom occurs "Not at all" and 3 represents that it occurs "Nearly every day".

Table A.3: Impacts of the treatments on depression (PHQ-9 score): heterogeneity

		X = Above Median:					X =
	Base Model (S.E.)	Baseline PHQ-9 (S.E.)	Age (S.E.)	Years of Education (S.E.)	Predicted PHQ-9 Drop (S.E.)	Expected E□cacy (S.E.)	Female Gender (S.E.)
Full Sample							
Treatment E□ect	−0.97**	−1.04*	0.23	−0.67	−1.30*	−0.90	−1.22
	(0.48)	(0.63)	(0.75)	(0.58)	(0.71)	(0.75)	(1.10)
X		2.12***	2.77***	−0.11	−1.38**	−1.09	1.71*
		(0.70)	(0.66)	(0.77)	(0.66)	(0.68)	(0.91)
Treatment * X		−0.22	−2.19**	−1.06	0.77	−0.12	0.33
		(1.00)	(0.99)	(1.23)	(0.98)	(1.00)	(1.08)
THPP							
Treatment E□ect	−0.04	−0.18	0.08	−0.22	−0.10	−0.46	
	(0.77)	(0.85)	(0.81)	(1.06)	(0.99)	(1.23)	
X		−0.63	0.33	−0.18	0.18	0.06	
		(1.19)	(2.40)	(0.99)	(1.00)	(1.00)	
Treatment * X		1.13	−2.28	0.40	0.14	0.64	
		(2.04)	(3.11)	(1.57)	(1.60)	(1.58)	
HAP							
Treatment E□ect	−1.43**	−1.84**	0.42	−0.84	−1.97**	−1.45	−1.22
	(0.61)	(0.92)	(1.54)	(0.68)	(0.90)	(0.91)	(1.10)
X		1.05	1.02	−0.26	−2.41***	−0.18	3.23***
		(0.83)	(1.13)	(1.18)	(0.81)	(0.84)	(0.96)
Treatment * X		0.48	−2.35	−2.98*	1.18	0.01	−0.17
		(1.24)	(1.67)	(1.54)	(1.21)	(1.23)	(1.30)

Notes: This table presents heterogeneity in the treatment effects on depression across samples, as measured in the follow-up study, several years after the initial interventions.

Each column reflects stratification of the sample below and above the median on a given characteristic. The last column splits the sample by gender.

The rows show the measured treatment effect, the estimated correlation between depression and characteristic 'X' independent of treatment status (in the control group), and the modifying effect of 'X' on the treatment effect, estimated using a single OLS regression.

The first panel pools participants from both trials. The second panel focuses on the THPP sample, which comprises only females. The third panel focuses on the HAP sample.

Table A.4: Impacts on beliefs about treatment effects (more detailed version)

	Full Sample		HAP		THPP	
OLS without controls	Control mean (S.D.)	Treatment Effect (S.E.)	Control mean (S.D.)	Treatment Effect (S.E.)	Control mean (S.D.)	Treatment Effect (S.E.)
Panel A: Effects after 3 Months						
True Remission from Depression (PHQ-9 < 10)	0.52	0.20***	0.39	0.25***	0.77	0.10**
(N = 711)	(0.50)	(0.03)	(0.49)	(0.04)	(0.42)	(0.05)
Control Group's Belief	0.64	0.12***	0.63	0.10***	0.65	0.13***
(N = 232)	(0.20)	(0.02)	(0.21)	(0.03)	(0.19)	(0.02)
Treatment Group's Belief	0.60	0.12***	0.59	0.10***	0.60	0.15***
(N = 218)	(0.22)	(0.02)	(0.22)	(0.03)	(0.21)	(0.03)
p-value (Control Group's Belief = Actual Data)	0.00	0.06	0.00	0.01	0.00	0.55
p-value (Treatment Group's Belief = Actual Data)	0.02	0.07	0.00	0.01	0.00	0.41
p-value (Treatment Group's Belief = Control Group's Belief)	0.03	0.93	0.16	0.87	0.09	0.70
Panel B: Effects after 6 Months/1 Year						
True Remission from Depression (PHQ-9 < 10)	0.61	0.11***	0.47	0.16***	0.85	0.03
(N = 699)	(0.49)	(0.03)	(0.50)	(0.05)	(0.35)	(0.04)
Control Group's Belief	0.70	0.06***	0.70	0.02	0.71	0.10***
(N = 232)	(0.21)	(0.02)	(0.23)	(0.03)	(0.19)	(0.03)
Treatment Group's Belief	0.62	0.16***	0.63	0.13***	0.62	0.20***
(N = 218)	(0.22)	(0.02)	(0.22)	(0.03)	(0.22)	(0.03)
p-value (Control Group's Belief = Actual Data)	0.00	0.20	0.00	0.03	0.00	0.17
p-value (Treatment Group's Belief = Actual Data)	0.63	0.33	0.00	0.65	0.00	0.00
p-value (Treatment Group's Belief = Control Group's Belief)	0.00	0.00	0.01	0.00	0.00	0.02
Panel C: Effects after 4/5 Years						
True Remission from Depression (PHQ-9 < 10)	0.63	0.08**	0.55	0.12**	0.78	0.00
(N = 589)	(0.48)	(0.04)	(0.50)	(0.05)	(0.41)	(0.06)
Control Group's Belief	0.70	0.07***	0.68	0.07**	0.73	0.08**
(N = 232)	(0.22)	(0.02)	(0.22)	(0.03)	(0.22)	(0.03)
Treatment Group's Belief	0.62	0.15***	0.62	0.13***	0.61	0.18***
(N = 218)	(0.24)	(0.02)	(0.25)	(0.03)	(0.23)	(0.03)
p-value (Control Group's Belief = Actual Data)	0.02	0.83	0.00	0.36	0.23	0.23
p-value (Treatment Group's Belief = Actual Data)	0.75	0.14	0.12	0.84	0.00	0.01
p-value (Treatment Group's Belief = Control Group's Belief)	0.00	0.01	0.02	0.08	0.00	0.03

Notes: Expanding Table 3, this table reports participants' beliefs about remission from depression (PHQ-9 < 10) in the treatment and control group, and the implied beliefs about the treatment effects on depression. Estimation uses the double machine learning approach of Chernozhukov et al. (2016).

Participants are asked, separately for the treatment group and the control group, out of 10 randomly selected individuals how many would have had their depression "reduced to healthy levels." We use their answers to construct their implied belief about the treatment effect on remission from depression.

For each time horizon, we report (i) the true (estimated) remission from depression; (ii) the control group's beliefs about remission from depression in the control group and about the treatment effect on it; (iii) the treatment group's beliefs about remission from depression in the control group and about the treatment effect on it.

The bottom three rows of each panel report the p-values corresponding to null hypotheses regarding whether or not the mean beliefs of treatment and control participants, respectively, are equal to each other or equal to the "true" estimated effect.

Table A.5: Construction of main outcomes

Outcome		Description
Overconfidence		
Initial Overconfidence	$	Overconfidence about own relative performance on the bracelet task, prior to observing any signals. Computed as the difference between the participant's prior about their probability of performing in the top half of their group, minus a full-information benchmark assuming randomly formed groups drawn from the population performance distribution.
Final Overconfidence	$	Overconfidence about own relative performance on the bracelet task, after observing five signals. Computed as the difference between the participant's posterior following five binary signals (true with probability $^2/_3$), minus the Bayesian posterior given the full-information benchmark and the observed signals.
Beliefs: Updating		
Response to Good News	$	Regression estimate of the coefficient on positive signals in the belief updating task. Bayesian benchmark $\beta_H = 1$.
Response to Bad News	$	Regression estimate of the coefficient on negative signals in the belief updating task. Bayesian benchmark $\beta_L = 1$.
Asymmetry	$	Difference between estimated β_H and β_L.
Conservatism	$	Regression estimate of the coefficient on signals, forcing symmetric response to good and bad news. Bayesian benchmark $\beta = 1$
Hiring Scheme Decisions		
Accept ability-based contract	$	Each participant is offered a choice between Rs. 300 for sure, or a risky job opportunity. Under the risky opportunity, if the participant's (not-yet-revealed) bracelet-making performance was in the top half, they receive a job making 1,000 bracelets over 1 month for a wage of Rs. 3,000. Otherwise, they receive no job and no money. Variable equals 1 if they chose the risky opportunity.
Patience		
Saving a Note Task	$	Each participant is given a Rs. 100 note at session 1, and told they will receive Rs. 30 more if they bring the exact same note to session 2. Variable equals 1 if they bring the note to session 2.
Willing to give up today for future		GPS question "How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future?" (0–10 scale).
Willing to complete task early		Survey question "How willing you are to complete tasks at the earliest, and not leave them for later/postpone them?" (0–10 scale). Adapted from question 7(v) in the web appendix to Falk et al. (2016).
Discounting δ_a: today vs 12 months		GPS questionnaire. Elicited discount factor between hypothetical monetary amounts: Rs. 200 paid today vs varying amounts in 12 months' time.
Discounting δ_b: 12 vs 24 months		Elicited discount factor between hypothetical monetary amounts: Rs. 240 paid in 12 months vs varying amounts in 24 months' time. Adapted from GPS, amounts increased by 20% and shifted 12 months into the future.
Present bias β: δ_a/δ_b		Present bias coefficient calculated from δ_a and δ_b.
Risk Tolerance		
Risk Aversion Task	$	Monetary amount Y at which the participant prefers a 50-50 lottery paying gain Rs. 200/gain Rs. Y, to a sure amount of Rs. 100. Higher values mean lower risk tolerance. 5 choices with $Y \in (80, 60, 40, 20, 0)$.
Loss Aversion Task	$	Monetary amount Z at which the participant is willing to accept a 50-50 lottery paying gain Rs. 100/lose Rs. Z. Higher values mean higher loss tolerance. 7 choices with $Z \in (20, 40, 60, 80, 100, 150, 200)$.
Willing to choose uncertain outcomes		GPS question "In general, how willing you are to choose uncertain outcomes in real life?" (0–10 scale).
Altruism		
Dictator Game	$	Participant receives Rs. 50 and decides how much to give to another participant in the study, $Y \in (0, 5, \ldots, 50)$, keeping the remainder for themselves.
Willing to do good without expecting return		GPS question "How willing are you to give to good causes without expecting anything in return?" (0–10 scale).

Notes: $ denotes incentivized tasks. "GPS" denotes the Global Preference Survey (Falk et al., 2016, 2018). Appendix B provides details of how tasks were divided across experimental sessions.

Table A.6: E□ects of treatment on deviations from Bayesian updating

	Full Sample		HAP		THPP	
	Control mean (S.E.)	Treatment E□ect (S.E.)	Control mean (S.E.)	Treatment E□ect (S.E.)	Control mean (S.E.)	Treatment E□ect (S.E.)
Panel A: Degenerate Prior						
Degenerate Prior Before First Signal	0.04	0.01	0.04	0.03	0.05	−0.03
	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)	(0.03)
N		576		385		191
Panel B: Degenerate Posterior						
Degenerate Posterior After Any Signal	0.08	−0.01	0.09	−0.01	0.05	−0.03**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Degenerate Posterior After Positive Signal	0.10	0.00	0.12	0.01	0.07	−0.02
	(0.01)	(0.02)	(0.01)	(0.03)	(0.01)	(0.03)
Degenerate Posterior After Negative Signal	0.06	−0.01	0.07	−0.01	0.04	−0.02
	(0.00)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)
N		2880		1925		955
Panel C: Never Update At All						
Never Update At All	0.13	−0.03	0.11	0.00	0.16	−0.07
	(0.02)	(0.03)	(0.02)	(0.03)	(0.04)	(0.05)
N		576		385		191
Panel D: Update with Posterior = Prior						
No Update After Any Signal	0.44	−0.05***	0.42	−0.04*	0.48	−0.08**
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)
No Update After Positive Signal	0.45	0.05	0.44	0.05	0.48	0.04
	(0.01)	(0.04)	(0.01)	(0.04)	(0.02)	(0.06)
No Update After Negative Signal	0.42	−0.08***	0.40	−0.07**	0.48	−0.10**
	(0.01)	(0.03)	(0.01)	(0.03)	(0.02)	(0.05)
N		2880		1925		955
Panel E: Update in the wrong direction						
Wrong Update After Any Signal	0.18	0.02	0.19	0.03	0.16	−0.01
	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
Wrong Update After Positive Signal	0.13	0.00	0.13	0.02	0.13	−0.04
	(0.01)	(0.03)	(0.01)	(0.04)	(0.01)	(0.05)
Wrong Update After Negative Signal	0.22	0.02	0.23	0.02	0.18	0.01
	(0.01)	(0.02)	(0.01)	(0.03)	(0.01)	(0.04)
N		2880		1925		955
Panel F: At least one irrational update						
At least one irrational update	0.68	0.01	0.70	0.05	0.63	−0.07
	(0.03)	(0.04)	(0.03)	(0.05)	(0.05)	(0.07)
N		576		385		191
Panel G: Comprehension score						
Correctly answered comprehension questions	15.36	−0.09	15.20	−0.10	15.68	−0.04
	(0.16)	(0.25)	(0.20)	(0.31)	(0.28)	(0.41)
N		576		385		191

Notes: This table reports estimated treatment effects on the deviations from Bayesian updating in the belief-updating task.

Participants report a prior belief. They then receive a sequence of five signals, each followed by a posterior belief elicitation. A belief update is a data point that combines a prior belief, a posterior belief and a signal received. Panels A, C, F, and G consider outcomes at the individual level, while panels B, D, and E consider outcomes at the belief update level. When estimating effects on belief updates, standard errors are clustered at the individual level.

Panel A considers the propensity to report a degenerate prior belief (0 or 1).

Panel B presents effects on the propensity to update beliefs to a degenerate posterior after any signal, after a positive signal or after a negative signal.

Panel C presents the propensity of individuals to never update their belief over the course of the five signals sequence.

Panel D considers the frequency of belief updates where beliefs remain unchanged, following either type of signal.

Panel E presents the frequency of belief updates in the wrong direction, i.e. revising a belief optimistically after a negative signal or pessimistically after a positive signal.

Panel F considers the proportion of participants whose belief updating behavior includes at least one of the above deviations: at least one non-update, at least one degenerate belief, or at least one update in the wrong direction.

Panel G presents average comprehension score on a series of comprehension question that are asked after explaining the belief updating task to participants. 19 questions were asked to participants and graded 0 or 1, resulting in a score out of 19 points.

Table A.7: Robustness of belief updating effects

	Full Sample		HAP		THPP	
	Control mean (S.E.)	Treatment Effect (S.E.)	Control mean (S.E.)	Treatment Effect (S.E.)	Control mean (S.E.)	Treatment Effect (S.E.)
Panel A: All Belief Updates with Non Extreme Prior and Posterior						
Response to Good News[a] (β_H)	0.73	−0.21**	0.73	−0.22*	0.67	−0.20
	(0.08)	(0.10)	(0.10)	(0.13)	(0.10)	(0.14)
Response to Bad News[b] (β_L)	−0.09	0.06	−0.13	0.03	−0.03	0.11
	(0.06)	(0.09)	(0.08)	(0.11)	(0.07)	(0.12)
Asymmetry[c] ($\beta_H - \beta_L$)	0.82	−0.27*	0.85	−0.25	0.69	−0.32
	(0.12)	(0.15)	(0.15)	(0.19)	(0.14)	(0.21)
Conservatism[d] ($\bar{\beta}$)	0.29	−0.06	0.27	−0.07	0.29	−0.02
	(0.04)	(0.05)	(0.05)	(0.07)	(0.06)	(0.08)
N	2620		1715		905	
Panel B: Drop Participants Who Never Update At All						
Response to Good News[a] (β_H)	0.82	−0.28**	0.84	−0.30**	0.75	−0.25
	(0.09)	(0.11)	(0.12)	(0.14)	(0.11)	(0.15)
Response to Bad News[b] (β_L)	−0.12	0.10	−0.15	0.05	−0.05	0.16
	(0.07)	(0.10)	(0.09)	(0.12)	(0.08)	(0.14)
Asymmetry[c] ($\beta_H - \beta_L$)	0.94	−0.38**	0.98	−0.35	0.80	−0.41*
	(0.13)	(0.17)	(0.18)	(0.22)	(0.15)	(0.22)
Conservatism[d] ($\bar{\beta}$)	0.32	−0.07	0.31	−0.10	0.33	−0.02
	(0.05)	(0.06)	(0.06)	(0.08)	(0.06)	(0.09)
N	2355		1555		800	
Panel C: Drop Updates Where Posterior = Prior						
Response to Good News[a] (β_H)	1.10	−0.36***	1.08	−0.39**	1.06	−0.31
	(0.11)	(0.13)	(0.14)	(0.17)	(0.14)	(0.20)
Response to Bad News[b] (β_L)	−0.18	0.14	−0.20	0.02	−0.11	0.31
	(0.09)	(0.13)	(0.11)	(0.16)	(0.14)	(0.20)
Asymmetry[c] ($\beta_H - \beta_L$)	1.28	−0.50**	1.28	−0.41	1.17	−0.62**
	(0.16)	(0.21)	(0.20)	(0.26)	(0.22)	(0.31)
Conservatism[d] ($\bar{\beta}$)	0.43	−0.10	0.40	−0.16	0.47	−0.01
	(0.07)	(0.09)	(0.09)	(0.11)	(0.10)	(0.14)
N	1563		1056		507	
Panel D: Drop Updates in the Wrong Direction						
Response to Good News[a] (β_H)	0.89	−0.18*	0.93	−0.17	0.79	−0.21
	(0.08)	(0.10)	(0.10)	(0.12)	(0.11)	(0.15)
Response to Bad News[b] (β_L)	0.35	0.11	0.39	0.06	0.28	0.17*
	(0.04)	(0.07)	(0.06)	(0.09)	(0.06)	(0.10)
Asymmetry[c] ($\beta_H - \beta_L$)	0.54	−0.29**	0.54	−0.23	0.51	−0.38**
	(0.09)	(0.12)	(0.12)	(0.15)	(0.12)	(0.17)
Conservatism[d] ($\bar{\beta}$)	0.63	−0.03	0.67	−0.05	0.53	−0.01
	(0.04)	(0.06)	(0.06)	(0.08)	(0.06)	(0.09)
N	2123		1356		767	
Panel E: Drop Participants Who Have at Least One Irrational Update						
Response to Good News[a] (β_H)	0.76	−0.14	0.80	−0.17	0.70	−0.08
	(0.09)	(0.13)	(0.13)	(0.18)	(0.12)	(0.19)
Response to Bad News[b] (β_L)	0.33	0.04	0.34	−0.03	0.32	0.12
	(0.05)	(0.09)	(0.08)	(0.10)	(0.07)	(0.14)
Asymmetry[c] ($\beta_H - \beta_L$)	0.43	−0.19	0.46	−0.13	0.38	−0.20
	(0.11)	(0.16)	(0.16)	(0.20)	(0.13)	(0.23)
Conservatism[d] ($\bar{\beta}$)	0.56	−0.05	0.58	−0.11	0.51	0.03
	(0.05)	(0.08)	(0.07)	(0.11)	(0.07)	(0.12)
N	915		530		385	

Notes: This table complements Figure 6 and presents robustness of treatment effects on the belief updating outcomes using different sub-samples of the full study population.

The first two columns show the impacts for the full sample, i.e., pooling the two trials. The following columns show the impacts for the two trials separately. Odd columns report control-group means, along with standard deviations in parentheses. Even columns report treatment effects along with standard errors in parentheses.

Coefficients come from belief updating regressions (equation (3)). β_H and β_L measure responsiveness to positive and negative signals, respectively. A Bayesian is characterized by $\beta_H = \beta_L = 1$. $\beta_H - \beta_L$ is the difference between responsiveness to positive and negative signals, and equals 0 for a Bayesian. β measures responsiveness forcing the response to positive and negative signals to be identical, which equals 1 for a Bayesian updater. Standard errors are clustered at the individual level as these regressions include multiple observations for each individual.

Panel A utilizes the primary analysis sample (full sample except observations with degenerate beliefs, since the likelihood ratios are not defined in this case). Panel B drops individuals who never update their beliefs. Panel C drops individual observations where the participant did not update (posterior equals prior). Panel D drops observations with wrong-signed updates (negative updates following good news or positive updates following bad news). The coefficients become more mechanically more positive relative to Panel A. Panel E drops individuals who ever do not update or update in the wrong direction.

Table A.8: Cross-sectional correlation of belief updating with depression (control participants only)

	Full Sample		HAP		THPP	
	Control mean (S.E.)	Remission E□ect (S.E.)	Control mean (S.E.)	Remission E□ect (S.E.)	Control mean (S.E.)	Remission E□ect (S.E.)
Panel A. Overconfidence						
Initial Overconfidence	0.14	0.06	0.19	0.07	−0.07	0.01
	(0.04)	(0.04)	(0.04)	(0.05)	(0.08)	(0.09)
Final Overconfidence	0.18	0.02	0.22	0.03	0.02	−0.04
	(0.04)	(0.05)	(0.04)	(0.06)	(0.10)	(0.10)
N	299		199		100	
Panel B. Beliefs-updating coe□cients						
Response to Good News (β_H)	0.68	0.08	0.66	0.12	0.63	0.05
	(0.11)	(0.16)	(0.12)	(0.20)	(0.24)	(0.27)
Response to Bad News (β_L)	−0.07	−0.04	−0.12	−0.01	0.09	−0.17
	(0.07)	(0.12)	(0.09)	(0.16)	(0.10)	(0.13)
Panel C. Asymmetry and conservatism						
Asymmetry ($\beta_H - \beta_L$)	0.75	0.12	0.78	0.14	0.54	0.22
	(0.13)	(0.22)	(0.16)	(0.30)	(0.27)	(0.32)
Conservatism ($\bar{\beta}$)	0.26	0.04	0.23	0.09	0.33	−0.05
	(0.06)	(0.08)	(0.07)	(0.11)	(0.11)	(0.13)
N	1354		887		467	

Notes: This table mirrors Table 4. Instead of estimating the (causal) effect of treatment on belief updating, we report the (non-causal) correlation between remission from depression and belief updating within the control group. Remission is defined as having a PHQ-9 score less than 10.

Panel A uses the double machine learning approach of Chernozhukov et al. (2016). Panels B and C use the belief-updating regression specification (3). The "Control mean" and "Treatment effect" columns report standard errors in parentheses.

Panel A reports correlations with the level of beliefs. Initial Overconfidence equals the participant's initial belief about their probability of being in the upper half of performance in their group of ten people, minus the full-information benchmark, computed assuming groups drawn from the population performance distribution. Final overconfidence equals the participant's belief after observing five signals, minus the Bayesian posterior given the full-information benchmark and the observed signals.

Panel B reports belief updating coefficients, which measure the change in the posterior likelihood ratio in response to signals, relative to the Bayesian benchmark. β_H measures the response to good news, and β_L the response to bad news. Bayesian updating implies $\beta_H = 1$, and $\beta_L = 1$.

Panel C reports transformations of the belief updating coefficients. Asymmetry measures the difference between the response to good and bad news. Bayesian updating implies $\beta_H \quad \beta_L = 0$. Conservatism measures the response to news, forcing the coefficients on good and bad news to be identical. Bayesian updating implies $\beta = 1$.

Table A.9: Cross-sectional correlation of preferences with depression (control participants only)

	Full Sample		HAP		THPP	
	Control mean (S.D.)	Remission E□ect (S.E.)	Control mean (S.D.)	Remission E□ect (S.E.)	Control mean (S.D.)	Remission E□ect (S.E.)
Panel A. Altruism						
Anderson Index	0.00	−0.05	0.00	−0.01	0.00	−0.19
	(1.00)	(0.12)	(1.00)	(0.14)	(1.00)	(0.29)
Dictator Game	15.87	1.00	15.92	1.43	15.68	−0.02
	(10.80)	(1.37)	(11.22)	(1.68)	(9.17)	(2.41)
Willing to do good without expecting return	7.63	−0.37	7.45	−0.34	8.36	−0.57
	(2.47)	(0.31)	(2.59)	(0.38)	(1.79)	(0.56)
Panel B. Patience						
Anderson Index	0.00	0.10	0.00	0.07	0.00	0.16
	(1.00)	(0.11)	(1.00)	(0.13)	(1.00)	(0.23)
Saving a Note Task	0.80	0.02	0.83	0.01	0.68	0.00
	(0.40)	(0.05)	(0.38)	(0.06)	(0.48)	(0.11)
Willing to give up today for future	6.91	0.57*	6.93	0.42	6.82	1.00*
	(2.61)	(0.32)	(2.63)	(0.38)	(2.56)	(0.60)
Willing to complete task early	8.28	−0.10	8.14	−0.15	8.86	−0.04
	(2.49)	(0.29)	(2.66)	(0.36)	(1.55)	(0.41)
Discounting δ_a: today vs 12 months	0.64	0.00	0.65	0.01	0.58	−0.01
	(0.23)	(0.03)	(0.24)	(0.03)	(0.18)	(0.04)
Discounting δ_b: 12 months vs 24 months	0.58	0.04	0.59	0.03	0.55	0.02
	(0.19)	(0.03)	(0.21)	(0.03)	(0.14)	(0.04)
Present bias β: δ_a/δ_b	1.13	−0.04	1.15	−0.04	1.04	−0.06
	(0.38)	(0.04)	(0.41)	(0.05)	(0.27)	(0.06)
Panel C. Risk tolerance						
Anderson Index	0.00	0.08	0.00	0.04	0.00	0.22
	(1.00)	(0.12)	(1.00)	(0.14)	(1.00)	(0.26)
Risk Aversion Task	56.24	−2.45	59.43	−4.99	43.64	5.45
	(31.06)	(3.76)	(31.15)	(4.49)	(27.87)	(7.20)
Loss Aversion Task	72.29	−1.64	71.72	−1.30	74.55	−2.94
	(67.03)	(8.19)	(70.60)	(10.10)	(51.87)	(14.00)
Willing to choose uncertain outcomes	7.01	0.26	7.28	−0.08	5.95	1.26**
	(2.75)	(0.33)	(2.79)	(0.40)	(2.36)	(0.56)
N	290		195		95	

Notes: This table mirrors Table 5. Instead of estimating the (causal) effect of treatment on preferences, we report the (non-causal) correlation between remission from depression and preferences within the Control group. Remission is defined as having a PHQ-9 score less than 10.

All outcomes are standardized to mean zero, standard deviation 1 in the control group. All estimates use the double machine learning approach of Chernozhukov et al. (2016). "Control mean" columns report standard deviations in parentheses. "Treatment Effect" columns report standard errors in parentheses. Stars refer to unadjusted two-sided p-values at thresholds 0.1, 0.05, and 0.01, respectively. Each panel reports an inverse covariance weighted index measure over the sub-components reported in that panel (Anderson, 2008).

Panel A reports measures of altruism. "Dictator game" is computed from the amount of money out of Rs. 50 that the participant chose to send to another participant instead of keeping for themselves. "Willing to do good without expecting return" is a self-evaluation measure of altruism.

Panel B reports measures of patience. "Note saved in saving a Note Task" records whether the participant saved a Rs. 100 banknote for one week, earning a Rs. 30 return. "Willing to give up today for future" and "Willing to complete task early" are self-evaluation measures of patience. Discounting and present bias parameters are computed from choices over hypothetical sooner/later monetary amounts. The Anderson index is constructed excluding the present bias parameter, which is a transformation of two other components.

Panel C reports measures of risk tolerance. "Risk tolerance" and "Loss tolerance" are based on the participant's switching point in incentivized risk/loss lottery choice tasks, aligned so that positive numbers represent higher risk tolerance and higher loss tolerance. "Willing to choose uncertain outcomes" is a self-evaluation measure of risk tolerance.

Table A.10: Determinants of hiring choice

	(1)	(2)	(3)	(4)	(5)	(6)
Treatment	0.014	0.021	0.020	0.022	0.022	0.016
	(0.040)	(0.040)	(0.040)	(0.040)	(0.040)	(0.040)
Final posterior		0.18*		0.069	0.069	0.044
		(0.089)		(0.099)	(0.098)	(0.100)
Adjusted prior			0.33**	0.29*	0.30**	0.24*
			(0.10)	(0.12)	(0.11)	(0.11)
Reservation wage (Rs.)					-0.000080***	
					(0.000021)	
Risk switching point (Rs.)					-0.0019**	
					(0.00060)	
Constant	0.63***	0.51***	0.41***	0.39***	0.61***	0.53***
	(0.028)	(0.065)	(0.073)	(0.079)	(0.089)	(0.093)
Reservation wage dummies	No	No	No	No	No	Yes
Risk switching point dummies	No	No	No	No	No	Yes
N	576	576	576	576	562	562

Notes: This table presents treatment effect on, and correlations with an indicator for whether the participant chose to apply to the risky job opportunity at the end of the belief-updating task.

Each participant is offered a choice between Rs. 300 for sure, or a risky job opportunity. Under the risky opportunity, if the participant's (not-yet-revealed) bracelet-making performance was in the top half, they receive a job making 1,000 bracelets over 1 month for a wage of Rs. 3,000. Otherwise, they receive no job and no money. The hiring choice indicator equals 1 if they chose the risky opportunity.

Each column presents the result of a regression of the hiring choice indicator on a set of covariates.

'Treatment' is a treatment dummy; 'Adjusted prior' and 'Final posterior' are the reported probability of being in the top half of performance as elicited using the beaker task, respectively, before and after all signals; 'Reservation wage' and 'Risk switching point' are calculated in Rupees using the participant's answers to the multiple price list tasks for reservation wage and risk aversion.

In column (6), dummies for all possible values of reservation wage and risk switching point are included.

Table A.11: Female empowerment

	Full Sample		HAP		THPP	
	Control mean (S.D.)	Treatment E□ect (S.E.)	Control mean (S.D.)	Treatment E□ect (S.E.)	Control mean (S.D.)	Treatment E□ect (S.E.)
Female Empowerment						
Anderson Index	0.00	0.05	0.00	0.16	0.00	−0.11
	(1.00)	(0.09)	(1.00)	(0.13)	(1.00)	(0.14)
Female decision - what to cook	0.82	0.03	0.78	0.10*	0.87	−0.06
	(0.38)	(0.04)	(0.42)	(0.05)	(0.33)	(0.06)
Female decision - whether buy expensive item	0.36	−0.08*	0.43	−0.12*	0.27	−0.03
	(0.48)	(0.05)	(0.50)	(0.06)	(0.45)	(0.07)
Female decision - number of children	0.35	0.00	0.35	0.11	0.36	−0.13*
	(0.48)	(0.05)	(0.48)	(0.07)	(0.48)	(0.07)
Female decision - what to do if respondent is sick	0.30	−0.05	0.30	−0.03	0.31	−0.09
	(0.46)	(0.04)	(0.46)	(0.06)	(0.46)	(0.06)
Female decision - whether buy land/property	0.22	−0.01	0.24	0.02	0.19	−0.06
	(0.41)	(0.04)	(0.43)	(0.06)	(0.39)	(0.05)
Female decision - how much to spend on Social Function	0.36	−0.03	0.41	0.02	0.29	−0.09
	(0.48)	(0.05)	(0.49)	(0.07)	(0.46)	(0.06)
Female decision - what to do if child is sick	0.48	−0.03	0.55	−0.02	0.39	−0.04
	(0.50)	(0.05)	(0.50)	(0.07)	(0.49)	(0.07)
Female decision - whom child should marry	0.25	−0.01	0.34	−0.05	0.14	0.03
	(0.43)	(0.04)	(0.48)	(0.06)	(0.35)	(0.05)
Whether go out by yourself	0.64	0.07	0.57	0.09	0.73	0.06
	(0.48)	(0.04)	(0.50)	(0.06)	(0.45)	(0.06)
Whether eats together	0.60	−0.01	0.50	0.01	0.73	−0.05
	(0.49)	(0.05)	(0.50)	(0.07)	(0.45)	(0.07)
Whether participant's name on any bank account	0.89	0.02	0.88	0.10***	0.91	−0.06
	(0.31)	(0.03)	(0.33)	(0.04)	(0.29)	(0.05)
Whether participant's name on ownership paper	0.29	−0.05	0.41	−0.12*	0.14	0.02
	(0.45)	(0.04)	(0.49)	(0.06)	(0.35)	(0.05)
Whether talk with husband about work	0.76	0.08**	0.67	0.14**	0.87	0.01
	(0.43)	(0.04)	(0.47)	(0.06)	(0.33)	(0.05)
Whether talk with husband about spending money	0.81	−0.03	0.74	0.02	0.89	−0.10*
	(0.39)	(0.04)	(0.44)	(0.06)	(0.31)	(0.05)
Whether talk with husband about things in community	0.53	0.09*	0.45	0.18***	0.63	0.01
	(0.50)	(0.05)	(0.50)	(0.07)	(0.48)	(0.07)
Whether have cash in hand	0.76	0.04	0.74	0.06	0.78	0.02
	(0.43)	(0.04)	(0.44)	(0.06)	(0.42)	(0.06)
Don't need permission - go to local health center	0.46	0.03	0.51	0.06	0.41	0.00
	(0.50)	(0.05)	(0.50)	(0.07)	(0.49)	(0.07)
Don't need permission - visit relatives/friends	0.42	0.03	0.46	0.01	0.36	0.05
	(0.49)	(0.05)	(0.50)	(0.06)	(0.48)	(0.07)
Don't need permission - go to kirana shop	0.79	0.04	0.80	0.09*	0.78	−0.01
	(0.41)	(0.04)	(0.40)	(0.05)	(0.42)	(0.06)
Don't need permission - go to a short distance	0.47	0.05	0.55	0.07	0.38	0.01
	(0.50)	(0.05)	(0.50)	(0.07)	(0.49)	(0.07)
Can go alone - go to local health center	0.69	0.05	0.76	−0.02	0.61	0.12*
	(0.46)	(0.04)	(0.43)	(0.05)	(0.49)	(0.07)
Can go alone - visit relatives/friends	0.68	0.07	0.77	0.05	0.57	0.10
	(0.47)	(0.04)	(0.42)	(0.05)	(0.50)	(0.07)
Can go alone - go to kirana shop	0.88	0.00	0.91	−0.04	0.84	0.05
	(0.33)	(0.03)	(0.29)	(0.04)	(0.37)	(0.05)
Can go alone - go to a short distance	0.65	0.05	0.76	0.00	0.52	0.10
	(0.48)	(0.04)	(0.43)	(0.06)	(0.50)	(0.07)
Whether belong to some organization	0.44	−0.05	0.57	−0.03	0.28	−0.08
	(0.50)	(0.05)	(0.50)	(0.07)	(0.45)	(0.06)
N	396		215		181	

Notes: This table presents treatment effects on female empowerment. All estimates use the double machine learning approach of Chernozhukov et al. (2016).

Samples are restricted to females married at baseline.

A higher index value corresponds to a higher level of female empowerment, e.g., more reported autonomy in a female's decisions.

Table A.12: Intimate partner violence

	Full Sample		HAP		THPP	
	Control mean (S.D.)	Treatment E□ect (S.E.)	Control mean (S.D.)	Treatment E□ect (S.E.)	Control mean (S.D.)	Treatment E□ect (S.E.)
IPV						
Anderson Index	0.00	−0.11	0.00	−0.12	0.00	−0.12
	(1.00)	(0.08)	(1.00)	(0.13)	(1.00)	(0.10)
Been beaten (dummy)	0.13	−0.01	0.14	−0.04	0.11	0.02
	(0.33)	(0.03)	(0.35)	(0.04)	(0.31)	(0.04)
Been beaten (number of times)	0.85	−0.32	0.86	−0.16	0.84	−0.51
	(3.74)	(0.30)	(3.65)	(0.45)	(3.85)	(0.36)
Been forced to have sex (dummy)	0.09	−0.06**	0.10	−0.05	0.09	−0.07**
	(0.29)	(0.02)	(0.30)	(0.04)	(0.29)	(0.03)
Been forced to have sex (number of times)	0.33	−0.10	0.27	0.00	0.39	−0.22
	(1.73)	(0.17)	(1.03)	(0.26)	(2.30)	(0.21)
N	412		221		191	

Notes: This table presents treatment effects on intimate partner violence (IPV) towards women. All estimates use the double machine learning approach of Chernozhukov et al. (2016).

Samples are restricted to females married at baseline.

A higher index value corresponds to a larger level of reported violence.

Table A.13: Sleep, loneliness, and locus of control

	Full Sample		HAP		THPP	
	Control mean (S.D.)	Treatment E□ect (S.E.)	Control mean (S.D.)	Treatment E□ect (S.E.)	Control mean (S.D.)	Treatment E□ect (S.E.)
Panel A: Sleep						
Anderson Index	0.00	0.20**	0.00	0.20*	0.00	0.25*
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.15)
Hours asleep	5.54	0.27*	5.34	0.36*	5.95	0.09
	(1.75)	(0.15)	(1.93)	(0.20)	(1.21)	(0.18)
Hours in bed but not asleep	1.93	−0.14	2.13	−0.08	1.53	−0.26*
	(1.58)	(0.13)	(1.71)	(0.18)	(1.21)	(0.15)
Sleep quality	2.68	0.20**	2.61	0.20*	2.83	0.22
	(1.01)	(0.08)	(1.05)	(0.11)	(0.92)	(0.14)
N	557		375		182	
Panel B: Loneliness						
Total Loneliness Score	0.00	−0.09	0.00	−0.13	0.00	−0.03
	(1.00)	(0.09)	(1.00)	(0.11)	(1.00)	(0.15)
In tune with the people around me	0.95	−0.06	1.01	−0.09	0.84	−0.02
	(0.73)	(0.06)	(0.77)	(0.08)	(0.64)	(0.10)
No one really knows me well	0.81	−0.02	0.83	−0.04	0.78	0.05
	(0.74)	(0.06)	(0.76)	(0.08)	(0.69)	(0.10)
Can find companionship	0.98	−0.03	0.94	0.05	1.04	−0.18
	(0.77)	(0.06)	(0.77)	(0.08)	(0.78)	(0.11)
People around me but not with me	1.00	−0.02	1.06	−0.08	0.87	0.10
	(0.81)	(0.07)	(0.81)	(0.08)	(0.82)	(0.12)
N	557		375		182	
Panel C: Locus of Control						
Total LoC Score	0.00	0.04	0.00	−0.02	0.00	0.19
	(1.00)	(0.08)	(1.00)	(0.10)	(1.00)	(0.14)
How my life goes depends on me	2.50	0.04	2.53	−0.05	2.44	0.25
	(1.07)	(0.09)	(1.09)	(0.11)	(1.03)	(0.16)
Achievement in life is about fate or luck	1.11	−0.05	1.09	−0.04	1.13	−0.07
	(0.95)	(0.08)	(0.96)	(0.10)	(0.94)	(0.14)
Other people have a controlling influence	1.67	0.08	1.52	0.04	1.96	0.15
	(1.33)	(0.10)	(1.33)	(0.13)	(1.29)	(0.18)
One has to work hard in order to succeed	3.27	0.07	3.24	0.08	3.32	0.05
	(0.94)	(0.08)	(0.89)	(0.09)	(1.04)	(0.15)
Doubt my own abilities if meet di□culties	2.07	−0.08	2.05	−0.05	2.11	−0.14
	(1.32)	(0.11)	(1.37)	(0.14)	(1.22)	(0.18)
Innate abilities are more important than e□orts	1.28	0.01	1.18	0.01	1.46	0.00
	(1.06)	(0.09)	(1.04)	(0.10)	(1.07)	(0.16)
Have little control over happening in my life	1.30	0.05	1.29	−0.07	1.33	0.27*
	(1.05)	(0.09)	(1.08)	(0.11)	(0.98)	(0.16)
N	566		378		188	

Notes: This table presents treatment effects on sleep, loneliness and locus of control. All estimates use the double machine learning approach of Chernozhukov et al. (2016).

Panel A reports self-reported measures of sleep. "Hours in bed but not asleep" enters negatively in the index, for which a higher score means higher sleep quality.

Panel B shows measures of loneliness, using a short version of the UCLA Loneliness Score. The loneliness score is standardized. A higher loneliness score corresponds to the participant feeling more lonely, and each component enters the score accordingly. The full Loneliness statements read: Statement 1: I feel in tune with the people around me. Statement 2: No one really knows me well. Statement 3: I can find companionship when I want it. Statement 4: People are around me but not with me.

Panel C reports measures of locus of control. A higher LoC score corresponds to a more internal locus of control, i.e. that the respondent feels more in control of their lives. Each component is coded to enter the score accordingly.

Table A.14: Expert survey composition

First survey Position	Field						
	Economics	Other	Pol. Sci.	Psychiatry	Psychology	Pub. Pol.	Total
Faculty	35	2	0	0	5	0	42
Grad student	69	14	12	1	6	4	106
Non-Academic Researcher	7	2	0	0	1	0	10
Other	2	4	0	1	0	1	8
Postdoc	22	4	0	1	0	0	27
Practitioner	1	0	0	1	0	0	2
Research Coordinator/Assistant	7	0	0	0	0	0	7
Undergrad	2	2	0	0	0	0	4
Total	145	28	12	4	12	5	206
Second survey Respondents	1	2	0	25	0	0	28
Total							234

Notes: This table presents a breakdown of the participants in our expert survey by field and by position within their field. The survey was fielded in two waves. The second wave focused on increasing representation among psychiatrists. Since the second wave did not assess faculty rank or position, it is reported separately from the first survey, though the results from both surveys are combined for analysis.

Table A.15: Expert survey results

	Actual point estimate (SDs)	Percentile of expert predictions 10th	25th	50th	75th	90th	p-value (estimate = median)	Share experts outside 90% CI	Estimate percentile among experts
Panel A: Treatment effects on main outcomes									
PHQ-9 Score	-.23	-.23	-.15	-.08	0	.1	.14	.5	9
Initial Overconfidence	-.05	.03	.1	.2	.26	.35	.01	.71	3
Patience index	.17	-.01	.04	.12	.2	.3	.62	.25	61
Risk tolerance index	.03	-.15	-.04	.05	.15	.29	.85	.26	47
Altruism index	.25	0	.04	.1	.2	.31	.15	.42	82
Consumption	.05	-.12	0	.06	.15	.22	.93	.17	43
Employment	-.01	-.12	.05	.12	.2	.32	.22	.42	13
Panel B: Belief updating ratios in control and treatment groups									
Control	-.17	.63	.75	1	1.82	2.09	0	1	0
Treatment	-.19	.51	.67	.79	1.09	1.56	0	1	0

Notes: This table presents results from our expert survey, giving more details beyond Figure 8.

The first column presents the actual point estimate in standard deviations for each outcome estimated using the double machine learning procedure of Chernozhukov et al. (2016), restricting to the HAP sample.

'Percentile of expert predictions' shows the distribution of expert predictions of the treatment effects in standard deviations.

p-values report outcomes from t-tests on whether or not the median expert prediction is different from the empirically-derived point estimates for each outcome, using the standard errors calculated from the point estimate.

The same estimates and standard errors are used to construct 90% confidence intervals around the estimate, and the share of experts outside of that interval is presented.

Finally, the estimate's percentile among expert predictions is presented in the last column.

B Appendix B: Implementation Details

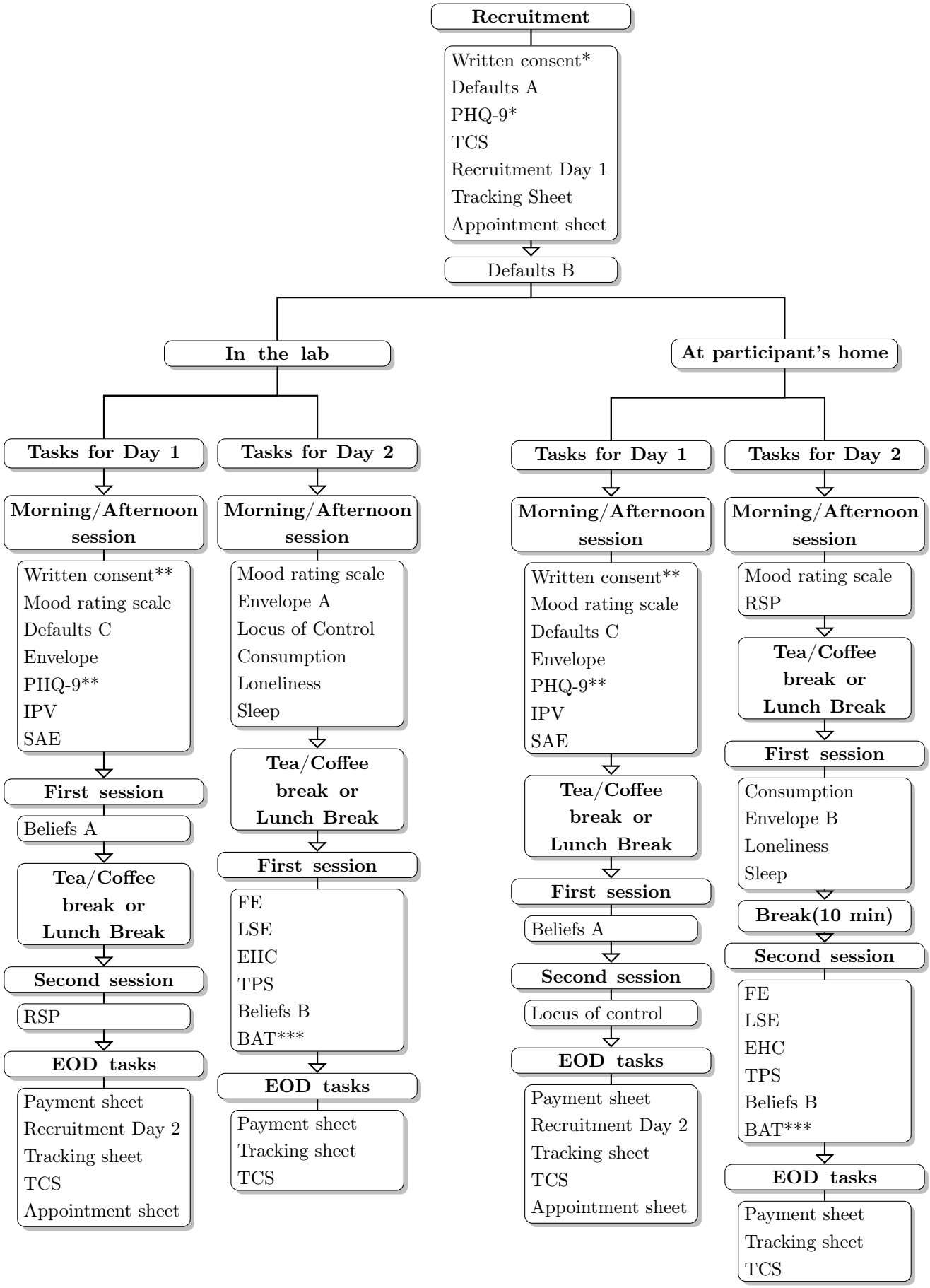
The study begins with the recruitment of the participant in which the participant is introduced to the study and their informed consent is obtained. This happens either over the phone or at participant's home, depending on how difficult it was to reach the participant during mobilization.

Experimental tasks and surveys took place over two days. Participants had the option to be interviewed at home or in our experimental lab, and the order of tasks was slightly different according to location, mainly driven by the fact that setting up the beliefs experiment took more time in participants' homes. Sessions could take place in the morning or afternoon. In addition, occasionally time constraints meant that not all survey modules could be completed in the time available.

The flowchart below shows the order of tasks. We use the following acronyms:

1. PHQ-9: Patient Health Questionnaire-9
2. IPV: Intimate Partner Violence
3. SAE: Serious Adverse Events
4. RSP: Risk and Social Preferences
5. TCS: Task Completion Sheet
6. FE: Female Empowerment
7. LSE: Labor Supply and Earnings
8. EHC: Education and Human Capital
9. TPS: Time Preferences Survey
10. BAT: Beliefs About Treatment
11. EOD: End of Day

In the flowchart, * shows that these tasks are supposed to be done only if the recruitment happened at participant's home, ** shows that these tasks are supposed to be done only if the recruitment happened over the phone, and *** shows that this survey was incorporated after the main study had started, so not all participants were asked these questions.


Recruitment
Written consent*
Defaults A
PHQ-9*
TCS
Recruitment Day 1
Tracking Sheet
Appointment sheet
Defaults B
In the lab
At participant's home
Tasks for Day 1
Tasks for Day 2
Tasks for Day 1
Tasks for Day 2
Morning/Afternoon session
Morning/Afternoon session
Morning/Afternoon session
Morning/Afternoon session
Written consent**
Mood rating scale
Defaults C
Envelope
PHQ-9**
IPV
SAE
Mood rating scale
Envelope A
Locus of Control
Consumption
Loneliness
Sleep
Written consent**
Mood rating scale
Defaults C
Envelope
PHQ-9**
IPV
SAE
Mood rating scale
RSP
First session
Beliefs A
Tea/Coffee break or Lunch Break
First session
FE
LSE
EHC
TPS
Beliefs B
BAT***
Tea/Coffee break or Lunch Break
First session
Beliefs A
Tea/Coffee break or Lunch Break
First session
Consumption
Envelope B
Loneliness
Sleep
Second session
RSP
EOD tasks
Payment sheet
Recruitment Day 2
Tracking sheet
TCS
Appointment sheet
EOD tasks
Payment sheet
Tracking sheet
TCS
Second session
Locus of control
EOD tasks
Payment sheet
Recruitment Day 2
Tracking sheet
TCS
Appointment sheet
Break(10 min)
Second session
FE
LSE
EHC
TPS
Beliefs B
BAT***
EOD tasks
Payment sheet
Tracking sheet
TCS

C Appendix C: Literature Review

Table A.16: Details of studies of psychological treatments' effects on depression

Authors	Paper Title	Depression Measure	Weeks since Treatment	Treatment Effect (SD)	Sample Size	Notes
Ali et al. 2003	The effectiveness of counseling on anxiety and depression by minimally trained counselors: a randomized controlled trial.	AKUADS	8	0.52	366	
Chen et al. 2000	Effects of support group intervention in postnatally distressed women: A controlled study in Taiwan	BDI	4	0.62	60	
Bolton et al. 2003	Group interpersonal psychotherapy for depression in rural Uganda: a randomized controlled trial	Depression scale	18	1.1	224	
Bolton et al. 2007	Interventions for Depression Symptoms Among Adolescent Survivors of War and Displacement in Northern Uganda	Local depression sympotom score	20	0.61	209	(IPT-G Group, Same Control Group)
Bolton et al. 2007	Interventions for Depression Symptoms Among Adolescent Survivors of War and Displacement in Northern Uganda	Local depression sympotom score	20	-0.16	209	(CP Group, Same Control Group)
Bolton et al. 2014a	A Transdiagnostic Community-Based Mental Health Treatment for Comorbid Disorders: Development and Outcomes of a Randomized Controlled Trial among Burmese Refugees in Thailand	Adapted HSCL-25 Depression scale	22	0.42	167	(CPT Group Group, Same Control Group)
Bolton et al. 2014a	A Transdiagnostic Community-Based Mental Health Treatment for Comorbid Disorders: Development and Outcomes of a Randomized Controlled Trial among Burmese Refugees in Thailand	Adapted HSCL-25 Depression scale	22	0.34	180	(BATD Group, Same Control Group)
Bolton et al. 2014b	A Transdiagnostic Community-Based Mental Health Treatment for Comorbid Disorders: Development and Outcomes of a Randomized Controlled Trial among Burmese Refugees in Thailand	Adapted HSCL-25 Depression scale	14	0.58	347	

Bass et al. 2006	Group interpersonal psychotherapy for depression in rural Uganda: 6-month outcomes	HSCL	16	1.54	284
Bass et al. 2006	Group interpersonal psychotherapy for depression in rural Uganda: 6-month outcomes	HSCL	40	1.38	216
Bass et al. 2013	Controlled Trial of Psychotherapy for Congolese Survivors of Sexual Violence	HSCL-25	16	1.5	405
Bass et al. 2013	Controlled Trial of Psychotherapy for Congolese Survivors of Sexual Violence	HSCL-25	40	1.33	405
Chibanda et al. 2014	Group problem-solving therapy for postnatal depression among HIV-positive and HIV-negative mothers in Zimbabwe	EPDS	12	0.67	58
Chibanda et al. 2016	E□ect of a primary care–based psychological intervention on symptoms of common mental disorders in Zimbabwe: a randomized clinical trial	SSQ-14	24	0.53	573
Rahman et al. 2008	Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: a cluster-randomised controlled trial	Hamilton	24	0.62	818
Rahman et al. 2008	Cognitive behaviour therapy-based intervention by community health workers for mothers with depression and their infants in rural Pakistan: a cluster-randomised controlled trial	Hamilton	48	0.82	798
Baranov et al. 2020	Maternal depression, women's empowerment, and parental investment: evidence from a randomized controlled trial	Hamilton	336	0.22	585
Weiss et al. 2015	Community-based mental health treatments for survivors of torture and militant attacks in Southern Iraq: a randomized control trial	HSCL-25	34	1.82	149
Weiss et al. 2015	Community-based mental health treatments for survivors of torture and militant attacks in Southern Iraq: a randomized control trial	HSCL-25	32	0.4	193
Tiwari et al. 2010	E□ect of an Advocacy Intervention on Mental Health in Chinese Women Survivors of Intimate Partner Violence	BDI-II (Chinese version)	12	0.21	200

Tiwari et al. 2010	E□ect of an Advocacy Intervention on Mental Health in Chinese Women Survivors of Intimate Partner Violence	BDI-II (Chinese version)	36	0.23	200
Rojas et al. 2007	Treatment of postnatal depression in low-income mothers in primary-care clinics in Santiago, Chile: a randomised controlled trial	EPDS	12	0.95	209
Rojas et al. 2007	Treatment of postnatal depression in low-income mothers in primary-care clinics in Santiago, Chile: a randomised controlled trial	EPDS	24	0.37	208
Maselko et al. 2020	E□ectiveness of a peer-delivered, psychosocial intervention on maternal depression and child development at 3 years postnatal: a cluster randomised trial in Pakistan	PHQ-9	144	0.13	572
Patel 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	BDI-II	12	0.57	495
Patel et al. 2010	E□ectiveness 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	CIS-R	24	0.064	673
Milani et al. 2015	E□ect of Telephone-Based Support on Postpartum Depression: A Randomized Controlled Trial	EPDS	6	0.56	54
Gao et al. 2010	Evaluation of an interpersonal-psychotherapy-oriented childbirth education programme for Chinese first-time childbearing women: A randomised controlled trial	EPDS	6	0.57	194
Gao et al. 2012	E□ects of an interpersonal-psychotherapy-oriented childbirth education programme for Chinese first-time childbearing women at 3-month follow up: Randomised controlled trial	EPDS	12	0.46	194
Gao et al. 2015	E□ects of an interpersonal-psychotherapy-oriented postnatal programme for Chinese first-time mothers: A randomized controlled trial	EPDS	6	0.23	180

Ho et al. 2009	E□ectiveness of a discharge education program in reducing the severity of postpartum depression: A randomized controlled evaluation study	EPDS	6	0.15	175	
Ho et al. 2009	E□ectiveness of a discharge education program in reducing the severity of postpartum depression: A randomized controlled evaluation study	EPDS	12	0.38	168	
Baker-Henningham et al. 2005	The e□ect of early stimulation on maternal depression: a cluster randomised controlled trial	CES-D	52	0.43	139	
Sikander et al. 2019	Delivering the Thinking Healthy Programme for perinatal depression through volunteer peers: a cluster randomised controlled trial in Pakistan	PHQ-9	24	0.13	453	
Rotheram-Borus et al. 2014	A Cluster Randomized Controlled Trial Evaluating the E□cacy of Peer Mentors to Support South African Women Living with HIV and Their Infants	GHQ	48	0.26	393	
Haushofer et al. 2020	The comparative impact of cash transfers and a psychotherapy program on psychological and economic well-being	GHQ	56	-0.03	4340	
Barker et al. 2021	Mental Health Therapy as a Core Strategy for Increasing Human Capital: Evidence from Ghana	Kessler	10	0.15	7412	
Me□ert et al. 2021	Interpersonal psychotherapy delivered by nonspecialists for depression and posttraumatic stress disorder among Kenyan HIV–positive women a□ected by gender-based violence: Randomized controlled trial	BDI-II	12	0.51	256	
Chowdhary et al. 2018	The Healthy Activity Program lay counsellor delivered treatment for severe depression in India: Systematic development and randomised evaluation	PHQ-9	8.6	0.45	55	
Petersen et al. 2014	A group-based counselling intervention for depression comorbid with HIV/AIDS using a task shifting approach in South Africa: A randomized controlled pilot study	PHQ-9	12	0.92	34	
Siddique et al. 2022	Forced Displacement, Mental Health, and Child Development: Evidence from the Rohingya Refugees	CESD-20	52	0.14	2845	