Different components of cognitive-behavioural therapy affect specific cognitive mechanisms

Agnes Norbury[1*], **Tobias U. Hauser**[2,3,4], **Stephen M. Fleming**[2,3,5], **Raymond J. Dolan**[2,3], **and Quentin J.M. Huys**[1,3]

1 Applied Computational Psychiatry Lab, Max Planck Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology and Mental Health Neuroscience Department, Division of Psychiatry, University College London, London, UK **2** Max Planck Centre for Computational Psychiatry and Ageing Research, Queen Square Institute of Neurology and Mental Health Neuroscience Department, Division of Psychiatry, University College London, London, UK **3** Wellcome Centre for Human Neuroimaging, University College London, London, UK **4** Department for Psychiatry and Psychotherapy, University of Tübingen, Germany **5** Department of Experimental Psychology, University College London, London, UK

*Corresponding Author: agnes.norbury@ucl.ac.uk

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

Abstract

Psychological therapies are among the most effective treatments for a range of common mental health problems – however, we still know relatively little about how exactly they improve symptoms. Here, we demonstrate the power of combing theory with computational methods to parse effects of different components of cognitive-behavioural therapies on to underlying mechanisms. Specifically, we present data from a series of randomized-controlled experiments testing the effects of components of behavioural and cognitive therapies on different cognitive processes, using well-validated behavioural measures and associated computational models (total N=807). We found that a goal-setting intervention, based on behavioural activation therapy, reliably and selectively reduced sensitivity to effort when deciding how to act to gain reward. By contrast, we found that a cognitive restructuring intervention, based on cognitive therapy, reliably and selectively reduced the tendency to attribute negative everyday events to self-related causes. Importantly, the effects of each intervention were specific to these respective measures. Our approach provides a basis for understanding how different elements of common psychotherapy programs work, which may enable theoretically-informed treatment targeting in the future.

Introduction

There is compelling evidence that psychotherapy programs as a whole are effective treatments for common mental health problems (Cuijpers et al., 2016, 2018; van Dis et al., 2020) (Figure 1a). However, psychotherapy programs are complex, multi-component interventions, and we still lack an understanding of *how* different components of these programs work (Kazdin, 2009; Holmes et al., 2014; Kazdin and Blase, 2011) (Figure 1b). Such insight is vital, as understanding the mechanisms underlying treatment response is one of the most promising routes to achieving many of the goals of mental health research – including increasing efficacy, engagement, and, ideally, theoretically-information treatment personalization (Huibers et al., 2021). Here, we argue that developments in the cognitive sciences concerning how to use robustly-designed behavioural tasks, in combination with rigorous modelling procedures that generate precise and reliable measures of cognitive processes, can accelerate progress towards these goals (Reiter et al., 2021; Huys et al., 2022).

In line with recent calls to the research community (Holmes et al., 2018; Wellcome, 2021), we take the pragmatic approach of starting from psychological therapies supported by a strong evidence base, and working back to theories regarding the mechanisms by which they work. Whilst cognitive and behavioural therapies are often administered together as part of the same treatment program (e.g., Clark 2018), they differ in underlying theory – for example, the primacy of behavioural *vs* cognitive changes in fostering improved mood (Beck et al., 1987; Martell et al., 2013). This distinction offers an opportunity to test whether these two types of interventions may work via different mechanisms – and whether there is specificity in their action via these proposed mechanisms.

Here, we present data from a series of studies investigating the mechanisms by which specific components of behavioural and cognitive therapies are proposed to work. We focus on a remote (online) setting, given the relative ease of delivering content in a standardized way, and the likely utility of a modular approach to treatment personalization for digitally-delivered therapies (see General Discussion). The first set of studies consisted of developing robust assessments of cognitive processes thought to be targeted by different components of cognitive and behavioural therapies (Figure 1c). Each assessment combined an optimized behavioural task with a theoretically-informed computational model, affording precise and reliable measurement of multiple different cognitive mechanisms (Figure 1d). Specifically, one set of measures was designed to probe constructs relevant to the goal-setting component of behavioural activation ("*how to decide when rewards are worth exerting effort for*"), and the other constructs relevant to the cognitive restructuring component of cognitive therapy ("*how to reason about likely causes of things that happen to us*"). In a second group of studies, we examined the extent to which these measures were sensitive to interventions based on each therapy component (Figure 1e). In a third study, we examined whether changes in cognitive mechanisms identified in the previous studies were *specific* to that particular intervention type (Figure 1f). Finally, we used data from studies two and three to explore to what extent individual differences in symptom profiles may relate to the magnitude of effects of each intervention on underlying cognitive mechanisms.

The overarching aim of this work is to demonstrate how creating reliable and acceptable measures of cognitive processes, drawn from relevant psychological theory, can identify mechanisms underlying psychological interventions. We believe our findings provide an important foundation towards establishing how real-world psychotherapy treatments may work, and who they are most likely to work for.

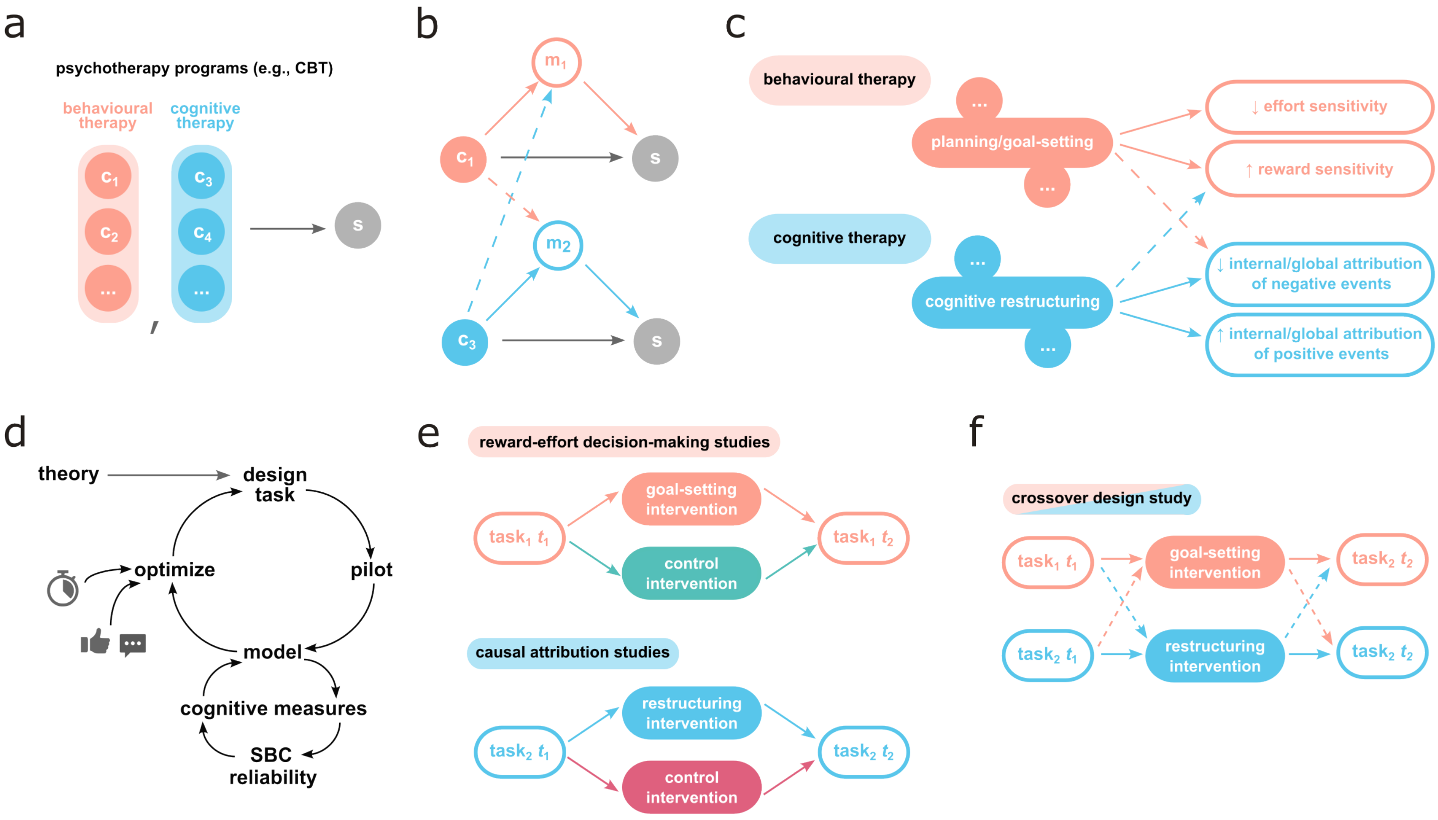


Figure 1: **The use of precise and reliable cognitive measures from computational cognitive science may help shed light on mechanisms targeted by components of common psychological therapies**. **a** At present, the majority of our causal knowledge regarding psychotherapy outcomes is at the level of how different treatment packages or programs affect symptom levels. c_x, different components of a given psychological treatment (e.g., behavioural or cognitive therapies; *orange and blue colours*), which may be further combined into psychotherapy treatment programs (e.g., cognitive-behavioural therapy, or CBT); *s*, symptoms; *solid arrows*, population-level evidence of causal influence. **b** Ideally, this knowledge can be further decomposed into a description of how different treatment components (c_x) affect symptoms via specific underlying mechanisms (m_x). Importantly, if different treatments work by at least partially distinct mechanisms, which can measure in a reliable way in a given individual, then this would yield an opportunity to develop personalized treatment packages. *dotted arrows*, effects not predicted under a specific mechanism model. **c** Potential mechanisms by which different components of psychological therapies improve symptoms may be drawn from underlying psychological theory. Here, we test whether psychological interventions based on components of behavioural and cognitive therapies are observed to influence different potential underlying cognitive mechanisms (*solid lines*), and whether these effects are specific to these mechanisms (*dotted lines*). **d** To generate precise and reliable estimates of potential cognitive mechanisms, task design and associated analysis methods first underwent several cycles of design optimization. This process included rigorous simulation-based calibration (SBC) analysis of model-based inference procedures and assessment of observed test-retest reliability of model-based measures. Key optimization targets alongside these measurement properties were task brevity and user-acceptability. *Continued on the next page*

Figure 1: Specifically, we developed a gamified reward-effort decision-making paradigm ($task_1$) that yields robust measurement of reward and effort sensitivity when deciding whether specific actions are worth taking, and a causal attribution task ($task_2$) that measures latent tendency to attribute positive and negative everyday events to internal (*vs* external) and global (*vs* specific) causes. **e** Next, we tested whether each of these sets of measures were sensitive to interventions based on relevant therapy components in a series of randomized, controlled experiments (i.e., the effects of a goal-setting intervention on reward/effort sensitivity, and the effects of a restructuring intervention on causal inference). **f** Finally, we used a cross-over design (where task and intervention conditions were independently randomly allocated) to test whether effects of interventions were specific to their proposed cognitive substrates. *NB* Arrows in panels d, e, f, represent experiment flow (over time, *t*), rather than causal influence.

Results

Developing useful measures for psychotherapy process research

Several considerations that motivated and guided our approach are worth outlining up-front. Firstly, in order to ensure reliable measurement of relevant cognitive mechanisms, each set of tasks and measures went through extensive rounds of design and analytic optimization prior to proceeding with the main studies (Figure 1d). To derive our computational measures, we used an analytical approach (Hierarchical Bayesian analysis; Gelman et al. 1995; Griffiths et al. 2008) previously shown to substantially increase the reliability of individual-level parameter estimates, by allowing information to be shared between relevant levels of analysis, and better accounting for measurement error (Katahira, 2016; Rouder and Haaf, 2019; Haines et al., 2020b; Brown et al., 2020). Secondly, for each prospective task design, rigorous testing of inference (model-fitting) procedures was carried out via simulation-based calibration (SBC; Talts et al. 2020), a general method for validating generative Bayesian algorithms using simulated datasets and posterior inference (Schad et al., 2021) (Methods, Figure S1). Following this analysis, the observed test-retest reliability of parameter estimates from the chosen design was explicitly assessed in our target population (Figure 2a,b).

Equally important as the above analytic considerations, useful individual difference measures should evoke robust differences between participants (Hedge et al., 2020; Zorowitz and Niv, 2023), and be acceptable (ideally engaging!) to their end users (Graham et al., 2019). Initial task development therefore proceeded via multiple informal cycles of 'user-in-the-loop' design optimization. Specifically, in order to maximise the magnitude of observable individual differences, care was taken to minimise range restriction (floor/ceiling) effects (Hedge et al., 2020; Zorowitz and Niv, 2023). In light of qualitative feedback from our participants and previous mental health studies using online tasks (Lee et al., 2023), major design optimization targets were task brevity (minimum trials to reliably detect target parameters), and basic levels of acceptability ("*I would be willing to play this game again in the future*"; Figure 2a,b). We attempted to increase the engagingness of our tasks via two different approaches: gamification (Cheng and Ebrahimi, 2023; Long et al., 2023), and providing opportunity for self-reflection or insight (Lee et al., 2023; Singh et al., 2021). Finally, where possible, we also tested measures for robustness to important sociodemographic differences between participants (Methods, Figure S2).

Although the above strategies necessarily involve some trade-offs against ideal psychometric properties and other important features (e.g., construct validity; see General Discussion), we believe that consideration of these issues at early stages is vital to the development of potentially clinically-useful measures (Paulus et al., 2016).

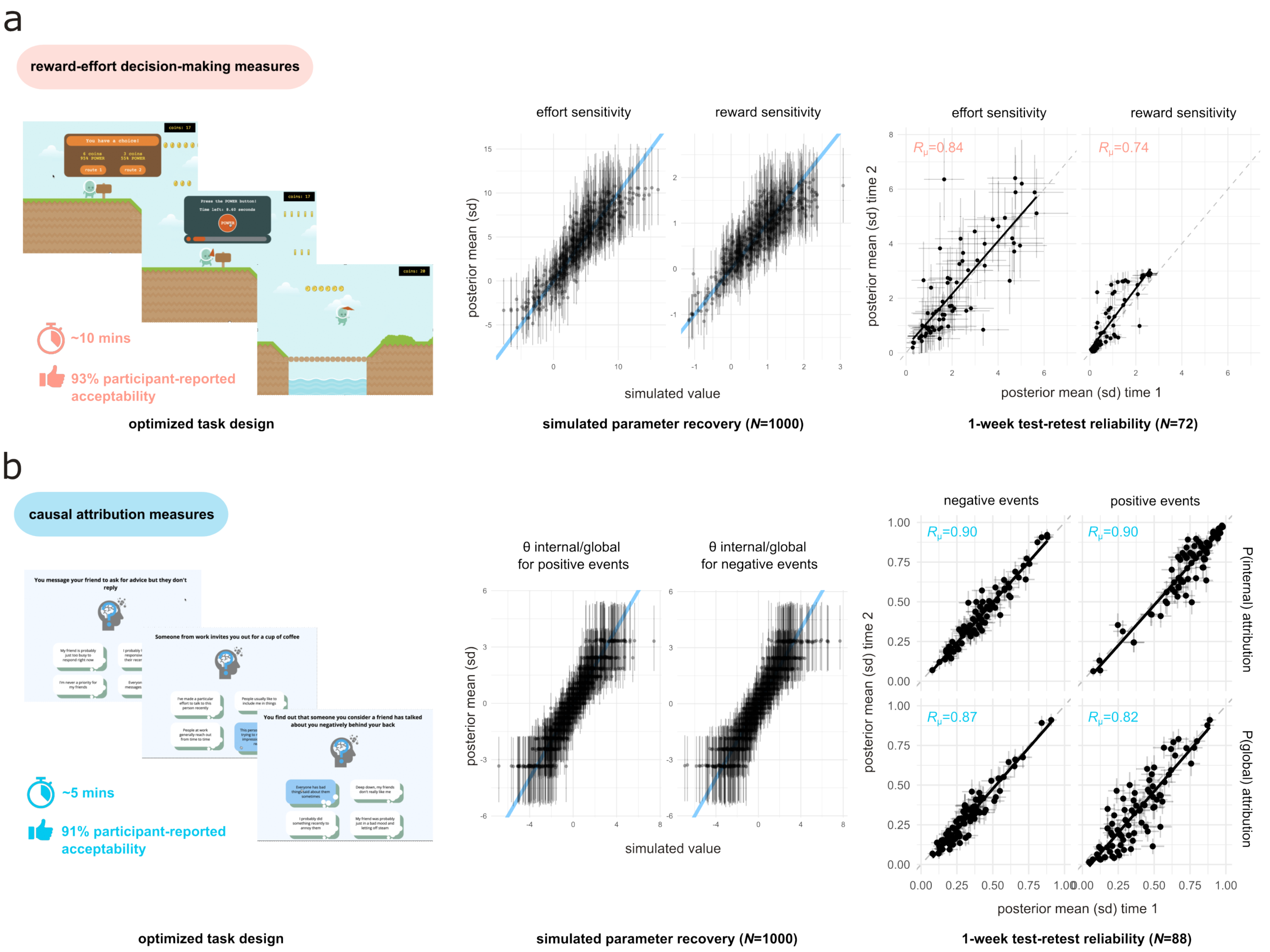


Figure 2: **Optimized task designs for measuring cognitive mechanisms relevant to goal-setting and cognitive restructuring components of behavioural activation and cognitive therapies.** **a** The gamified reward-effort decision-making task, which measures sensitivity to required effort and potential reward when deciding between different effortful actions (play a demo version here). **b** The causal attribution task, which measures latent tendency to attribute positive and negative everyday events to internal (*vs* external) and global (*vs* specific) causes (play a demo version here). *Left*, screenshots of the final task versions, alongside average task completion times and % user-acceptability ratings. *Centre*, *right*, psychometric properties of derived cognitive measures (independent parameter recovery during simulation-based calibration analysis and observed test-retest reliability).θ, parameters describing latent tendency to endorse internal/global attributions for positive and negative events; R_μ, posterior mean estimates for observed test-retest reliability of each model parameter.

Testing effects of interventions based on different components of cognitive-behavioural therapy on their proposed cognitive mechanisms

To test whether interventions derived from different components of behavioural activation and cognitive restructuring therapies impact their proposed cognitive mechanisms, we next conducted a set of studies in which participants completed the relevant task-based assessment twice, with 1:1 random assignment to either the active intervention or a well-matched control intervention in between (i.e., a mixed within/between-subjects design; Figure 1e). In all cases, initial discovery experiments were followed up with replication tests, to assess the reliability of findings.

For all studies, participants were recruited from an online research participation platform (Prolific), were required to be based in the UK, 18-65 years old, and fluent in English. At the end of the study, participants completed demographic and psychological symptom self-report measures (see Methods). Samples showed some evidence of self-selection for interest in mental health research, given on average 45% of participants reported previous treatment for a mental health problem, and moderate endorsement of current depression and social anxiety symptoms (Table S1, Figure S3). Samples were relatively well-balanced in terms of age, gender, and neurodiversity, but were predominantly White (Table S1).

Effects of a goal-setting intervention derived from behavioural activation therapy on reward-effort decision-making

Role of goal-setting in behavioural activation therapy. The use of activity-scheduling and goal-setting exercises is a core element of behavioural activation therapy for low mood (Martell et al., 2013). Acting according to a plan, rather than relying on internal state or mood, is thought to increase the likelihood of both acting and subsequently experiencing natural rewards, resulting in a positive reinforcement loop that serves to promote further activity and reward experience (Quigley and Dobson, 2017). In theory, acting according to a predetermined plan could boost activity levels either by making potential rewards more salient (*increasing reward sensitivity*), or by lowering the perceived level of effort required (*decreasing effort sensitivity*), when deciding if a particular action is worth taking (Reiter et al., 2021; Huys et al., 2022) (Figure 1c).

Investigating the effects of goal-setting on reward-effort decision-making. Here, we made use of the fact that reward-effort decision-making has been well-studied in cognitive neuroscience (e.g., Treadway et al. 2009; Bonnelle et al. 2015; Berwian et al. 2020). Starting from a previously-validated task design (Berwian et al., 2020), in conjunction with the recent introduction of online game engines into behavioural neuroscience research (Wise and Dolan, 2020), we developed a gamified task that was short, acceptable to users, and could reliably identify reward and effort sensitivity parameters from choice data (Figure 2a, Methods). Briefly, on each trial participants were asked to choose between two options, which varied both in terms of required physical effort (fast presses on a mouse or touchscreen) and offered reward amount (number of game coins, which were converted into a cash bonus at the end of the study). Choices were always non-dominated (the higher reward option required greater effort), except for two 'catch' trials used as internal attention checks (see Methods). After choosing an option, participants had to exert the required effort within a time-limit (10s) to gain the reward. After each block of trials (four per task), participants were asked to rate their sense of achievement on successful effort exertion, sense of pleasure in gaining rewards, and boredom levels, using an interactive slider.

The goal-setting intervention consisted of text describing the importance of setting realistic (achievable) goals, followed by a short comprehension quiz (see Methods). When completing the game for the second time, participants in the goal-setting condition were asked to set a goal (number of coins they would like to earn, out of the maximum possible available) prior to completing each block. Within each block, progress towards their goal was then tracked visually across trials. The control intervention consisted of matched-length information about different kinds of computer games, also followed by a quiz (for intervention reading times and quiz results, see Figure S5c,d). At the start of each block, control participants were asked to rate how much they enjoyed different kinds of games, but the task remained otherwise identical.

Goal-setting decreases effort sensitivity during reward-effort decision-making. In both initial discovery (*N*=100) and replication samples (*N*=102), linear mixed-effects modelling of individual trial data revealed a significant interaction between intervention condition and timepoint (pre *vs* post intervention) on proportionate choice of higher-effort higher-reward options ($F_{1,8697} = 14.5, F_{1,8871} = 34.8; p < 0.001$) – with greater choice of higher-effort options at time 2 in the goal-setting group (Supplementary Results, Figure S4). Analysis via our pre-specified Hierarchical Bayesian analysis model (as developed during the task design-optimization process) revealed that, in both samples, this was due to a specific decrease in effort sensitivity at time 2 for individuals who completed the goal-setting intervention(posterior means for group-level effect of the goal-setting intervention on effort sensitivity at time 2=-0.57 [90% posterior Credible Interval (CI) -0.91,-0.23], -0.32 [90% posterior CI -0.65,-0.004]; Figure 3, Table S2). Mean posterior predictive accuracy of the model for each sample was 0.81 [SD 0.16] and 0.83 [SD 0.15], and pseudo-r^2 values (reflecting relative proportion of variance in choice behaviour explained, compared to a chance model) were 0.51, 0.49, respectively.

Goal-setting changes subjective evaluation of effort expenditure and reward receipt. In line with the theory that goal-setting leads to a decrease in effort sensitivity when deciding to act, in turn leading to greater experience of reward, in both samples participants in the goal-setting condition reported greater sense of achievement on successful effort exertion ($F_{1,97} = 21.0, F_{1,100} = 23.7$), greater pleasure on gaining rewards ($F_{1,97} = 20.4, F_{1,100} = 12.7$), and lower boredom levels, during the second game ($F_{1,97} = 33.8, F_{1,100} = 4.2$; all $p < 0.05$; Supplementary Results, Figure S4).

Emphasizing the importance of setting achievable goals leads to increased effort expenditure over time. Consistent with the importance behavioural activation therapy places on both setting achievable goals and gradual increasing effort expenditure over time (Martell et al., 2013), we found that participants in the goal-setting condition tended to both exceed their goals within each block, and increase the ambitiousness of their goals across task blocks ($F_{2.4,236} = 19, F_{2.5,245} = 8.9; p < 0.001$; Supplementary Results, Figure S5).

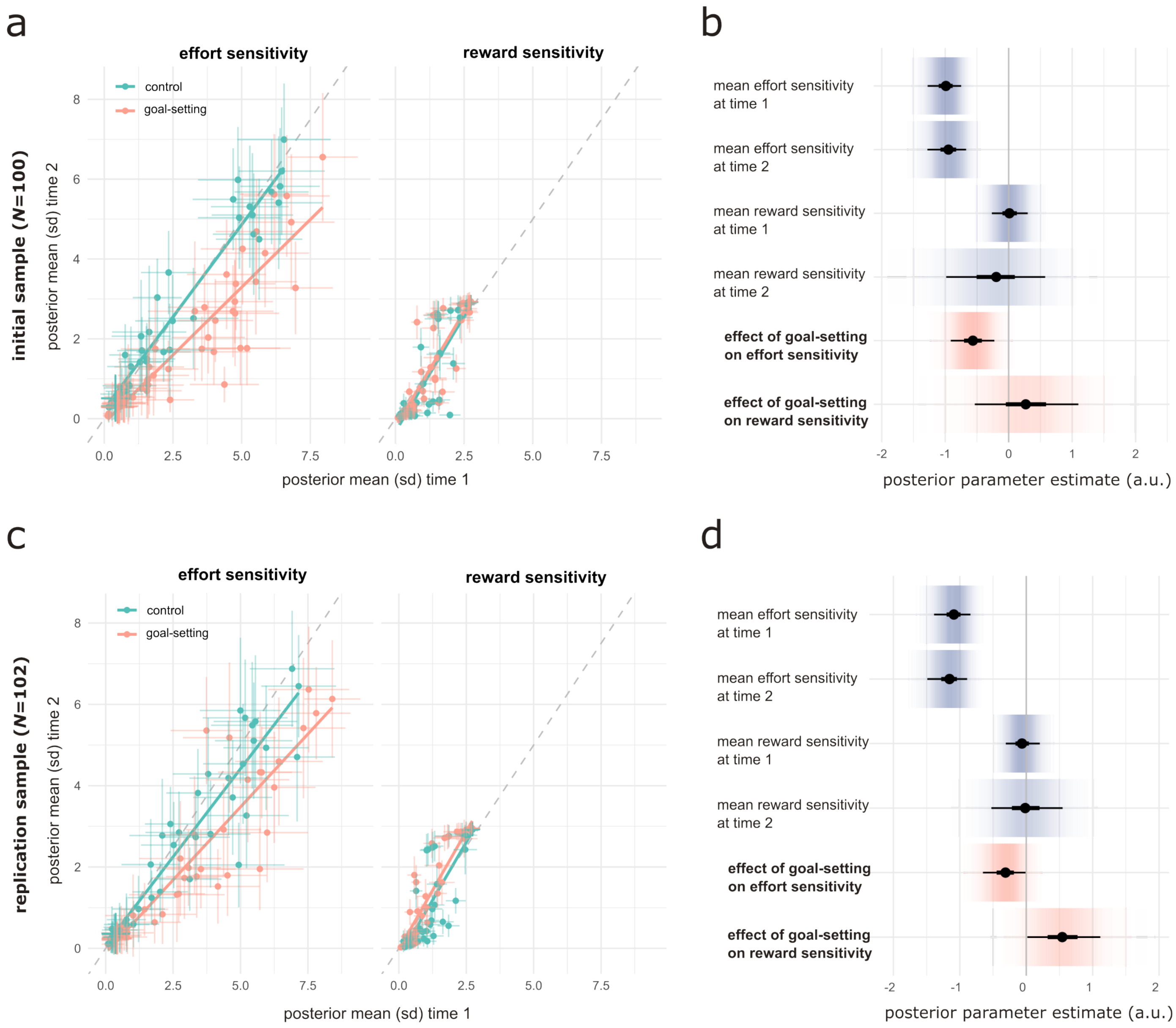


Figure 3: **A goal-setting intervention based on principles of behavioural activation therapy resulted in a selective decrease in effort sensitivity during reward-effort decision-making, compared to a well-matched control condition.** **a** Posterior mean (and SD) parameter estimates for each participant at time 1 (pre-intervention) vs time 1 (post-intervention), by intervention condition, in the initial discovery sample (goal-setting intervention, N=49; control intervention, N=51). Lines of best fit for posterior mean parameter estimates at time 1 vs time 2 for individuals in each intervention group are plotted for illustration purposes. **b** Posterior parameter estimates for group means (over all participants/intervention conditions) for each parameter at each time point, and the additional effects of goal-setting intervention in active group participants at time 2, in the initial discovery sample. Thick inner lines represent 50%, and thin outer lines represent 90% Credible Intervals, the point estimate is the mean, and shading represents posterior probability density, *a.u.*, arbitrary units. **c** The same plot as (a), for the independent replication sample (goal-setting, N=50; control, N=52). **d** The same plot as (b), for the independent replication sample.

Effects of a cognitive restructuring intervention on causal attribution of positive and negative events

Role of cognitive restructuring in cognitive therapy. A core idea underlying cognitive therapy is that it is often how we interpret things that happen to us, rather than the events themselves, that shapes how we end up feeling (Beck et al., 1987). In particular, learned helplessness theory suggests that, in some individuals, persistent low mood results from a heightened tendency to attribute negative events to causes which are *internal* (related to the self, compared the outside world), *global* (likely to be active in all situations, rather than this specific one alone), and *stable* (likely to persist in time, rather than change in the future) (Abramson et al., 1978). Therefore, a key focus of cognitive restructuring is training individuals to identify unhelpful attributions, and practising consideration of alternative and helpful explanations ('reappraisal') (Clark, 2022).

Whilst there is robust evidence of heightened attribution of positive events to internal and global causes in healthy individuals (an effect which has been interpreted as a self-serving or self-protective bias), overly internal and global attributions of negative events have been identified in currently depressed individuals, and predicts future depressed mood (Mezulis et al., 2004; Pearson et al., 2015). However, it is not clear 1) the extent to which addressing these different dimensions (internality, globality) is important in cognitive restructuring, and 2) the extent to which improvements in mood relate to a decreased tendency to make 'depressogenic' attributions (internal/global attributions of negative events), *versus* increased use of self-protective or compensatory strategies (internal/global attributions of positive events) (Barber and DeRubeis, 1989; Huibers et al., 2021) (Figure 1c).

Investigating the effects of cognitive restructuring on causal attribution. Here, we present data from a novel hybrid self-report/task measure ('causal attribution task'), developed from an analysis of previous scenario-based attribution tasks, item-response theory-based optimization, and consideration of sensitivity to potential sociodemographic moderators (age, gender, functional disability/neurodivergence, and minoritized group status; Methods, Figure S2). Briefly, participants were presented a series of brief descriptions of events, and asked to choose which of four listed causal explanations they thought the most likely, if such an event had happened to them. Half the events were positive and half negative, and the four potential explanations varied orthogonally in terms of describing internal (*vs* external) and global (*vs* specific) causes. Extensive pilot testing revealed that data from two alternative task versions could be used to reliably identify parameters governing probability of endorsement of an internal (*vs* external) and global (*vs* specific) causes, separately for positively and negatively valenced events (Methods, Figure 2b).

The cognitive restructuring intervention consisted of information about a cognitive model of mood (link between interpretations of events and feelings), interactive exercises identifying helpful and unhelpful attributions of the same events, inviting people to practise generating alternative explanations for recent events in their own lives, and a summary comprehension quiz (Methods). The control intervention was based on materials from emotion-focused therapy (Greenberg, 2015), and was closely matched in terms of length, interactivity, and self-relevant exercise content – although, importantly, it did not contain reference to cognitive interpretations influencing feelings or include reappraisal activities (e.g., reflection on whether a particular emotional reaction is helpful or not; for intervention completion times see Figure S6c,d). According to

Cognitive restructuring decreases tendency to attribute negative events to internal (self-related) causes. In both initial discovery (N=100) and replication samples (N=100), linear mixed-effects modelling of individual trial data revealed a significant interaction between intervention condition and time-point (pre *vs* post intervention) on proportionate choice of internal attributions for negative events ($F_{1,6294} = 10.9, F_{1,6294} = 5.0$; both $p < 0.03$) – with lower choice of internal attributions for negative events at time 2 in the cognitive restructuring group (Supplementary Results, Figure S6). Analysis via our pre-specified Hierarchical Bayesian model revealed that, in both samples, this was due to a decrease in the model parameter describing the latent tendency of individuals to internally-attribute negative events following the restructuring intervention (posterior means for group-level effect of the cognitive restructuring intervention at time 2=-0.56 [90% posterior CI -0.87,-0.24], -0.34 [90% posterior CI -0.61,-0.05]; Figure 4, Table S3). Mean posterior predictive accuracy of the model for each sample was 0.74 [SD 0.11] and 0.73 [SD 0.11] for internal attributions, and 0.69 [SD 0.11] and 0.68 [SD 0.11] for global attributions. Pseudo-r^2 values were 0.64, 0.64, for internal attributions, and 0.59, 0.58 for global attributions, respectively.

Effects of interventions based on different components of cognitive-behavioural therapy on their proposed cognitive mechanisms: Interim summary

In two parallel sets of studies, we found 1) evidence that a goal-setting intervention based on principles of behavioural activation therapy reliably reduced sensitivity to required effort (but not reward) levels when choosing between different actions, and 2) that a restructuring intervention based on cognitive therapy reliably reduced a tendency to attribute negative events to self-related or internal causes (an aspect of attributional style thought to contribute to symptoms of low mood), but did not impact a tendency to make overly-general or global attributions.

However, it is not possible to tell on the basis of results so far whether the effects of each interventions were *specific* to the task administered in each study – or whether each intervention's effects might 'spill over' to other cognitive domains (Figure 1b).

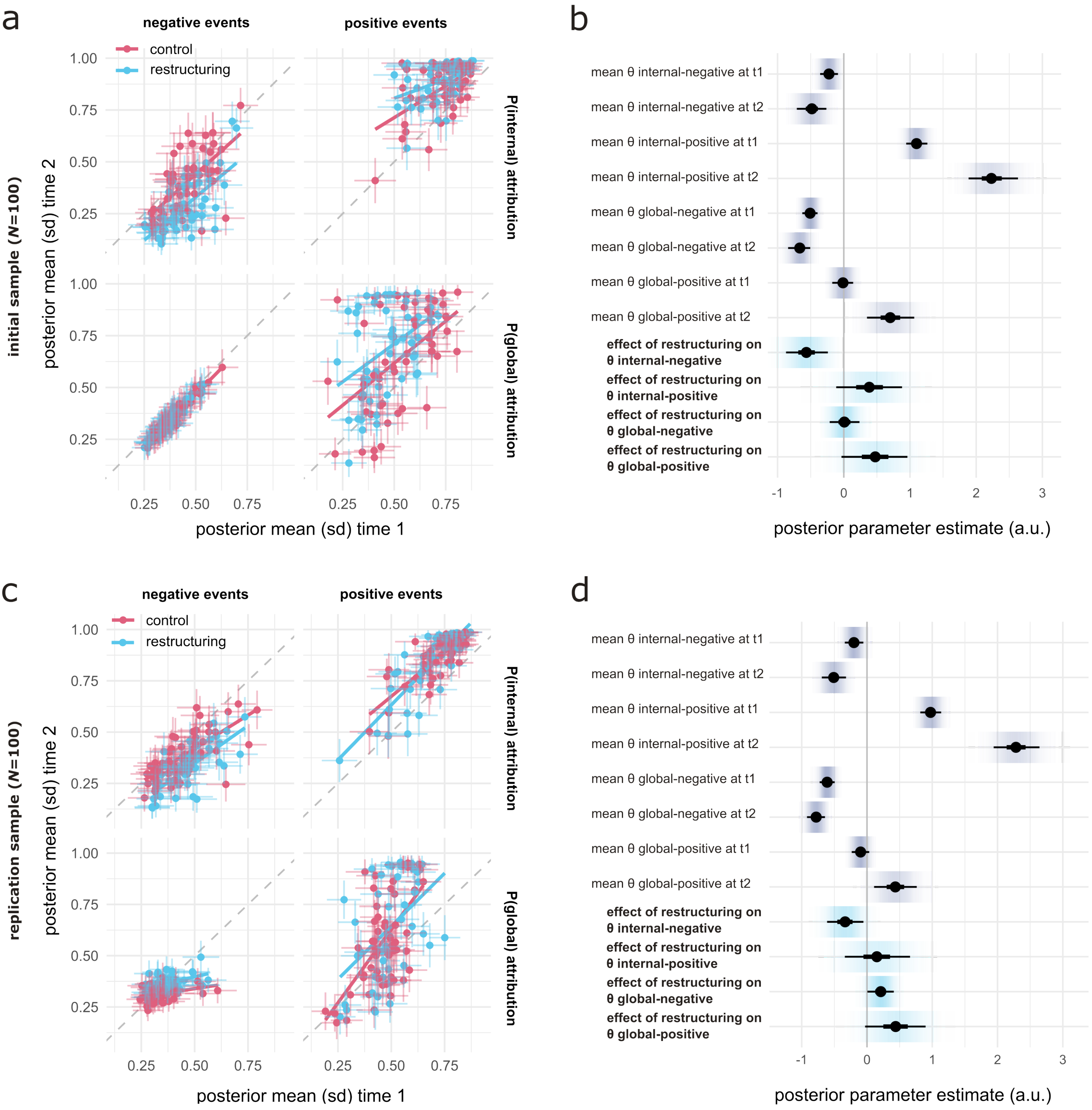


Figure 4: **A cognitive restructuring intervention based on cognitive therapy resulted in decreased internal attribution of negative events, compared to a well-matched control condition**. **a** Posterior mean (and SD) parameter estimates for each participant at time 1 (pre-intervention) and time 2 (post-intervention), by intervention group, in the initial discovery sample (cognitive restructuring intervention, N=49; control intervention, N=51). Lines of best fit for mean time 1 *vs* time 2 estimates for individuals in each group are plotted for illustration purposes. *Continued on the next page*

Figure 4: **b** Posterior parameter estimates for group means (over all participants/intervention conditions) for each parameter at each time point, and the additional effect of intervention in cognitive restructuring group participants at time 2, in the initial discovery sample. Thick inner lines represent 50%, and thin outer lines represent 90% Credible Intervals, the point estimate is the mean, and shading represents posterior probability density, *a.u.*, arbitrary units. **c** The same plot as (a), for the independent replication sample (cognitive restructuring, N=44; control, N=56). **d** The same plot as (b), for the independent replication sample. θ, parameters describing latent tendency to endorse internal/global attributions for positive and negative events.

Specificity of interventions to proposed cognitive mechanisms

To test whether effects of our interventions were specific to their proposed cognitive mechanisms, we next carried out a study using a 2 × 2 intervention × task crossover design (Figure 1f). Specifically, participants were separately randomized to task and intervention conditions, in order to investigate the effects of goal-setting *vs* cognitive-restructuring on reward-effort decision-making, and cognitive-restructuring *vs* goal-setting on causal attribution. Participants were recruited as previously, and are described in Table S1.

Goal-setting but not cognitive restructuring affects effort sensitivity during reward-effort decision-making. For crossover study participants who were randomized to complete the reward-effort decision-making task (N=197), Hierarchical Bayesian analysis revealed that goal-setting but not cognitive restructuring resulted in decreased effort sensitivity during reward-effort decision-making (posterior mean for group-level effect of goal-setting *vs* restructuring =-0.48 [90%CI -0.74,-0.22], Figure 5a,b, Table S4).

Cognitive restructuring but not goal-setting affects internal attribution of negative events. For crossover study participants who were randomized to complete the causal attribution task (N=208), Hierarchical Bayesian analysis revealed that the cognitive restructuring but not goal-setting intervention resulted in reduced internal attribution of negative events (posterior mean for group-level effect of restructuring *vs* goal-setting on negative events =-0.28 [90%CI -0.44,-0.12], Figure 5c,d, Table S4) Further, in this sample, cognitive restructuring was associated with increased internal attribution of positive events (posterior mean for group-level effect of restructuring *vs* goal-setting on positive events =0.46 [90%CI 0.22,0.70]).

Of note, under this analysis framework, effects common to both intervention conditions would be expressed as changes in group-level parameter means between time 1 and time 2 - however posterior distributions (90% Credible Intervals) for group means were overlapping for all parameters across time-points (Figure 5b,d). Therefore, data from this study provided not only a further replication of the effects found in the first set of studies, but showed that the effects of each intervention appeared specific to their relevant theoretically-informed task and parameter measures.

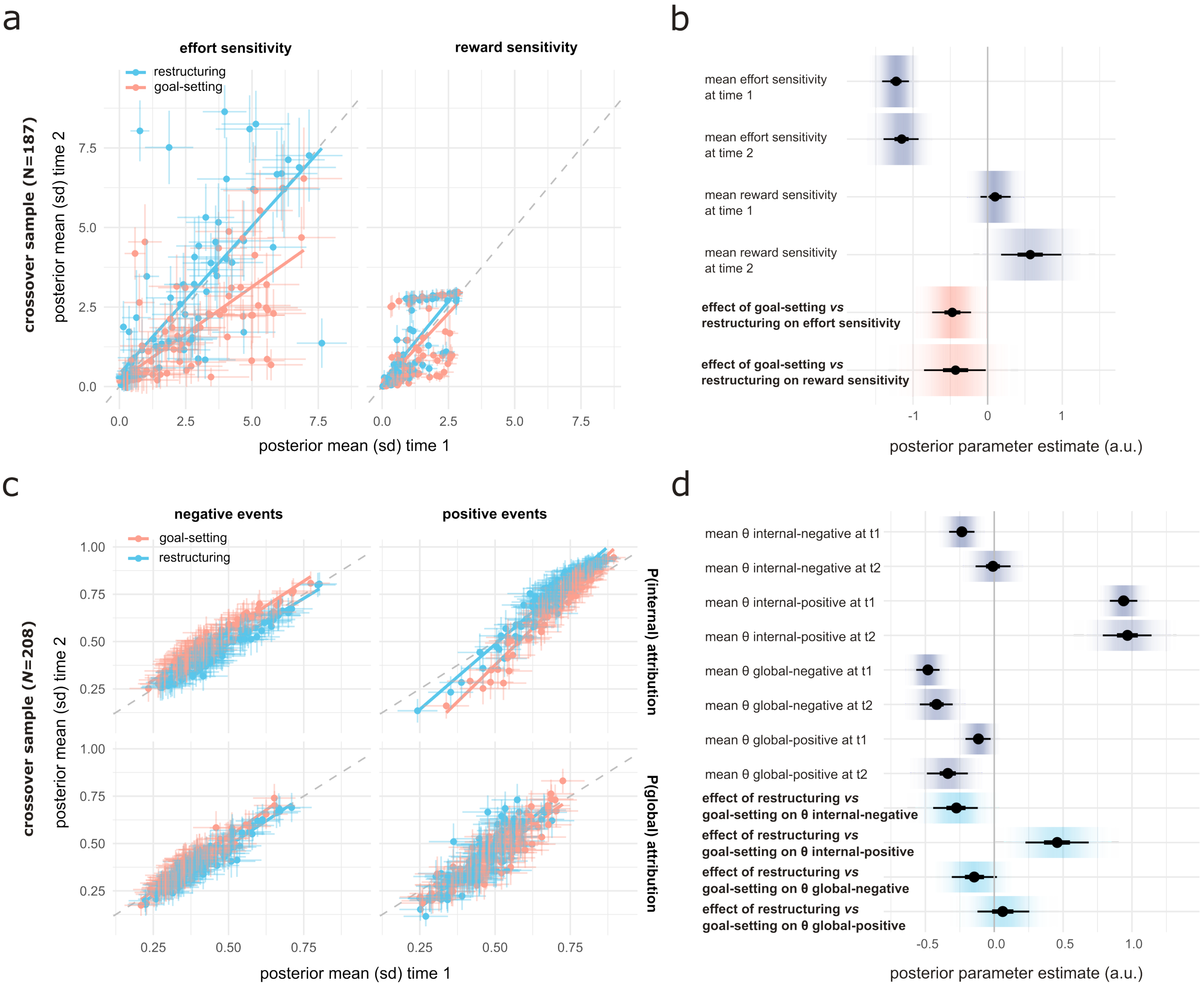


Figure 5: **In a crossover design, effects of goal-setting and cognitive restructuring interventions were found to be specific to their relevant cognitive mechanisms.** **a** Posterior mean (and SD) parameter estimates for each participant at time 1 (pre-intervention) and time 2 (post-intervention), by intervention group, for the crossover study participants randomized to the reward-effort decision-making task (goal-setting, N=99; cognitive restructuring, N=88). **b** Posterior parameter estimates for group means (over all participants) for each parameter at each time point, and the additional effect of the goal-setting intervention at time 2, in the crossover study participants who completed the reward-effort decision-making task. Compared to restructuring, goal-setting reduced effort sensitivity. Thick inner lines represent 50%, and thin outer lines represent 90% credible intervals, the point estimate is the mean, and shading represents posterior probability density, *a.u.*, arbitrary units. **c** The same plot as (a), for the crossover study participants who were randomized to the causal attribution task (cognitive restructuring, N=106; goal-setting, N=102). **d** The same plot as (b), for the additional effect of the cognitive restructuring intervention at time 2, in the crossover study participants who completed the causal attribution task. Compared to goal-setting, restructuring reduced internal attribution of negative events, and increased internal attribution of positive events.

Relating magnitude of intervention effects to individual symptom profiles

Finally, we conducted an exploratory analysis to determine if individual differences in psychological symptom profiles might moderate the effects of interventions on our cognitive measures. To increase power, initial discovery and replication samples from the sets of studies described above were first combined for each task. We then sought to determine if any effects in these combined samples were replicated in the crossover study data (where comparison interventions were less well-matched in terms of e.g., length, interactivity).

Heterogeneity of treatment effects analysis. Across tasks and measures, we found evidence of moderate response variation in terms of change in mean effort sensitivity following the goal-setting intervention (point estimate for SD of individual responses=0.43 [95%CI 0.32,0.55]), and mean tendency to attribute positive events to internal causes following the cognitive restructuring intervention (point estimate for SD of individual responses=0.40 [95%CI 0.04,0.76]).

Joint modelling of task and self-report data. In order to test if symptom profiles were related to magnitude of either of these responses, individual symptom data were combined into the previously described behavioural task analysis models. Following Haines (2021), within the joint model, individual item self-report data were analysed using item response theory (IRT). Specifically, we hypothesized the existence of two latent traits in the symptom data, labelled '*behavioural amotivation*' (symptoms of anhedonia and behavioural apathy: constructed from Apathy Motivation Index behavioural amotivation items and PHQ9 items indexing anhedonia and lethargic symptoms) and '*negative cognition*' (negative self-beliefs associated with depressed mood: constructed from Dysfunctional Attitude Scale items and PHQ9 items indexing feelings of hopelessness and failure; see Methods). These traits were chosen on the basis of proposals that behavioural treatments might be more effective for clinical presentations dominated by the former, and cognitive treatments for the latter (e.g., Beck et al. 1987; Forbes 2020). Figure 6a,b shows that the pattern of item contributions to each latent trait estimate was relatively stable across samples (for the highest discriminability items in each sample see Supplementary Results).

Individual differences in the effect of goal-setting on reward-effort decision-making. In the combined goal-setting *vs* control intervention sample (*N*=195), higher amotivation estimates were associated with both greater effort sensitivity at baseline (posterior parameter estimate for group-level β weight of trait amotivation estimates on time 1 effort sensitivity estimates, β_{BASE}=0.24 [90%CI 0.06,0.44]), and greater magnitude of response to the goal-setting intervention (posterior estimate for β weight of amotivation on group-level active intervention effect, β_{INT}=-0.37 [90%CI -0.60,-0.15], Figure 6c, Table S5). The direction (but not magnitude) of these effects was replicated in the less well-controlled crossover sample, where amotivation and negative cognition estimates could be included in the same model (*N*=185, β_{BASE}=0.16 [90%CI -0.12,0.51], β_{INT}=-0.11 [90%CI -0.62,0.37], Figure 6d, Table S5). Evidence that individuals higher in amotivation differed in baseline effort sensitivity and showed greater response to the goal-setting intervention was therefore somewhat inconclusive.

Individual differences in the effect of cognitive restructuring on causal attribution. In the combined restructuring *vs* control intervention sample (*N*=200), higher negative cognition estimates were associated with a lower tendency to attribute positive events to internal causes at baseline (β_{BASE}=-0.16 [90%CI -0.29,-0.04]), but not magnitude of change in this measure following restructuring (β_{INT}=-0.24 [90%CI -0.55,0.07], Figure 6e, Table S5). In the crossover sample (N=205), there was borderline evidence that both higher amotivation symptoms and higher negative cognition were associated with lower tendency to internally attribute positive events at baseline (β_{BASE}=-0.18 [90%CI -0.43,0.01], -0.20 [90%CI -0.46,0.01]). There was again no relationship between negative cognition and change on this measure following the restructuring intervention, but there was evidence of a negative relationship with amotivation (β_{INT}=-0.45 [90%CI -0.97,-0.04], Figure 6f, Table S5). This suggests that whilst symptoms of both amotivation and negative cognition are associated with lower baseline self-protective attributional tendencies, only greater amotivation symptoms were associated with response on this measure to an intervention based on cognitive restructuring - with greater amotivation relating to smaller increases in internal-positive attribution tendency.

In summary, we found some evidence for higher amotivation symptoms relating to greater response to a goal-setting intervention based on behavioural activation therapy, but a smaller response to a restructuring intervention based on cognitive therapy, in terms of change in underlying cognitive mechanisms. However, we caution that these results are very preliminary and will require replication in the future work.

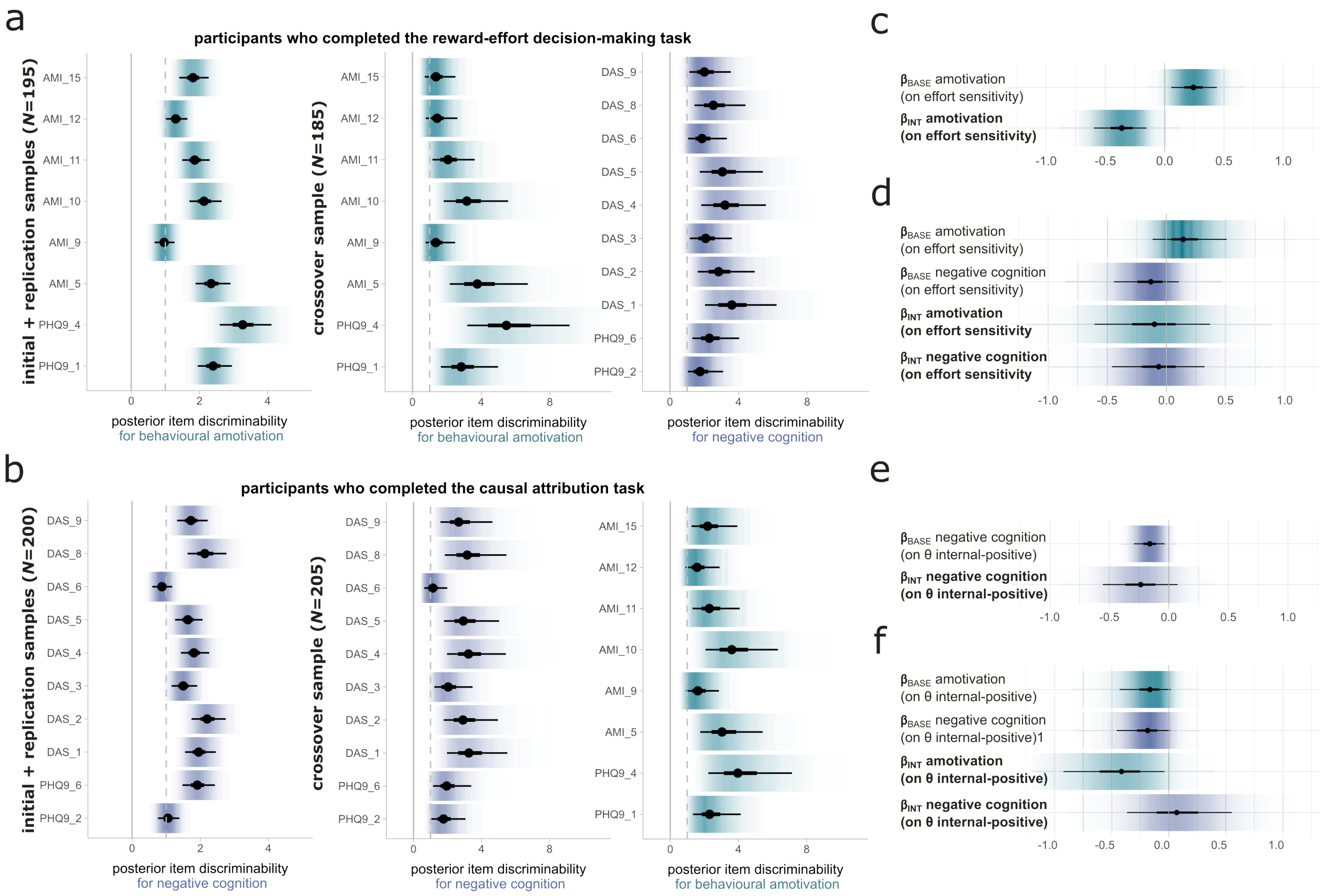


Figure 6: **Relationships between psychological symptoms and magnitude of intervention effects, in joint models of behavioural and self-report data**. *Continued on the next page.*

Figure 6: **a** *Left*, posterior item discriminability estimates (IRT model parameters describing how well an item differentiates between individuals scoring high and low on a latent trait) for behavioural amotivation in the combined reward-effort decision-making samples. Top-discriminating items for behavioural amotivation included "little interest or pleasure in doing things", "feeling tired or having little energy", and "I don't like to laze around" (reverse-scored). *Centre, right*; posterior discriminability estimates for amotivation and negative cognition in crossover study participants who completed the reward-effort decision-making task. Top discriminating items for negative cognition included "If other people know what you are really like, they will think less of you", and "If I don't set the highest standards for myself, I am likely to end up a second-rate person". The dotted lines represents a threshold of posterior discriminability > 1, which can be understood as representing a meaningful contribution to latent trait estimates. *AMI_x*, Apathy Motivation Index behavioural subscale items; *PHQ_1*, "little interest or pleasure in doing things"; *PHQ9_4*, "feeling tired or having little energy". *DAS_x*, Dysfunctional Attitude Scale (short form) items; *PHQ_2*, "feeling down, depressed, or hopeless"; *PHQ9_6*, "feeling bad about yourself or that you are a failure". **b** The same plot as (a), for trait negative cognition in the combined causal-attribution task samples (*left*), and negative cognition and behavioural amotivation in crossover study participants who completed the causal-attribution task (*centre; right*). **c** Posterior estimates from the joint self-report-task model in the combined reward-effort decision-making sample for the influence of behavioural amotivation on effort sensitivity at baseline (β_{BASE} amotivation), and on the effect of the goal-setting intervention on effort sensitivity at time 2 (β_{INT} amotivation) (arbitrary units). **d** The same plot at (b), for the crossover study participants who completed the reward-effort decision-making task, including the influence of negative cognition on baseline and intervention-induced changes in effort sensitivity (β_{BASE} negative cognition, β_{INT} negative cognition). **e** The same plot as (c), for influence of negative cognition on baseline and intervention-induced changes on internal attribution of positive events in the combined causal-attribution task sample. **f** The same plot as (d), for the influence of behavioural amotivation and negative cognition on baseline and intervention-induced changes in internal-positive attributions in crossover study participants who completed the causal-attribution task. In all panels, thick inner lines represent 50%, and thin outer lines represent 90% credible intervals, the point estimate is the mean, and shading represents posterior probability density.

General Discussion

Over the last half century, there have been many calls for research into mechanisms by which existing effective psychotherapy treatments work (Kazdin, 2009; Huibers et al., 2021). Large individual patient data meta-analyses have provided some hints of differences in effects between treatments, and in different groups of individuals given the similar treatments (Cuijpers et al., 2019, 2022; Furukawa et al., 2021). Recent analyses of large-scale intensively-sampled mood data has also shown that symptom clusters representing anhedonia/lethargic symptoms and depressed mood/feelings of worthlessness exhibit different dynamic properties within and between individuals – which may represent different opportunities for intervention (Ebrahimi et al., 2021). However, it remains largely unclear which individuals are more likely to benefit from different kinds of treatment – in particular, cognitive *vs* behavioural therapies – and this is an active area of ongoing research (Craske, 2022; Driessen et al., 2022). Further, a key issue in psychotherapy process research is distinguishing causal relationships from correlates of treatment response (Lorenzo-Luaces et al., 2015; Eronen, 2020). This is critical, as only the former are likely to support the longer-term goal of truly effective treatment personalization. Here, we show that using well-validated cognitive measures, in conjunction with experimental designs capable of supporting causal inference, we can test directly whether different proposed mechanisms are impacted by interventions derived from distinct components of psychological therapies.

We found that a goal-setting intervention, that included education about the importance of setting achievable goals and salient visual tracking of progress towards goals, reliably led to increased selection of higher-effort/higher-reward actions. Model-based analysis revealed that this was due a selective reduction in sensitivity to required effort levels (but not sensitivity to potential rewards), when deciding how to act (Figure 4). Significantly, this change in decision-making was was accompanied by an increased sense of achievement for actions and experienced pleasure for rewards – suggesting not only that goal-setting decreased subjective weighting of effort but that the resulting energizing of overall action levels may be sufficient to kick-start a positive reinforcement cycle through which behavioural activation therapy is thought to improve mood (Quigley and Dobson, 2017; Huys et al., 2022). This implies that a focus on setting achievable rewards (which are gradually increased over time), and active monitoring of completion of activities (e.g., via monitoring forms) may be key active ingredients of behavioural activation therapy. It is not clear from our current results which particular features of our intervention (education about achievable goals, pre-commitment to a specific target, and monitoring of progress towards this target) were most potent in effecting this change, but this can be further dissected in future work. It will also be important to test if the effects identified here generalize from in-game actions and rewards to the kinds of everyday effortful activities and rewards employed in a therapeutic context (see below).

We also found that a restructuring intervention, that included education about a cognitive model of low mood (”*thoughts affect feelings*”) and reappraisal practice, reliably reduced a tendency to attribute negative events to internal (self-related) causes, whilst not robustly affecting a tendency to assign events to overly-general or global causes (Figure 5). Both heightened internal and global attributional styles are implicated in depressed mood (Mezulis et al., 2004; Pearson et al., 2015), and indeed we observed associations between both these tendencies and depression symptoms and negative self-beliefs in our samples (Figure 6, Figure S7). We note that, in general, participants found the internal-external dimension of choice options easier to parse than the global-specific dimension, which may explain the lack of robust effects on this measure. It is not currently clear whether this is a limitation of our task materials or reflects a more general difficulty in understanding this aspect of attributional style, something that can be usefully explored in future work (e.g., Mason et al. 2023). Further, it is an open question whether expression of these kinds of belief is a cause or consequence of low mood (Cristea et al., 2015; Ezawa and Hollon, 2023). Here, we provide initial evidence that cognitive restructuring directly impacts attributional choice in a realistic scenario-based task that is robust to sociodemographic differences. In the future, this kind of measure may enable more precise and reliable tracking of changes in causal attribution over the course of treatment, and determination of whether or not this predates symptom change.

A critical aspect of our results is our demonstration that changes in theoretically-derived cognitive measures were specific to relevant interventions. This is a vital step towards an eventual goal of providing more targeted or personalized psychotherapy treatment, as if different cognitive processes are affected by multiple treatment components to the same extent, then it would render it hard to leverage differential administration or dosage of components to address relative deficits (or capitalize on relative strengths) on the basis of measurements of these processes (Eronen, 2020). Finally, we presented preliminary evidence that symptoms of behavioural amotivation (anhedonia and lethargy) may relate to greater responses to goal-setting, and lesser responses to cognitive restructuring. This accords with theoretical notions that behavioural treatments may be preferable for clinical presentations dominated by this kind of symptom profile (Beck et al., 1987; Forbes, 2020) – although these findings should be interpreted with caution as they did not replicate fully across samples.

The major limitation of these initial proof-of-concept results is that they concern the effects of custom interventions based on components of psychological therapies, as opposed to modules of real, proven to be effective, cognitive and behavioural treatments. Extension of our findings to this context is therefore a critical next step in constructing a chain of evidence that unpacks the mechanisms by which real-world therapies work. Such a translation would enable us to complete a vital link that relates change in cognitive mechanisms to parallel change in psychological symptoms following treatment completion. It remains possible there are too many differences between our toy interventions and actual psychotherapies (even highly controlled digitally-delivered content) for our results to hold. However, we believe that initial evidence of replicable effects of therapy-derived interventions on theory-based mechanisms, and, in particular, evidence of specific effects of these interventions, represents a foundational step prior to embedding such tests in resource-intensive contexts, such as clinical trials or treatment programs (Paulus et al., 2016).

An important lesson learned over the course of these studies is that the development of 'good' measures of cognitive processes fundamentally involves the management of various competing trade-off factors (Zorowitz and Niv, 2023). Specifically, increasing user engagement via gamification strategies (e.g., our reward-effort task) may involve a trade-off between noisiness of data, and face or construct validity. Conversely, measures with increased face validity (e.g., our scenario-based causal attribution measure) may involve a different degree of insight than more behavioural tasks, where individual differences in interpretation or understanding of the state-space may be a source of noise. Optimal points for these trade-offs may be hard to judge on the basis of isolated quantitative measurements (such as test-retest reliability), and better understood in the context of qualitative input from future end-users (Bear et al., 2022).

In conclusion, digital therapies can help reduce the treatment gap in mental health service provision (Thornicroft et al., 2017; Torous et al., 2021), in particular for underserved populations (Schueller et al., 2019). However, increasing user engagement is likely to be key for greater uptake of digital therapeutics (Graham et al., 2019; Borghouts et al., 2021). Promising targets for increasing engagement with such services include increasing value to end users (e.g., providing knowledge back), and evidence of personalization of content (Szinay et al., 2020). We argue that greater knowledge about the mechanisms via which established psychological treatments work is an important step towards achieving these goals (Craske et al., 2023).

Methods

Data and materials availability statement

Code for implementing both tasks and analyses described here is available at the study github repository. Anonymized data will be deposited at the same location upon acceptance of this manuscript for peer-reviewed publication.

Ethical approval

All participants gave written informed consent and all studies were approved by the UCL Research Ethics Committee (project ID 21029/001).

General methods

All analyses were carried out in `R` version 4.1.2 (The R Foundation for Statistical Computing, 2021), using `RStudio` version 2022.02.0 (RStudio, PBC, 2022).

Hierarchical Bayesian modelling

Model evaluation and fit procedures were carried out according to Bayesian workflow recommendations (Gelman et al., 2020; Schad et al., 2021), with results of Bayesian analyses reported in accordance with recent guidelines (Kruschke, 2021). Model parameters were estimated using Markov-Chain Monte Carlo (MCMC) sampling as implemented in Stan 2.21.0 (Carpenter et al., 2017), using `RStan` 2.21.3 (Stan Development Team, 2021). MCMC chains were initiated with random starting values, and posterior distributions were formed using 4 chains of 2000 iterations, with 1000 discarded warm-up samples (i.e., 4000 kept iterations per model). Convergence of sampling chains was assessed via inspection of trace plots and Gelman-Rubin ($\hat{R}$) statistics for each parameter (Gelman and Rubin, 1992). Assessment for sampling difficulties and parameter collinearity was via inspection of bivariate marginal posterior distributions between pairs of parameters. All models used generic weakly-informative priors (see Supplementary Methods).

At the initial task/measure development stage, different models of the same data were compared via leave-one-out cross-validation, using the R package `loo` (Vehtari et al., 2017). Given our concern with optimizing for task brevity (whilst considering multiple potential mechanisms of behavioural change; see main text), priority was accorded to simple models with few parameters – with a model comparison metric based on out-of-sample prediction guarding against overfitting. For the main analyses reported here, two model-agnostic 'goodness-of-fit' measures are reported. Posterior predictive accuracy was calculated as the match between replicated choice data generated stochastically from posterior parameter estimates and task trial arrays, and the observed data from each participant (means and SDs across participants are reported). Pseudo-r^2 statistics, which reflect the amount of variance explained by the model relative to a model of pure chance, were calculated as $1-L/C$, where L is the summed log likelihood over participants, and C is the chance likelihood of observing responses (for two choice options, $log(0.5)t$) (Daw, 2011).

For experimental effects of interest (e.g., the group-level effect of receiving the active vs control intervention on parameter estimates at time 2), parameters were assessed using 90% credible intervals (CIs), with a 90% CI excluding zero interpreted as representing evidence for a meaningful contribution to posterior parameter estimates. Although this choice of threshold is somewhat arbitrary, it follows conventions in the literature, and recommendations of use of a <95% CI for sample sizes less than 10,000 (McElreath, 2016). Distributions of posterior parameter estimates and CIs were visualized using the R packages `bayesplot` (Gabry et al., 2019) and `tidybayes` (Kay, 2022).

Simulation-based calibration analysis

Simulation-based calibration (SBC) analysis was used to validate our modelling and inference procedures for both tasks and sets of measures (Talts et al., 2020). Briefly, this involves generating draws from the prior predictive distribution of a generative model (creating N simulated datasets), then fitting the model to each simulated dataset and obtaining D independent draws from the model posterior. For each parameter of interest, the rank of the simulated value within the posterior draws is then calculated. If the data generation and inference procedure works as expected, then the resulting ranks should be uniformly distributed across $[0, D]$ (Modrák et al., 2022). Here, we generated N=1000 datasets based on independent draws from the prior distributions of each parameter, which were specified generously based on the empirical posterior estimates of parameter distributions observed in pilot data. We then took D=2000 posterior draws (after discarding 1000 warm-up samples), across two sampling chains. Graphical summaries of SBC results were generated using the R package `SBC` (Kim et al., 2023).

Test-retest reliability analysis

Recent discussions highlight adequate test-retest reliability as a prerequisite for detection of true individual differences in a measure (Hedge et al., 2018; Haines et al., 2020b; Brown et al., 2020; Zorowitz and Niv, 2023). Here, we estimated test-retest reliability using the approach described in (Rouder and Haaf, 2019; Haines et al., 2020b). Specifically, data from two time points (repeat test administration in the same sample of participants) were fit using a single hierarchical model, with separate group means for each parameter at each time point, and individual parameter estimates at each time point assumed to be drawn from a multivariate normal distribution, and a uniform prior over $[-1, 1]$ on correlation of individual values across time-points (see Equation 3, Equation 9). Posterior R values for correlation of individual parameter estimates across time-points are then reported as an estimate of test-retest reliability, that sufficiently takes into account both relatedness of different measurements and measurement error (precision) of individual estimates.

Self-reported demographic and clinical information

At the end of each study, participants completed a set of brief self-report measures to provide information about their recent experience of mental health symptoms, and other relevant sociodemographic information. Symptoms of low mood were measured using The 9-item Patient Health Questionnaire (PHQ9) (Kroenke et al., 2001). We also included the 3-item Social Phobia Inventory (miniSPIN), a brief measure of social anxiety symptoms (Connor et al., 2001), given our previous observations that social anxiety is relatively elevated in Prolific samples. The Apathy Motivation Index (AMI), which measures apathy and amotivation across behavioural, social, and emotional domains (Ang et al., 2017), was included for reward-effort decision-making samples, given the hypothesis that behavioural activation therapy may be particularly effective for individuals with disrupted reward or effort processing (Forbes, 2020; Reiter et al., 2021). The Dysfunctional Attitudes Scale (short form) (DAS), a measure of negative self-beliefs observed in some depressed people (Beevers et al., 2007), was included for causal attribution task samples, as it has previously been shown to be sensitive to cognitive treatment of low mood (Cristea et al., 2015).

The demographic measure included questions about participant gender identity, age, neurodivergence (defined as “a term for when someone processes or learns information in a different way to that which is considered ’typical’: common examples include autism and ADHD”), previous treatment for a mental health problem, disability across World Health Organization Disability Assessment 2.0 domains of functioning (World Health Organization, 2012), and financial, housing, and employment status (given these factors have previously been shown to relate to treatment outcomes for depression; Buckman et al. (2022). All self-report batteries included two infrequency items (in which some responses are logically invalid or highly improbable), in order to detect potential inattentive responding (Zorowitz et al., 2021). Participants were required to provide correct responses to both items in order to be included in analyses including self-report data.

Reward-effort decision-making studies

Reward-effort decision-making task

Code for implementing the version of the task described here and a link to a demo version of the game is available here. The task was coded in javascript using `phaser` 3.23.0, a framework for creating HTML5 games for desktop and mobile devices (https://phaser.io/; Photon Storm, 2020).

Participants were informed that they were travelling through a strange land, covered in rivers and streams. At regular points along their journey, they would be required to power up their magic umbrella, in order to fly across the water. At each crossing point, they could choose between different routes. Different routes would allow them to collect different numbers of coins (with total coins converted into a cash bonus at the end of the study), but required different amounts of effort to cross. For each route, they would have to press or click quickly an on-screen ‘power’ button, until they reached the required effort level to cross. Effort levels were presented as percentages of maximal power, which (unknown to participants) was individually calibrated at the start of the study during a series of practice trials, designed to elicit maximal possible effort levels (press rates) during the time limit (10s). In order to avoid ‘gaming’ of practice trials, a minimal effort level was also applied.

The main task consisted of 44 choice trials divided into 4 blocks. This included two ‘catch’ (non-dominated choice) trials, where the highest reward was offered for the lowest effort level. In order to be included in the analysis, participants were required to select the ‘correct’ answer on at least 1 of the catch trials. At the end of each block, participants rated their sense of achievement upon successful effort exertion, sense of pleasure upon gaining rewards, and overall boredom levels, using an interactive slider.

Interventions

The full content of the goal-setting and control interventions (described in the main text) is available here.

Participants

The initial (discovery) sample consisted of N=100 participants (N=0 excluded from behavioural data analysis), and the replication sample consisted of N=102 participants (N=0 excluded). A total of N=5 participants were excluded from analyses that included self-report data, for providing improbable answers to infrequency (catch) items.

Initial statistical analysis

Preliminary statistical analysis of choice behaviour was a via mixed-effects logistic regression model, as implemented in `lme4` (Bates et al., 2015). Individual choices were categorized as to whether or not the higher effort/higher reward option was chosen on each trial, and modelled as

$$\text{choice} \sim \text{interventionCondition} * \text{taskNo} + \text{trialNo} + (1|\text{subID}) \tag{1}$$

Where appropriate, pairwise differences were assessed using follow-up t-tests using the Tukey adjustment for multiple comparisons, as implemented in the R package `emmeans`.

Hierarchical Bayesian analysis

The most parsimonious model of choice behaviour, taking into account parameter recovery from the optimized task design and model comparison results in pilot datasets (see above), was a simple linear model with two free participant-level parameters representing reward and effort sensitivity.

$$V_{i,s,t} = rewSens_s * reward_{i,s,t} - effSens_s * effort_{i,s,t} \tag{2}$$

where V is the value of each choice option (i) for each trial (t) and session (s; time 1 or time 2), based on the reward offered (*reward*), required effort (*effort*) and participant reward (*rewSens*) and effort sensitivity (*effSens*) parameters for that time point. As described above, we assumed that task parameters across time points (pre- vs post-intervention) were drawn from multivariate normal distributions.

$$\begin{aligned}
\begin{bmatrix} rewSens_1 \\ rewSens_2 \end{bmatrix} &\sim MVNormal\left(\begin{bmatrix} rewSens_{\mu,1} \\ rewSens_{\mu,2} \end{bmatrix}, \sigma_{rewSens}\right) \\
\begin{bmatrix} effSens_1 \\ effSens_2 \end{bmatrix} &\sim MVNormal\left(\begin{bmatrix} effSens_{\mu,1} \\ effSens_{\mu,2} \end{bmatrix}, \sigma_{effSens}\right)
\end{aligned} \tag{3}$$

where $effSens_{\mu,s}$ and $rewSens_{\mu,s}$ are the group-level means for each parameter and time-point, and σ is the covariance between individual-level parameters across time-points (prior correlation between time-points was set to be uniform over [-1,1], using an $LKJ(1)$ prior). Choice values ($V_{i,s,t}$) were assumed to map onto observed choice data (y) using a simple Bernoulli likelihood function.

$$y_{p,s,t} \sim Bern(logit(V_{2,s,t} - V_{1,s,t})) \tag{4}$$

Participant-level parameter estimates were constructed using non-centered reparameterization in order to separate the hierarchical parameters and lower-level parameters in the prior (Papaspiliopoulos et al., 2007). For each parameter (e.g., ϕ) and time point s, participant-level estimates ($\phi_{p,s}$) were constructed from a group mean ($\phi_{\mu,s}$) and an individual offset ($\tilde{\phi}_{p,s}$). The between-subjects effects of intervention group were then modelled as:

$$
\begin{aligned}
\phi_{p,1} &= \phi_{\mu,1} + \tilde{\phi}_{p,1} \\
\phi_{p,2} &= \begin{cases} \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{INT}, & \text{if active intervention} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2}, & \text{otherwise} \end{cases}
\end{aligned}
\tag{5}
$$

where ϕ_{INT} is a group-level parameter describing potential effects of allocation to the active intervention on parameter estimates at time 2. For all models, the priors for effects of active intervention on parameter estimates were centered on 0 ($\phi_{INT} \sim N(0, 1)$). For full details of parameter constraints and model priors see Supplementary Methods.

Causal attribution studies

Causal attribution task

Code for implementing the task and a link to a demo version is available here. The task was coded in javascript using the `jsPsych` library, version 7.2.1 (de Leeuw, 2015).

Participants were instructed that during the task they would be asked to imagine themselves in various everyday situations. For each situation, they were asked to picture the situation described as clearly as they could ("*as if the events were happening to them right now*"), and then choose which of several possible explanations listed below they thought most likely. Specifically, participants were informed that, although events can have multiple different causes, they should choose the explanation they thought closest to the main reason the event happened, if it had actually happened to them.

Participants were presented with 32 event scenarios (16 positive and 16 negative events, randomly interleaved), divided into two blocks. Event scenarios were based on analysis of the previous literature (Alloy et al., 2000; Kinderman and Bentall, 1996; Wisco and Nolen-Hoeksema, 2010) and drawn from interpersonal (e.g., ”Someone you are close to tells you that they admire you”), professional/academic ("You and your friends do a general knowledge quiz and you get the lowest score"), and general life-functioning domains ("You fix something around the house that you have been meaning to get done for a while"). For each event, participants were asked to choose between four response options that varied orthogonally in terms of internal-external and global-specific explanation types, derived from examples provided in (Abramson et al., 1978). For example, for the event ”You find out that someone you consider a friend has talked about you negatively behind your back", possible explanations were ”Deep down, my friends don't really like me" (*internal-global*), ”I probably did something recently to annoy them" (*internal-specific*), ”Everyone has bad things said about them sometimes" (*external-global*), and ”My friend was probably just in a bad mood and letting off steam" (*external-specific*). We chose to focus on these two attribution dimensions as these have been most reliably linked in the past to low mood symptoms (Mezulis et al., 2004; Pearson et al., 2015). Full details of scenarios by event type (valence; interpersonal/not), category (close relationship/friends/contemporaries/colleagues; general performance-professional/general performance-academic; general life functioning), and possible attributions are available here.

Event scenarios for the final task version were chosen on the basis of analysis of a fuller (128) item set during pilot testing. Specifically, responses to the full item set were collected in $N=100$ participants, and the data modelled using Item-Response Theory. Subsets of items with the highest discriminability parameters for latent tendencies to make internal and global attributions for positive and negative events were then derived, ensuring final item sets with balanced positive/negative event frequencies and that all items contribute meaningfully to trait parameter estimates (posterior mean discriminability >1). We further conducted internal consistency, split-half analysis of attribution type counts, and test-retest reliability analysis of our trait parameter estimates, in order to ensure consistent responding across event types and over time (see main text and Supplementary Results). Given the novelty of this task, we also sought to validate the derived measures by relating them to negative self-beliefs as measured by the DAS, and current levels of depression and social anxiety symptoms (Supplementary Results). Finally, given the likelihood that responses to realistic social/professional scenarios might be influenced by individual and social factors, we examined if trait parameter estimates varied substantially according to various relevant measures (e.g., age, functional disability, minoritized status; see Supplementary Results).

The final item set did not include catch trials, but we applied the following exclusion rules to participants' choice data: median response time was required to be >2s, and proportionate choice of each response option position (e.g., top-left) was required to be <75% (participants were aware of these rules prior to completing the task, and informed that their compensation for taking part in the study may depend on these rules; different response options were displayed randomly in each position on each trial).

Interventions

Taking inspiration from materials described in Yeager et al. (2022), both active and control interventions were in the form of a series of interactive worksheets, requiring participants to select answers from multiple potential options during worked examples, and provide input based on recent positive and negative experiences from their own lives. The full content of the cognitive restructuring and control interventions (described in the main text) is available here.

Participants

The initial (discovery) sample consisted of $N=100$ participants, and the replication sample of $N=100$ participants (0 were excluded from either sample based on task data according to the above criteria). Across these samples, no participants additionally were excluded from analyses including self-report data.

Initial statistical analysis

Preliminary statistical analysis of choice behaviour was via mixed-effects logistic regression models. Individual choices on each trial were categorized according to whether an internal (*vs* external), and global (*vs* specific) attribution was selected, and the two orthogonal choice dimensions were separately modelled as:

$$choice_{internal} \sim interventionCondition * itemValence * taskNo + (1|subID) \tag{6}$$
$$choice_{global} \sim interventionCondition * itemValence * taskNo + (1|subID) \tag{7}$$

Hierarchical Bayesian analysis

For analysis of causal attribution task data, we used a simple model based on single-parameter IRT model to infer parameters governing a tendency to make internal and global attributions, based on a non-linear analysis of the pattern of responses across trials. Given evidence of valence-related asymmetry in attribution tendencies in both our data and the wider literature (see Supplementary Results, main text), separate parameters were used to describe internal and global attribution tendencies for positive and negative events. Specifically, participants' choices on each trial were coded along two dimensions, according to whether an internal (*vs* external) or global (v*s* specific) response option was chosen ($y\_internal$ and y_global, respectively), with the resulting data analysed within a single hierarchical model with 4 free participant-level parameters.

$$\begin{gathered} y\_internal_{p,s,v} \sim Bern(\theta_{internal,p,s,v}) \\ y\_global_{p,s,v} \sim Bern(\theta_{global,p,s,v}) \end{gathered} \tag{8}$$

where $\theta_{internal,p,s,v}$ and $\theta_{global,p,s,v}$ represent the latent traits governing a participant (*p*)'s tendency to make an internal or global attribution at that time point or session (*s*), separately for positively and negatively valenced (*v*) event scenarios. We chose this simple model as it maps intuitively onto concepts from attributional style theory (Abramson et al., 1978), on evidence that it accounted well for participants' choices in pilot data, and on the basis that final task items were chosen based on a more complex 2PL IRT analysis of a larger item set, in order to ensure good discriminability for our traits of interest (see Supplementary Results).

Given pilot data showing correlations between individuals' tendencies to make global and internal attributions for positive and negative events (Supplementary Results, Figure S7), and in order to allow maximum information to contribute to individual parameter estimates, we also assumed that individual tendencies to make internal and global attributions for each type of event were drawn from a multivariate normal distribution (allowing for direct estimation of covariance between attribution types within each session).

$$\begin{gathered} \begin{bmatrix} \theta_{internal,1,neg} \\ \theta_{global,1,neg} \\ \theta_{internal,2,neg} \\ \theta_{global,2,neg} \end{bmatrix} \sim MVNormal\left(\begin{bmatrix} \theta_{internal,\mu,1,neg} \\ \theta_{global,\mu,1,neg} \\ \theta_{internal,\mu,2,neg} \\ \theta_{global,\mu,2,neg} \end{bmatrix}, \sigma_{\theta,neg} \right) \\ \begin{bmatrix} \theta_{internal,1,pos} \\ \theta_{global,1,pos} \\ \theta_{internal,2,pos} \\ \theta_{global,2,pos} \end{bmatrix} \sim MVNormal\left(\begin{bmatrix} \theta_{internal,\mu,1,pos} \\ \theta_{global,\mu,1,pos} \\ \theta_{internal,\mu,2,pos} \\ \theta_{global,\mu,2,pos} \end{bmatrix}, \sigma_{\theta,pos} \right) \end{gathered} \tag{9}$$

where $\theta_{internal,\mu,s,v}$ and $\theta_{internal,\mu,s,v}$ are the group-level means for each parameter and time-point (modelled separately for positive, *pos*, and negative, *neg*, events), and σ is the covariance between individual-level parameters across attribution types and time points. For full descriptions of parameter constrains and model priors see Supplementary Methods.

Effects of belonging to the active intervention condition on parameter estimates at time 2 were modelled as described in Equation 5.

Crossover study

For the crossover study, participants were randomly assigned to experimental conditions in a 2*2 factorial design of task (reward-effort decision-making or causal attribution) and intervention (goal-setting or cognitive restructuring) condition. Tasks and intervention materials were as described previously.

Analysis was via the same hierarchical models of each task as described above, with the effect of active intervention at time 2 now representing the effect of allocation to the goal-setting *vs* cognitive restructuring intervention on task measures, rather than either intervention alone *vs* a well-matched control.

Participants

N=400 total participants were recruited for the crossover study. N=192 were randomized to complete the reward-effort decision-making task, with N=5 excluded from behaviour-only analyses on the basis of catch trial performance. N=208 were randomized to the causal attribution task, with no participants excluded from behavioural analyses. A further N=5 participants were excluded from analyses that included self-report data, on the basis of response to infrequency items.

Heterogeneity of treatment effects analysis

Before examining individual differences related to magnitude of intervention effects, we first sought to determine if we had evidence across samples of significant individual differences in responses to active compared to control interventions (Hopkins, 2015; Norbury and Seymour, 2018). This analysis involves comparing standard deviations of change scores in the active and control groups, in order to assess evidence for greater variance in outcomes in the active intervention group (since we assume control arm change score variance represents effects of individual variability over time and measurement error).

Change scores were defined as differences in mean posterior parameter estimate between time points, and change scores in each arm were standardized by the SD of baseline (pre-treatment) posterior means. The standard deviations of individual responses to the active treatment were then calculated as $SD_{IR} = \sqrt{SD_{Act}{}^2 - SD_{Con}{}^2}$; where SD_{Act} and SD_{Con} are the standardized standard deviations of the change scores in the active and control groups, respectively. Confidence limits for SD_{IR} were obtained by assuming its sampling variance is normally distributed, $SD_{IRse} = \sqrt{2 * (SD_{Act}{}^4/DF_{Act} + SD_{Con}{}^4/DF_{Con})}$, such that the 95% CI was calculated as $SD_{IR} \pm 1.96 * SD_{IRse}$. DF_{Act} and DF_{Con} are the degrees of freedom of the standard deviations in the two groups ($N-1$). Where standardized SDs are used, 0.1, 0.3, 0.6, represent thresholds for small, moderate, and large individual response effects.

Joint modelling of self-report and task data

Self-reported symptom data model

Behavioural amotivation trait estimates were constructed from the 6 AMI behavioural amotivation subscale items plus the PHQ9 items "*little interest or pleasure in doing things*" and "*feeling tired or having little energy*". Negative cognition estimates were constructed from the 8 DAS short-form items plus the PHQ9 items "*feeling down, depressed, or hopeless*" and "*feeling bad about yourself or that you are a failure*" (Figure 6a,b).

In order to construct individual trait estimates, self-report data were analysed via a Graded Response Model (GRM) (Samejima, 1969) – a form of IRT model that was developed to make use of ordinal responses such as ordered Likert scales (essentially, an ordered logistic extension of the model described in Equation S1). Given our relatively limited N ($\sim$200 per sample), we allowed items to contribute to their hypothetical latent trait only (i.e., we fit two parallel unidimensional GRMs, rather than a more complex multidimensional GRM). This process yields approximately normally distributed latent trait estimates.

Combining self-report and task behaviour data

Joint modelling allows maximum use of participant-level data, whilst retaining information about uncertainty or precision of each kind of measurement (Turner et al., 2017; Haines et al., 2020a; Haines, 2021; Hopkins et al., 2021).

For the joint model, individual estimates for trait amotivation (θ_A) and/or trait negative cognition (θ_N; constructed as above), were allowed to influence the effect of intervention on time 2 parameter estimates (ϕ_{INT}) found to show evidence of heterogeneous individual responses via the inclusion of additional β weight parameters (β_{INT}; see Haines et al. 2020a; Hopkins et al. 2021) for previous examples of this approach). These β weights can interpreted similarly as in a standard regression model, with the group-level intervention effect (ϕ_{INT}) now representing the intercept (see below).

In order to account for potential regression-to-the-mean effects caused by baseline associations between task performance and self-reported clinical symptoms (see e.g., Figure S7), joint models also included β weights for the same parameter estimate at time 1 (β_{BASE}).

$$\begin{aligned}
\phi_{p,1} &= \phi_{\mu,1} + \tilde{\phi}_{p,1} + \beta_{BASE}\theta_{A/N} \\
\phi_{p,2} &= \begin{cases} \phi_{\mu,2} + \tilde{\phi}_{p,2} + \phi_{INT} + \beta_{INT}\theta_{A/N}, & \text{if active intervention} \\ \phi_{\mu,2} + \tilde{\phi}_{p,2}, & \text{otherwise} \end{cases}
\end{aligned} \tag{10}$$

Posterior estimates for β weights with a 90% credible interval that excluded zero were taken as evidence that the trait estimates were meaningfully related to the effect of interest.

Acknowledgements

This research was funded by a research grant from Koa Health to QJMH, SF and RD and a grant from Carigest S.A to QJMH. We acknowledge support by the UCLH NIHR BRC. TUH is supported by a Sir Henry Dale Fellowship (211155/Z/18/Z; 211155/Z/18/B; 224051/Z/21) from Wellcome and The Royal Society. SMF is a CIFAR Fellow in the Brain, Mind and Consciousness Program, and funded by a Wellcome/Royal Society Sir Henry Dale Fellowship (206648/Z/17/Z) and a Philip Leverhulme Prize from the Leverhulme Trust. QJMH acknowledges grant funding from the Wellcome Trust (221826/Z/20/Z). The Max Planck-UCL Centre for Computational Psychiatry and Ageing Research is a joint initiative supported by University College London and the Max Planck Society.

AN declares no competing interests. QJMH has obtained fees and options for consultancies for Aya Technologies and Alto Neuroscience. TUH has obtained fees and options for consultancies for Limbic Ltd.

References

Abramson, L. Y., Seligman, M. E., and Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, 87:49–74.

Alloy, L. B., Abramson, L. Y., Hogan, M. E., Whitehouse, W. G., Rose, D. T., Robinson, M. S., Kim, R. S., and Lapkin, J. B. (2000). The Temple-Wisconsin Cognitive Vulnerability to Depression Project: Lifetime history of Axis I psychopathology in individuals at high and low cognitive risk for depression. *Journal of Abnormal Psychology*, 109:403–418.

Ang, Y.-S., Lockwood, P., Apps, M. A. J., Muhammed, K., and Husain, M. (2017). Distinct Subtypes of Apathy Revealed by the Apathy Motivation Index. *PLOS ONE*, 12(1):e0169938.

Barber, J. P. and DeRubeis, R. J. (1989). On second thought: Where the action is in cognitive therapy for depression. *Cognitive Therapy and Research*, 13(5):441–457.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67:1–48.

Bear, H. A., Nunes, L. A., DeJesus, J., Liverpool, S., Moltrecht, B., Neelakantan, L., Harriss, E., Watkins, E., and Fazel, M. (2022). Determination of Markers of Successful Implementation of Mental Health Apps for Young People: Systematic Review. *Journal of Medical Internet Research*, 24(11):e40347.

Beck, A. T., Rush, A. J., Shaw, B. F., and Emery, G. (1987). *Cognitive Therapy of Depression*. Guildford Press.

Beevers, C. G., Strong, D. R., Meyer, B., Pilkonis, P. A., and Miller, I. W. (2007). Efficiently assessing negative cognition in depression: An item response theory analysis of the Dysfunctional Attitude Scale. *Psychological Assessment*, 19:199–209.

Berwian, I. M., Wenzel, J. G., Collins, A. G. E., Seifritz, E., Stephan, K. E., Walter, H., and Huys, Q. J. M. (2020). Computational Mechanisms of Effort and Reward Decisions in Patients With Depression and Their Association With Relapse After Antidepressant Discontinuation. *JAMA Psychiatry*, 77(5):513–522.

Bonnelle, V., Veromann, K.-R., Burnett Heyes, S., Lo Sterzo, E., Manohar, S., and Husain, M. (2015). Characterization of reward and effort mechanisms in apathy. *Journal of Physiology-Paris*, 109(1):16–26.

Borghouts, J., Eikey, E., Mark, G., Leon, C. D., Schueller, S. M., Schneider, M., Stadnick, N., Zheng, K., Mukamel, D., and Sorkin, D. H. (2021). Barriers to and Facilitators of User Engagement With Digital Mental Health Interventions: Systematic Review. *Journal of Medical Internet Research*, 23(3):e24387.

Brown, V. M., Chen, J., Gillan, C. M., and Price, R. B. (2020). Improving the Reliability of Computational Analyses: Model-Based Planning and Its Relationship With Compulsivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 5(6):601–609.

Buckman, J. E. J., Saunders, R., Stott, J., Cohen, Z. D., Arundell, L.-L., Eley, T. C., Hollon, S. D., Kendrick, T., Ambler, G., Watkins, E., Gilbody, S., Kessler, D., Wiles, N., Richards, D., Brabyn, S., Littlewood, E., DeRubeis, R. J., Lewis, G., and Pilling, S. (2022). Socioeconomic Indicators of Treatment Prognosis for Adults With Depression: A Systematic Review and Individual Patient Data Meta-analysis. *JAMA Psychiatry*, 79(5):406–416.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76:1–32.

Cheng, C. and Ebrahimi, O. V. (2023). A meta-analytic review of gamified interventions in mental health enhancement. *Computers in Human Behavior*, 141:107621.

Clark, D. A. (2022). Cognitive Reappraisal. *Cognitive and Behavioral Practice*, 29(3):564–566.

Clark, D. M. (2018). Realising the Mass Public Benefit of Evidence-Based Psychological Therapies: The IAPT Program. *Annual review of clinical psychology*, 14:159–183.

Connor, K. M., Kobak, K. A., Churchill, L. E., Katzelnick, D., and Davidson, J. R. (2001). Mini-SPIN: A brief screening assessment for generalized social anxiety disorder. *Depression and Anxiety*, 14(2):137–140.

Craske, M. (2022). Screening and Treatment for Anxiety & Depression (S.T.A.N.D): Alacrity Center Signature Project on Triaging and Adapting to Level of Care. Clinical trial registration NCT05591937, clinicaltrials.gov.

Craske, M. G., Herzallah, M. M., Nusslock, R., and Patel, V. (2023). From neural circuits to communities: an integrative multidisciplinary roadmap for global mental health. *Nature Mental Health*, 1(1):12–24.

Cristea, I. A., Huibers, M. J. H., David, D., Hollon, S. D., Andersson, G., and Cuijpers, P. (2015). The effects of cognitive behavior therapy for adult depression on dysfunctional thinking: A meta-analysis. *Clinical Psychology Review*, 42:62–71.

Cuijpers, P., Ciharova, M., Quero, S., Miguel, C., Driessen, E., Harrer, M., Purgato, M., Ebert, D., and Karyotaki, E. (2022). The Contribution of "Individual Participant Data" Meta-Analyses of Psychotherapies for Depression to the Development of Personalized Treatments: A Systematic Review. *Journal of Personalized Medicine*, 12(1):93.

Cuijpers, P., Cristea, I. A., Karyotaki, E., Reijnders, M., and Huibers, M. J. (2016). How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World Psychiatry*, 15(3):245–258.

Cuijpers, P., Huibers, M. J., and Reijnders (2019). The role of common factors in psychotherapy outcome. *Annual Review of Clinical Psychology*, 15:207–231.

Cuijpers, P., Karyotaki, E., Reijnders, M., and Huibers, M. J. H. (2018). Who benefits from psychotherapies for adult depression? A meta-analytic update of the evidence. *Cognitive Behaviour Therapy*, 47(2):91–106.

Daw, N. D. (2011). Trial-by-trial data analysis using computational models. In *Decision Making, Affect, and Learning: Attention and Performance XXIII*. Oxford University Press.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, 47:1–12. Place: Germany Publisher: Springer.

Driessen, E., Cohen, Z. D., Lorenzo-Luaces, L., Hollon, S. D., Richards, D. A., Dobson, K. S., Dimidjian, S., Delgadillo, J., Vázquez, F. L., McNamara, K., Horan, J. J., Gardner, P., Oei, T. P., Mehta, A. H. P., Twisk, J. W. R., Cristea, I. A., and Cuijpers, P. (2022). Efficacy and moderators of cognitive therapy versus behavioural activation for adults with depression:

study protocol of a systematic review and meta-analysis of individual participant data. *BJPsych Open*, 8(5):e154.

Ebrahimi, O. V., Burger, J., Hoffart, A., and Johnson, S. U. (2021). Within- and across-day patterns of interplay between depressive symptoms and related psychopathological processes: a dynamic network approach during the COVID-19 pandemic. *BMC Medicine*, 19(1):317.

Eronen, M. I. (2020). Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, 59:100785.

Ezawa, I. D. and Hollon, S. D. (2023). Cognitive restructuring and psychotherapy outcome: A meta-analytic review. *Psychotherapy*.

Forbes, C. N. (2020). New directions in behavioral activation: Using findings from basic science and translational neuroscience to inform the exploration of potential mechanisms of change. *Clinical Psychology Review*, 79:101860.

Furukawa, T. A., Suganuma, A., Ostinelli, E. G., Andersson, G., Beevers, C. G., Shumake, J., Berger, T., Boele, F. W., Buntrock, C., Carlbring, P., Choi, I., Christensen, H., Mackinnon, A., Dahne, J., Huibers, M. J. H., Ebert, D. D., Farrer, L., Forand, N. R., Strunk, D. R., Ezawa, I. D., Forsell, E., Kaldo, V., Geraedts, A., Gilbody, S., Littlewood, E., Brabyn, S., Hadjistavropoulos, H. D., Schneider, L. H., Johansson, R., Kenter, R., Kivi, M., Björkelund, C., Kleiboer, A., Riper, H., Klein, J. P., Schröder, J., Meyer, B., Moritz, S., Bücker, L., Lintvedt, O., Johansson, P., Lundgren, J., Milgrom, J., Gemmill, A. W., Mohr, D. C., Montero-Marin, J., Garcia-Campayo, J., Nobis, S., Zarski, A.-C., O'Moore, K., Williams, A. D., Newby, J. M., Perini, S., Phillips, R., Schneider, J., Pots, W., Pugh, N. E., Richards, D., Rosso, I. M., Rauch, S. L., Sheeber, L. B., Smith, J., Spek, V., Pop, V. J., Ünlü, B., Bastelaar, K. M. P. v., Luenen, S. v., Garnefski, N., Kraaij, V., Vernmark, K., Warmerdam, L., Straten, A. v., Zagorscak, P., Knaevelsrud, C., Heinrich, M., Miguel, C., Cipriani, A., Efthimiou, O., Karyotaki, E., and Cuijpers, P. (2021). Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. *The Lancet Psychiatry*, 8(6):500–511.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020). Bayesian Workflow.

Graham, A. K., Lattie, E. G., and Mohr, D. C. (2019). Experimental Therapeutics for Digital Mental Health. *JAMA Psychiatry*, 76(12):1223–1224.

Greenberg, L. S. (2015). *Emotion-focused therapy: Coaching clients to work through their feelings, 2nd ed.* Emotion-focused therapy: Coaching clients to work through their feelings, 2nd ed. American Psychological Association, Washington, DC, US.

Griffiths, T. L., Kemp, C., and Tenenbaum, J. B. (2008). *Bayesian Models of Cognition*. Cambridge Univ Press, Cambridge.

Haberman, S. J. (2008). When Can Subscores Have Value? *Journal of Educational and Behavioral Statistics*, 33(2):204–229.

Haines, N. (2021). *Integrating Trait and Neurocognitive Mechanisms of Externalizing Psychopathology: A Joint Modeling Framework for Measuring Impulsive Behavior*. PhD thesis, The Ohio State University.

Haines, N., Beauchaine, T. P., Galdo, M., Rogers, A. H., Hahn, H., Pitt, M. A., Myung, J. I., Turner, B. M., and Ahn, W.-Y. (2020a). Anxiety Modulates Preference for Immediate Rewards Among Trait-Impulsive Individuals: A Hierarchical Bayesian Analysis. *Clinical Psychological Science*, 8(6):1017–1036.

Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., and Turner, B. M. (2020b). Theoretically Informed Generative Models Can Advance the Psychological and Brain Sciences: Lessons from the Reliability Paradox.

Hedge, C., Bompas, A., and Sumner, P. (2020). Task Reliability Considerations in Computational Psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 5(9):837–839.

Hedge, C., Powell, G., and Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, 50(3):1166–1186.

Holmes, E. A., Craske, M. G., and Graybiel, A. M. (2014). Psychological treatments: A call for mental-health science. *Nature*, 511(7509):287–289.

Holmes, E. A., Ghaderi, A., Harmer, C. J., Ramchandani, P. G., Cuijpers, P., Morrison, A. P., Roiser, J. P., Bockting, C. L. H., O'Connor, R. C., Shafran, R., Moulds, M. L., and Craske, M. G. (2018). The Lancet Psychiatry Commission on psychological treatments research in tomorrow's science. *The Lancet Psychiatry*, 5(3):237–286.

Hopkins, A. K., Dolan, R., Button, K. S., and Moutoussis, M. (2021). A Reduced Self-Positive Belief Underpins Greater Sensitivity to Negative Evaluation in Socially Anxious Individuals. *Computational Psychiatry*, 5(1):21–37.

Hopkins, W. G. (2015). Individual responses made easy. *Journal of Applied Physiology*, 118(12):1444–1446.

Huibers, M. J. H., Lorenzo-Luaces, L., Cuijpers, P., and Kazantzis, N. (2021). On the Road to Personalized Psychotherapy: A Research Agenda Based on Cognitive Behavior Therapy for Depression. *Frontiers in Psychiatry*, 11.

Huys, Q. J. M., Russek, E. M., Abitante, G., Kahnt, T., and Gollan, J. K. (2022). Components of Behavioral Activation Therapy for Depression Engage Specific Reinforcement Learning Mechanisms in a Pilot Study. *Computational Psychiatry*, 6(1):238–255.

Katahira, K. (2016). How hierarchical models improve point estimates of model parameters at the individual level. *Journal of Mathematical Psychology*, 73:37–58.

Kay, M. (2022). tidybayes: Tidy Data and Geoms for Bayesian Models.

Kazdin, A. E. (2009). Understanding how and why psychotherapy leads to change. *Psychotherapy Research*, 19(4-5):418–428.

Kazdin, A. E. and Blase, S. L. (2011). Rebooting Psychotherapy Research and Practice to Reduce the Burden of Mental Illness. *Perspectives on Psychological Science*, 6(1):21–37.

Kim, S., Moon, H., Modrák, M., and Säilynoja, T. (2023). SBC: Simulation Based Calibration for rstan/cmdstanr models.

Kinderman, P. and Bentall, R. P. (1996). A new measure of causal locus: the internal, personal and situational attributions questionnaire. *Personality and Individual Differences*, 20(2):261–264.

Kroenke, K., Spitzer, R. L., and Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, 16(9):606–613.

Kroenke, K., Spitzer, R. L., and Williams, J. B. W. (2003). The Patient Health Questionnaire-2: validity of a two-item depression screener. *Medical care*, 41(11).

Kruschke, J. K. (2021). Bayesian Analysis Reporting Guidelines. *Nature Human Behaviour*, 5(10):1282–1291.

Lee, C. T., Palacios, J., Richards, D., Hanlon, A. K., Lynch, K., Harty, S., Claus, N., Swords, L., O'Keane, V., Stephan, K. E., and Gillan, C. M. (2023). The Precision in Psychiatry (PIP) study: Testing an internet-based methodology for accelerating research in treatment prediction and personalisation. *BMC Psychiatry*, 23(1):25.

Long, B., Simson, J., Buxó-Lugo, A., Watson, D. G., and Mehr, S. A. (2023). How games can make behavioural science better. *Nature*, 613(7944):433–436.

Lorenzo-Luaces, L., German, R. E., and DeRubeis, R. J. (2015). It's complicated: The relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clinical Psychology Review*, 41:3–15.

Martell, C. R., Dimidjian, S., and Herman-Dunn, R. (2013). *Behavioral Activation for Depression: A Clinician's Guide*. Guilford Press.

Mason, J., Pownall, M., Palmer, A., and Azevedo, F. (2023). Investigating Lay Perceptions of Psychological Measures: A Registered Report.

McElreath, R. (2016). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, New York.

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., and Hankin, B. L. (2004). Is There a Universal Positivity Bias in Attributions? A Meta-Analytic Review of Individual, Developmental, and Cultural Differences in the Self-Serving Attributional Bias. *Psychological Bulletin*, 130:711–747.

Modrák, M., Moon, A. H., Kim, S., Bürkner, P., Huurre, N., Faltejsková, K., Gelman, A., and Vehtari, A. (2022). Simulation-Based Calibration Checking for Bayesian Computation: The Choice of Test Quantities Shapes Sensitivity.

Norbury, A. and Seymour, B. (2018). Response heterogeneity: Challenges for personalised medicine and big data approaches in psychiatry and chronic pain. *F1000Research*, 7:55.

Papaspiliopoulos, O., Roberts, G. O., and Sköld, M. (2007). A General Framework for the Parametrization of Hierarchical Models. *Statistical Science*, 22(1):59–73.

Paulus, M. P., Huys, Q. J. M., and Maia, T. V. (2016). A Roadmap for the Development of Applied Computational Psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(5):386–392.

Pearson, R. M., Heron, J., Button, K., Bentall, R. P., Fernyhough, C., Mahedy, L., Bowes, L., and Lewis, G. (2015). Cognitive styles and future depressed mood in early adulthood: The importance of global attributions. *Journal of Affective Disorders*, 171:60–67.

Quigley, L. and Dobson, K. S. (2017). Chapter 12 - Behavioral Activation Treatments for Depression. In Hofmann, S. G. and Asmundson, G. J. G., editors, *The Science of Cognitive Behavioral Therapy*, pages 291–318. Academic Press, San Diego.

Reiter, A. M., Atiya, N. A., Berwian, I. M., and Huys, Q. J. (2021). Neuro-cognitive processes as mediators of psychological treatment effects. *Current Opinion in Behavioral Sciences*, 38:103–109.

Rouder, J. N. and Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, 26(2):452–467.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, 34:100–100.

Schad, D. J., Betancourt, M., and Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, 26:103–126.

Schueller, S. M., Hunter, J. F., Figueroa, C., and Aguilera, A. (2019). Use of Digital Mental Health for Marginalized and Underserved Populations. *Current Treatment Options in Psychiatry*, 6(3):243–255.

Singh, S., Strong, R. W., Jung, L., Li, F. H., Grinspoon, L., Scheuer, L. S., Passell, E. J., Martini, P., Chaytor, N., Soble, J. R., and Germine, L. (2021). The TestMyBrain Digital Neuropsychology

Toolkit: Development and Psychometric Characteristics. *Journal of Clinical and Experimental Neuropsychology*, 43(8):786–795.

Szinay, D., Jones, A., Chadborn, T., Brown, J., and Naughton, F. (2020). Influences on the Uptake of and Engagement With Health and Well-Being Smartphone Apps: Systematic Review. *Journal of Medical Internet Research*, 22(5):e17572.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2020). Validating Bayesian Inference Algorithms with Simulation-Based Calibration.

Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Andrade, L., Borges, G., Bruffaerts, R., Bunting, B., Almeida, J. M. C. d., Florescu, S., Girolamo, G. d., Gureje, O., Haro, J. M., He, Y., Hinkov, H., Karam, E., Kawakami, N., Lee, S., Navarro-Mateu, F., Piazza, M., Posada-Villa, J., Galvis, Y. T. d., and Kessler, R. C. (2017). Undertreatment of people with major depressive disorder in 21 countries. *The British Journal of Psychiatry*, 210(2):119–124.

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., Carvalho, A. F., Keshavan, M., Linardon, J., and Firth, J. (2021). The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, 20(3):318–335.

Treadway, M. T., Buckholtz, J. W., Schwartzman, A. N., Lambert, W. E., and Zald, D. H. (2009). Worth the ‘EEfRT’? The Effort Expenditure for Rewards Task as an Objective Measure of Motivation and Anhedonia. *PLOS ONE*, 4(8):e6598.

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., and Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, 76:65–79.

van Dis, E. A. M., van Veen, S. C., Hagenaars, M. A., Batelaan, N. M., Bockting, C. L. H., van den Heuvel, R. M., Cuijpers, P., and Engelhard, I. M. (2020). Long-term Outcomes of Cognitive Behavioral Therapy for Anxiety-Related Disorders: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, 77(3):265–273.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432.

Wellcome (2021). What science has shown can help young people with anxiety and depression. Identifying and reviewing the ’active ingredients’ of effective interventions: Part 1. Technical report.

Wisco, B. E. and Nolen-Hoeksema, S. (2010). Interpretation bias and depressive symptoms: The role of self-relevance. *Behaviour Research and Therapy*, 48(11):1113–1122.

Wise, T. and Dolan, R. J. (2020). Associations between aversive learning processes and transdiagnostic psychiatric symptoms in a general population sample. *Nature Communications*, 11(1):4179.

World Health Organization (2012). WHO Disability Assessment Schedule (WHODAS 2.0).

Yeager, D. S., Bryan, C. J., Gross, J. J., Murray, J. S., Krettek Cobb, D., H. F. Santos, P., Gravelding, H., Johnson, M., and Jamieson, J. P. (2022). A synergistic mindsets intervention protects adolescents from stress. *Nature*, 607(7919):512–520.

Zorowitz, S. and Niv, Y. (2023). Improving the reliability of cognitive task measures: A narrative review. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 0(0). Publisher: Elsevier.

Zorowitz, S., Niv, Y., and Bennett, D. (2021). Inattentive responding can induce spurious associations between task behavior and symptom measures.

Supplementary Material

Supplementary Methods

Hierarchical Bayesian modelling of reward-effort decision-making task data

Priors for group-level parameter means were specified using standard normal distributions, $\phi_{\mu,s} \sim N(0,1)$. Priors for group-level parameter standard deviations were specified as $\phi_{\sigma,s} \sim Cauchy(0,1)$. Priors for individual participant deviations from group-level parameter estimates (effort and reward sensitivity) were also specified using standard normal distributions ($\tilde{\phi}_{p,s} \sim N(0,1)$). The prior over the correlation matrix relating parameter estimates across sessions was set to be uniform over $[-1,1]$ using an $LKJ(1)$ prior.

The priors for group-level effects of intervention on parameter estimates at time 2 (ϕ_{INT}), and group-level beta weights governing influence of latent traits on effects of interest (β_{INT}, β_{BASE}), were also specified as standard normal distributions (i.e., centred on zero).

Individual parameter estimates for effort and reward sensitivity were constrained to be positive and constrained to be in the range $[0,10]$ and $[0,3]$, respectively, on the basis of empirical posterior distributions observed in pilot data and values considered to observable based on the range of effort and reward values in our test trial array (this explains the plateauing of inferred values outside this range in simulation-based calibration analysis; Figure 1d).

Hierarchical Bayesian modelling of causal attribution task data

Priors for group-level parameter means were specified using standard normal distributions, $\phi_{\mu,s} \sim N(0,1)$. Priors for group-level parameter standard deviations were specified as $\phi_{\sigma,s} \sim cauchy(0,1)$. Priors for individual participant deviations from group-level parameter estimates ($\theta_{internal,p,s,neg}$, $\theta_{internal,p,s,pos}$, $\theta_{global,p,s,neg}$, $\theta_{global,p,s,pos}$) were also specified using standard normal distributions ($\tilde{\phi}_{p,s} \sim N(0,1)$). The prior over the correlation matrix relating parameter estimates across sessions was set to be uniform over $[-1,1]$ using an $LKJ(1)$ prior.

The priors for group-level effects of intervention on parameter estimates at time 2 (ϕ_{INT}), and group-level beta weights governing influence of latent traits on effects of interest (β_{INT}, β_{BASE}), were also specified as standard normal distributions (i.e., centred on zero).

Individual parameter estimates for latent traits governing tendency to attribute positive and negative events to internal and global causes were unconstrained but passed to the Bernoulli observation function (Equation 8) using an inverse logit transform, scaling probability of endorsement to the range $[0,1]$ (see e.g., Figure 1de).

Supplementary Results

Linear mixed-effects analysis of task behaviour

Effects of planning/goal-setting on reward-effort choice behaviour. Choice of higher-effort/ higher-reward options on each trial were analysed via linear mixed-effects models, with the within-subjects factors of time (pre *vs* post intervention) and trial number, and between-subjects factor of intervention group (planning *vs* control). In both initial and replication samples, there was a significant group*time interaction ($F_{1,8697} = 15.4; F_{1,8871} = 34.8$; both $p < 0.001$; Figure S4a,b). Follow-up pairwise comparisons with Tukey correction for multiple comparisons revealed that in the initial sample, this was due to a decrease in high-effort/high-reward choice options from time 1 to time 2 in the control ($t_{8967} = -3.96, p < 0.001$), but not the planning intervention group ($t_{8967} = 1.20, p > 0.5$). In the replication sample, this was due to a decrease in higher effort choices from time 1 to time 2 in the control group ($t_{8871} = -3.77, p < 0.001$), but increase in higher effort choices in the planning group ($t_{8871} = 4.57, p < 0.001$).

Effects of planning/goal-setting on self-reported sense of achievement, pleasure, and boredom during the reward-effort decision-making task. Participants were asked to rate their sense of achievement on successful effort exertion, pleasure on gaining rewards, and boredom levels, following each four blocks of the task. Ratings data were analysed via linear mixed-effects models, with the within-subjects factors of time (pre *vs* post intervention) and block number, and between-subjects factor of intervention group (planning *vs* control). In both initial and replication samples, there were significant group*time interaction effects on sense of achievement ($F_{1,679} = 59.3, F_{1,700} = 34.0$), pleasure on gaining rewards ($F_{1,679} = 58.8, F_{1,700} = 62.3$), and boredom ($F_{1,679} = 102.0, F_{1,700} = 14.9$; all $p < 0.001$; Figure S4c,d). Follow-up pairwise comparisons with Tukey correction for multiple comparisons revealed that in the initial sample, this was due to higher sense of achievement, higher pleasure on reward receipt, and lower boredom, at time 2 in the planning vs control group ($t_{108} = 3.35, t_{108} = 3.82, t_{103} = -3.75$; all $p < 0.01$). In the replication sample, this was due to higher sense of achievement and pleasure on reward receipt, at time 2 in the planning group ($t_{111} = 2.98, t_{110} = 2.91; p < 0.03$), and smaller increase in self-reported boredom at time 2 in the planning vs control group ($t_{700} = 10.5, t_{700} = 4.84$; both $p < 0.001$).

Goal-setting behaviour in planning intervention participants. Post-intervention, prior to each block of the task, participants in the planning intervention were asked to set an achievable goal for the amount of reward they would like to earn within that block. Participants in the control condition were asked to enter a liking rating for different kinds of online games. Participants in the planning condition tended to exceed their goals for each block (achieved > goal reward; Figure S5a,b). Analysis of ratings data via repeated-measures ANOVA with the within-subjects factor of block and between-subjects factor of intervention group revealed that in both initial and replication samples there was a significant interaction between intervention group and block number on answers ($F_{2.4,236} = 19, F_{2.5,245} = 8.9; p < 0.001$). Specifically, participants in the planning group increased the ambitiousness of their goals across blocks (difference between first and last blocks: $t_{59} = 3.70, t_{50} = 3.14; p < 0.01$), whereas participants in the control group did not show a systematic effect of block on liking ratings (difference between first and last block: $t_{41} = -2.78, t_{52} = -1.84, p < 0.07$; Figure S5a,b).

Effects of psychoeducation/reappraisal practice on attribution choice. Choices of internal (*vs* external) and global (*vs* specific) explanations for each scenario with analysed via linear mixed-effects models with the within-subjects factors of scenario valence (positive or negative events) and time (pre *vs* post intervention), and the within-subjects factor of intervention group (psychoeducation *vs* control). In both initial discovery and replication samples, there were found to be significant interactions between time, item valence, and intervention group on frequency of internal attributions ($F_{1,6294} = 10.9, F_{1,6294} = 5.0$; both $p < 0.03$, Figure S6). Follow-up pairwise comparisons with Tukey correction for multiple comparisons revealed that this was due to a decrease in internal attribution of negative events between time 1 and time 2 in the psychoeducation group ($t_{6294} = -7.3, t_{6294} = 6.0, p < 0.001$), but not control group ($t_{6294} = -2.7, t_{6294} = 3.0, p > 0.05$). Conversely, both groups showed increased frequency of internal attributions for positive events at time 2 (psychoeducation group: $t_{6294} = 7.5, t_{6294} = 6.6$; control group: $t_{6294} = 5.6, t_{6294} = 6.5$; all $p < 0.001$). There was evidence of an interaction between time, item valence, and intervention group on frequency of choice of global attributions in the initial discovery ($F_{1,6294} = 6.1, p = 0.014$), but not the replication sample ($F_{1,6294} = 0.6, p > 0.4$).

Hierarchical Bayesian model-based analysis of task behaviour

Descriptions of posterior group-level parameter estimates and relevant sampling diagnostics for analyses presented in Figures 2-5 are listed in Tables S2-5. For each parameter, tables describe the posterior mean, standard error of the posterior mean, and 90% posterior probability quantiles. In accordance with recommendations for reporting of Bayesian analyses, tables also include the effective sample size for each parameter (an estimate of the number of independent draws from the posterior distribution of that parameter) and $\hat{R}$ (Gelman-Rubin) statistics, which index convergence across different sampling chains.

IRT-modelling of self-report data

Posterior discriminability parameters (the IRT parameter describing how well each item differentiates between individuals high and low on latent trait scores) for each sample are depicted in Figure 6a,b. For the 'amotivation' trait, the top-discriminating items across all samples were PH9 1,4 ("*little interest or pleasure in doing things*"; "*feeling tired or having little energy*"), and AMI 5,10 ("*I make decisions firmly and without hesitation*"; "*I don't like to laze around*"; both reverse-scored). For the 'negative cognition' trait, there was more variance in the top-discriminating items across samples, as many items had similar posterior discriminability estimates. Top items across samples were mainly from the DAS scale, and included DAS 1,4,5 ("*If I don't set the highest standards for myself, I am likely to end up a second-rate person*", "*I am nothing if a person I love doesn't love me*", "*If other people know what you are really like, they will think less of you*").

Development and validation of the causal attribution task

Due to our use of a novel scenario battery and response option structure, extensive pilot work was carried out during the development of the causal attribution task. Specifically, data from an initial pilot sample (*N*=102) was initially collected on the full set of items (128 total). In order to assess if we were able to consistently measure tendency to attribute events to negative and positive events to internal and global causes, data were first analysed using a simple subscores approach from classical test theory. This involved summing counts of internal and global attributions across items (64 positive, 64 negative), and then adjusting sums using the approach described in (Haberman, 2008), which accounts for the existence of measurement error in observations, using the R package `subscore`. The split-half reliability for each subscore was then calculated across N=1000 random splits of the data, to generate average split-half correlation scores, using the R package `multicon`. Internal reliability statistics (Cronbach's α) were also calculated for each score, for comparison with traditional questionnaire-based measures of attributional style (Figure S7a).

Within-individuals, we observed that internal attributions for positive and negative events tended to be moderately negatively correlated (R=-0.26, Figure S7b), supporting the interpretation that individuals may vary in their tendency to express self-protective bias (i.e., heightened internal attribution of positive events, coupled with lower internal attribution of negative events). In the global-specific domain, there was a weak positive correlation between tendency to globally attribute positive and negative events (R=0.16, suggesting a more general preference for global vs specific explanations.

External validation via association with clinical scores (the 2-item version of the PHQ9, the PHQ2 (Kroenke et al., 2003), DAS, and miniSPIN total scores). We observed moderate correlations between tendency to attribute negative events to internal causes and negative self-belief (DAS) and depressed mood (PHQ2) scores (Rs=0.26-0.35). Relationships with internal attributions of positive events and in the global domain tended to be small and weak - although there was some evidence that participants higher in social anxiety (miniSPIN) scores tended to attribute positive events to internal causes less often, and negative events to global causes more often (Figure S7c,d).

The two equivalent 32-item versions of the task were then developed from using 2PL IRT modelling of the full test set. Specifically,

$$P(X = 1 \mid \theta_p, \alpha_i, \beta_i) = \exp^{\alpha_i(\theta_p \beta_i)} / (1 + \exp^{\alpha_i(\theta_p \beta_i)}) \tag{S1}$$

where $P(X = 1)$ represents the probability of choosing a global or internal attribution for each item *i* (modelled separately for positive and negative events), α is the discriminability parameter (governing how well each item differentiates between individuals high or low on the latent trait of interest), β is the difficulty parameter (governing how high an individual must be on the latent trait in order to positively endorse the item), and θ is the participant (*p*)-level latent trait estimate (here, 'globality' and 'internality' for positive and negative events, respectively). Intuitively, the above equation describes a logistic function relating trait estimates to probability of endorsement of each item, with α values governing the slope of the function, and β values its left-right translation.

After fitting the 2PL model to the full item set, items were ranked in terms of their posterior discriminability estimates for trait internality and globality (separately for positive and negative events), and the top-ranked items alternately assigned to form two equivalent test sets, such that they both consisted of 16 negative and 16 positive items. All posterior mean discriminability estimates for included items exceeded 1 (i.e., we had evidence that they meaningfully contributed to the construction of trait estimates).

Finally, simulation-based calibration analysis of the 32-item versions and associated inference model was then carried out, and observed test-retest reliability of task parameters from the two versions was formally assessed (see main text).

Supplementary Figures

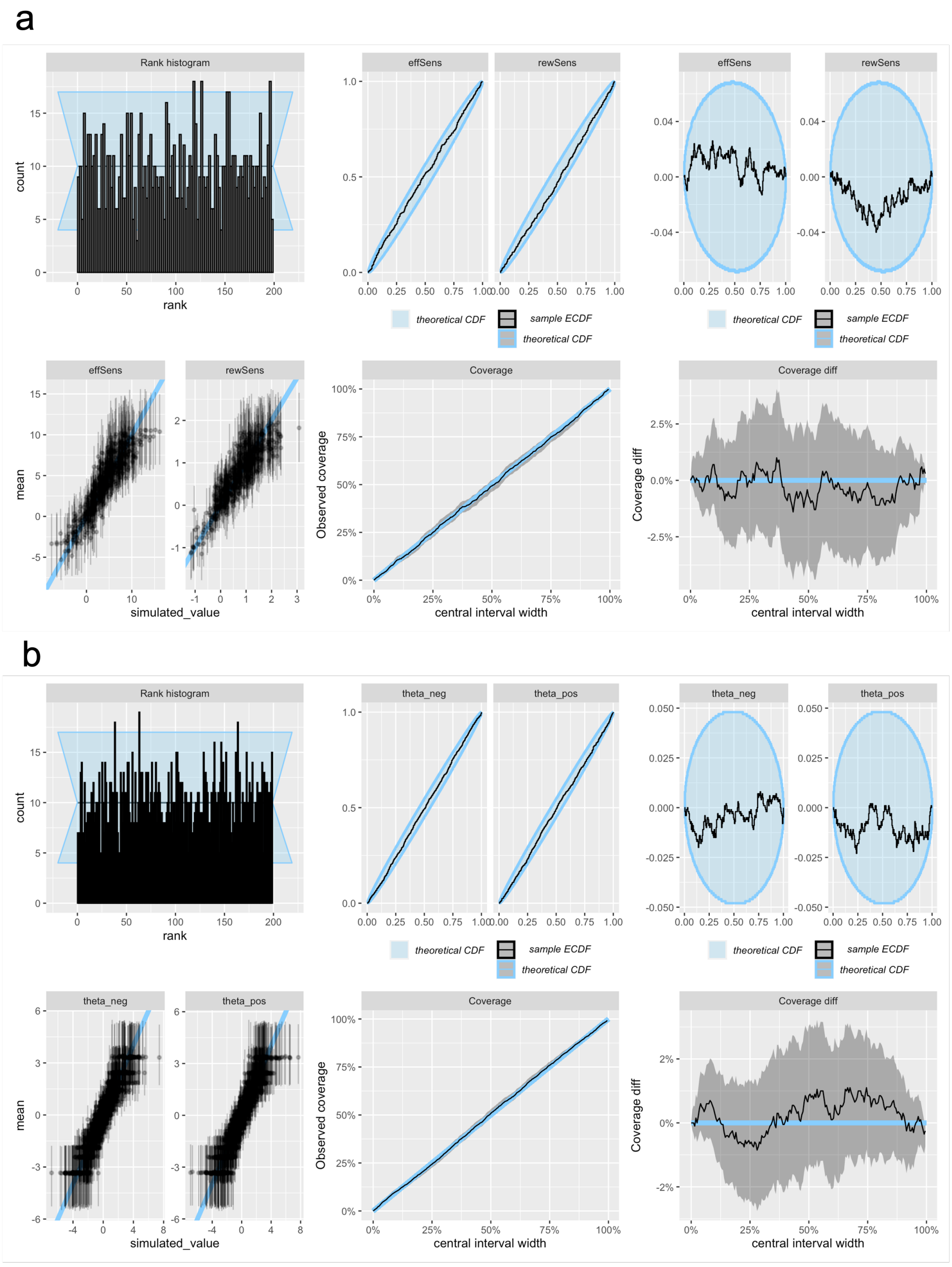


Figure S1: **Graphical summary of simulation-based calibration analysis of data generation and model fit procedures for both tasks. a** Reward-effort decision-making task. *effSens*, effort sensitivity parameter; *rewSens*, reward sensitivity parameter. **b** Causal-attribution task. *theta_neg*, parameter governing latent tendency to make internal/global attributions of negative events; *theta_pos*, parameter governing latent tendency to make internal/global attributions of positive events. Plots in each panel are rank histograms (check for uniformity of posterior draw ranks; horizontal black line=expected average count, blue trapezoid=approximate 95% interval for expected deviations), (E)CDF, (empirical) cumulative distribution functions (blue ellipse=region outlining expected 95% deviations; top-right plots show are rotated by 45° for easier visualisation of deviations), and coverage plots (which show the proportion of true variable values that fall within the 95% posterior credible intervals for each parameter).

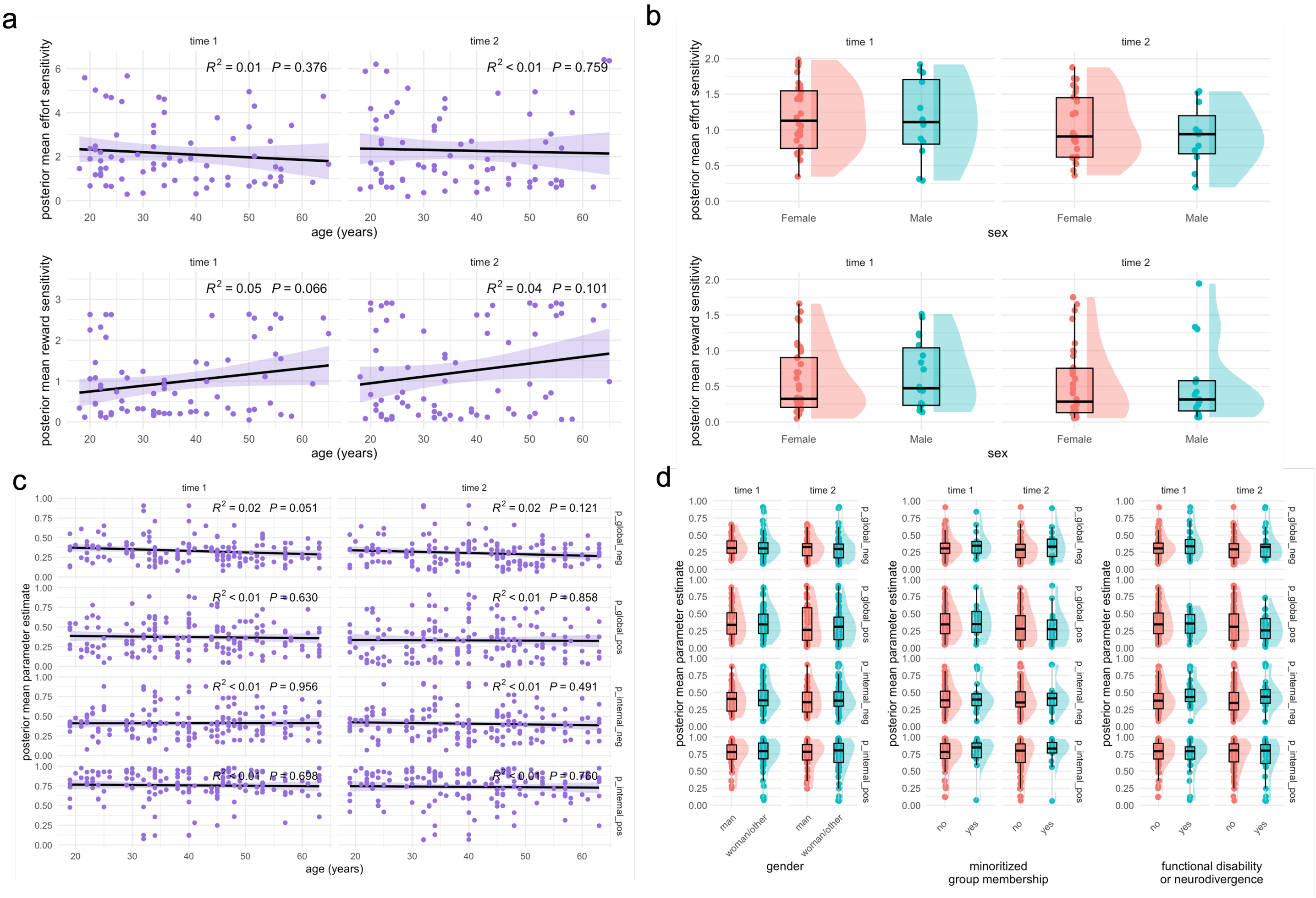


Figure S2: **Parameter estimates according to age and other sociodemographic information, in pilot data samples.** **a** Posterior mean effort and reward sensitivity estimates by age in years, for the reward-effort decision-making task test-retest reliability sample (N=72). **b** Posterior mean effort and reward sensitivity estimates by sex, for the reward-effort decision-making task. **c** Posterior parameter estimates by age, for the causal attribution task test-retest reliability sample (N=88). *p_global_neg/pos*, probability of attributing a negative/positive event to a global cause; *p_internal_neg/pos*, probability of attributing a negative/positive event to an internal cause. **d** Posterior parameter estimates for the causal attribution task by gender identity (man, or woman/non-binary/other), minoritized group status (identifying as belonging to a group that may lead to greater risk of being discriminated against or experiencing prejudice in social or professional situations), and functional disability or neurodivergence (disability or form of neurodivergence that affects ability to concentrate for extended periods of time, perform physically effortful activities, read/write/do maths, deal with people you don't know, or other form or impact on psychosocial functioning).

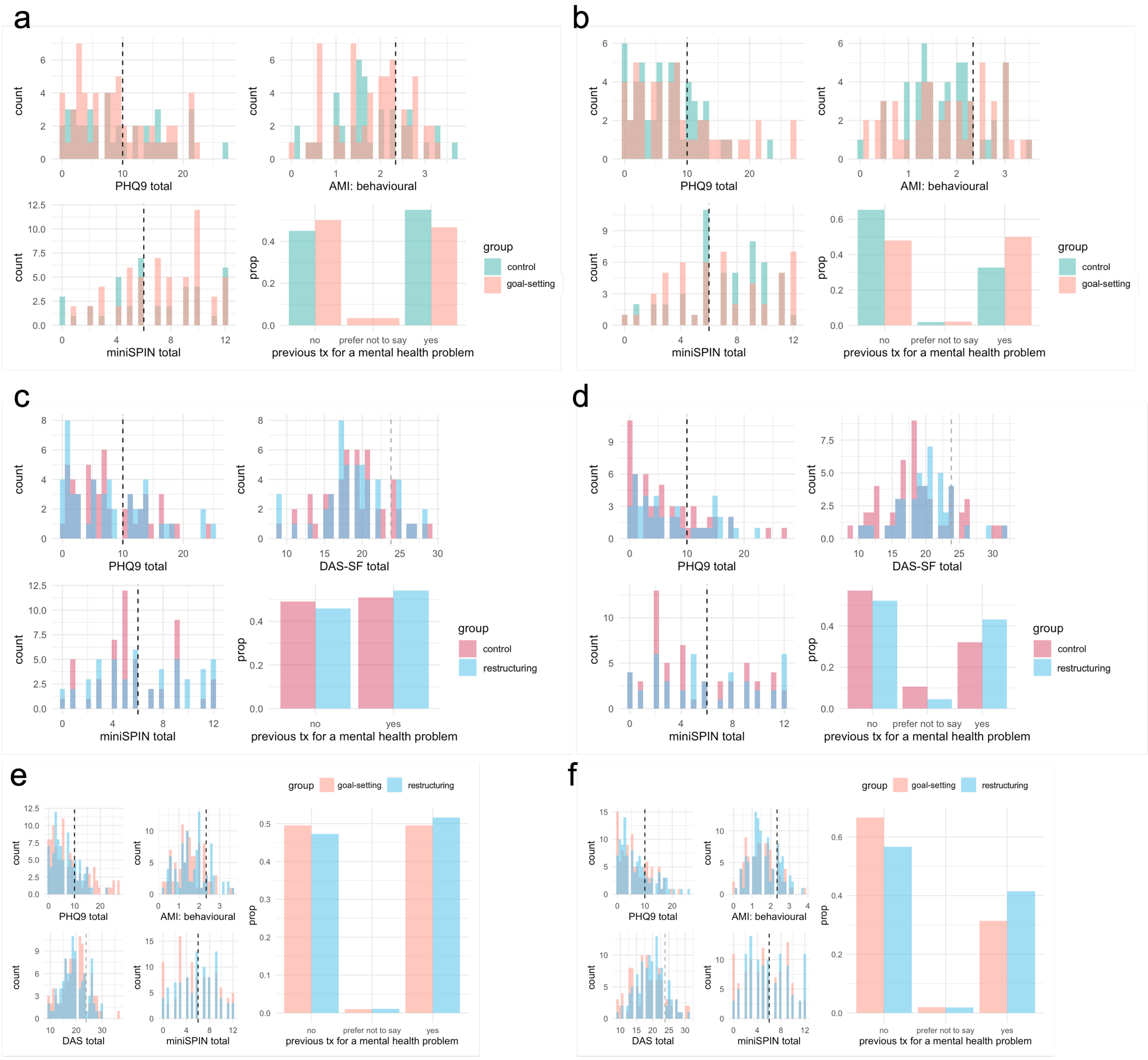


Figure S3: **Self-reported psychological symptom scores for study participants.** **a** Reward-effort decision-making initial discovery sample. **b** Reward-effort decision-making replication sample. **c** Causal attribution task initial discovery sample. **d** Causal attribution task replication sample. **e** Crossover study participants who completed the reward-effort decision-making task. **f** Crossover study participants who completed the causal attribution task. *PHQ9 total*, Physician's Health Questionnaire 9-item measure of depressed mood total score. *AMI: behavioural*, Apathy Motivational Index behavioural amotivation subscale score. *miniSPIN total*, mini Social Phobia Inventory total score. *DAS-SF total*, Dysfunctional Attitude Scale short-form total score. Black dotted lines represent previously-published cut-off scores for clinically-significant levels of symptoms. For the DAS-SF, where no such cut-off score is available, grey dotted lines represent mean scores in previously-published samples of depressed in-patients. Participants were also asked if they had ever previously received treatment (tx) for a mental health problem (see Table S1).

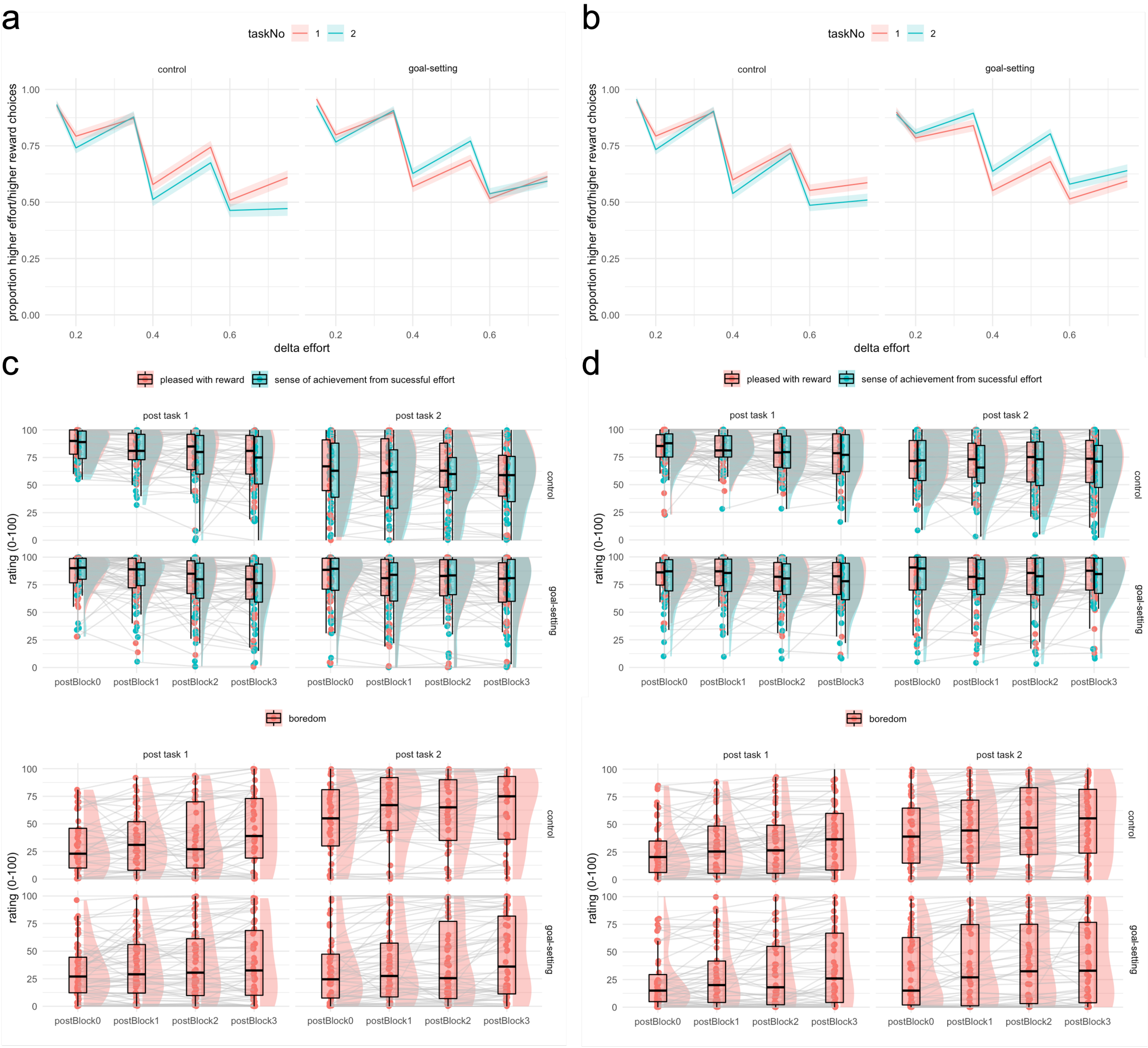


Figure S4: **Choice data and self-report ratings from the reward-effort decision-making task.** **a** Proportionate choice of higher effort/higher reward options by difference in required effort level between choices (delta effort), pre- and post-intervention (taskNo 1, taskNo 2), by intervention condition, in the initial discovery sample. **b** The same data as in (a), in the replication sample. **c** Self-reported ratings data collected after each block of the task, by time and intervention condition. Pleased with reward, "During the task, did you feel PLEASED when you collected the coins?"; sense of achievement from successful effort, "During the task, did you feel A SENSE OF ACHIEVEMENT when you collected the coins?"; boredom, "During the task, did you feel BORED?". **d** The same data as in (c), in the replication sample.

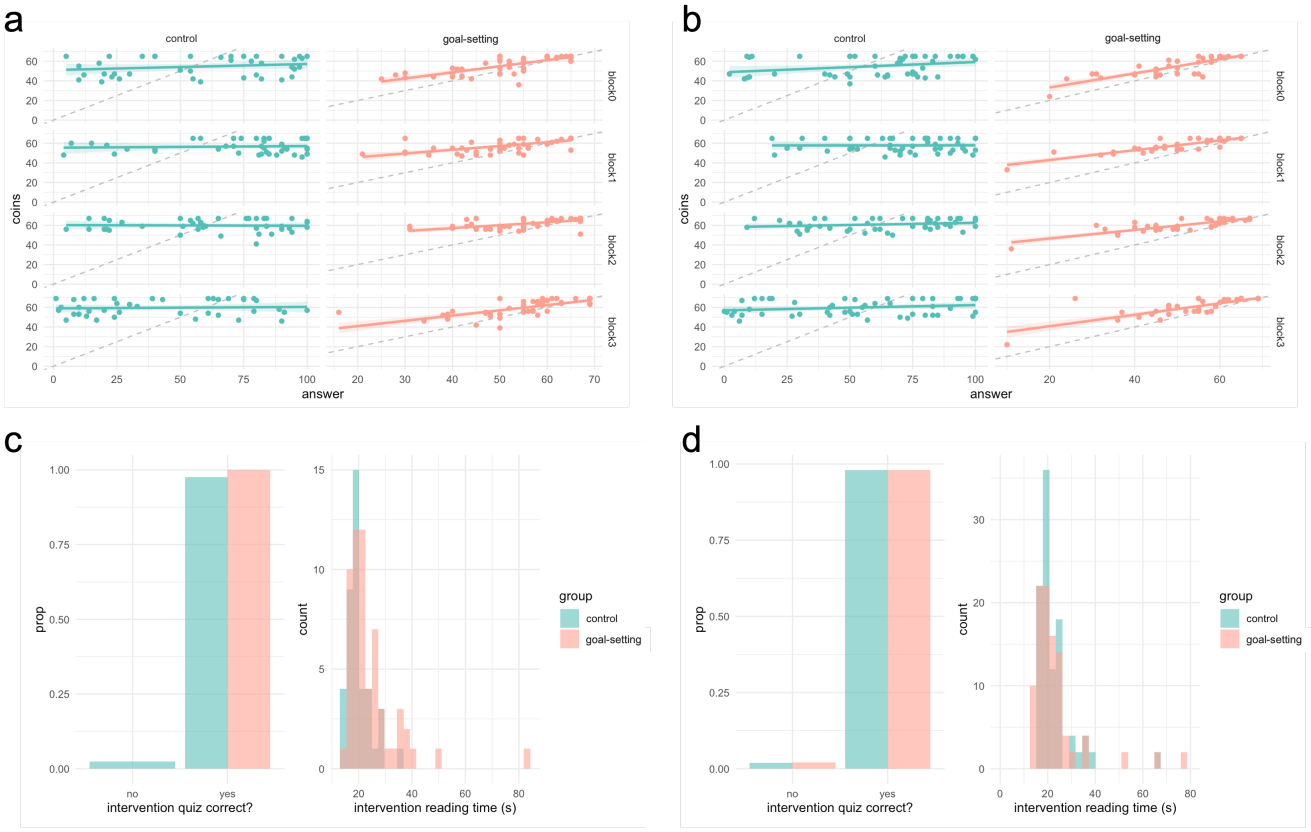


Figure S5: **Goal-setting and intervention reading time data for the reward-effort decision-making task. a** Goals for each block of the task (at time 2) for participants in the initial discovery sample. Participants were invited to set a goal prior to completing each block, given the information that the maximum available reward per block was 69 coins, if they chose the highest effort option every time. For the control condition, answer values represent liking ratings for different types of computer games. Coins represent the actual reward earned by participants in that block. **b** The same information as (a), for the replication sample. **c** *Left*, proportion of participants who provided the correct answer to the multiple-choice comprehension quiz, which followed the intervention text (participants were allowed to return and re-read the text prior to answering). *Right*, time spent reading the intervention text (a single screen of information), in discovery sample participants. **d** The same information as in (c), for the replication sample participants.

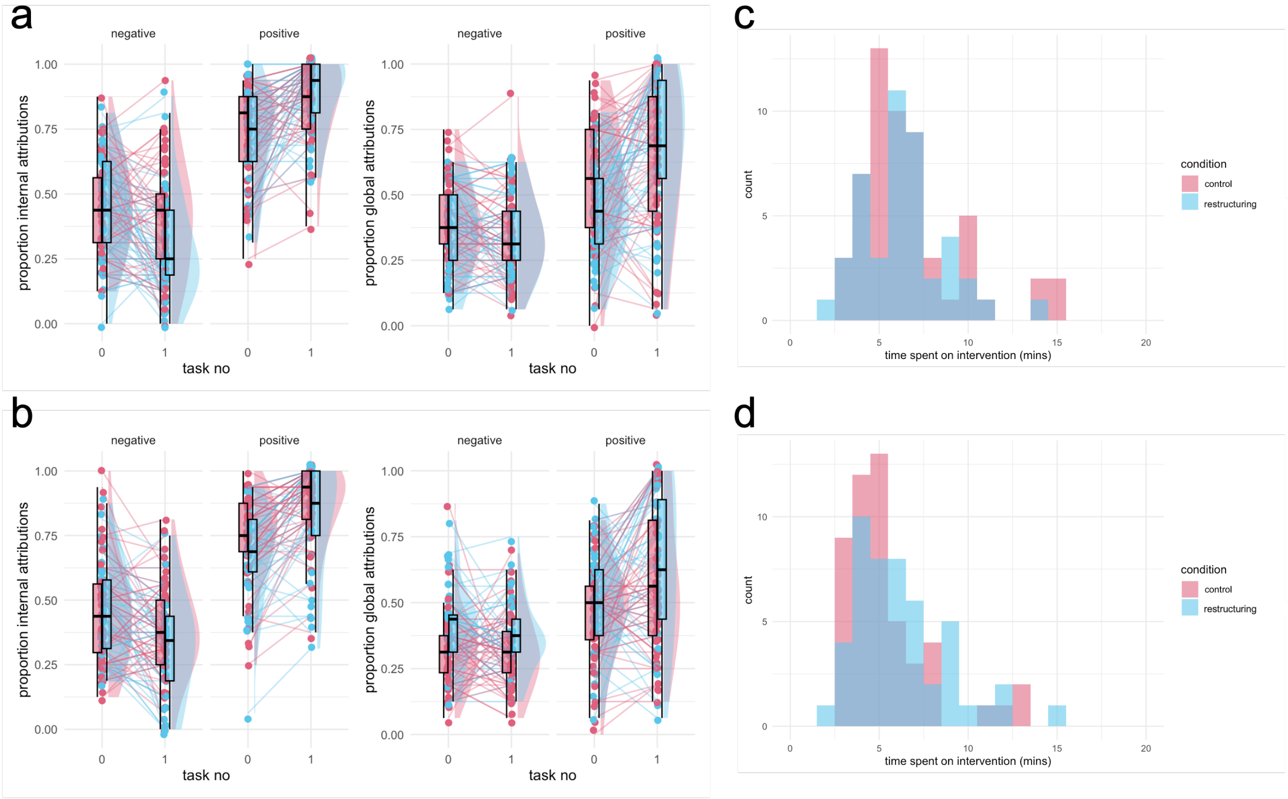
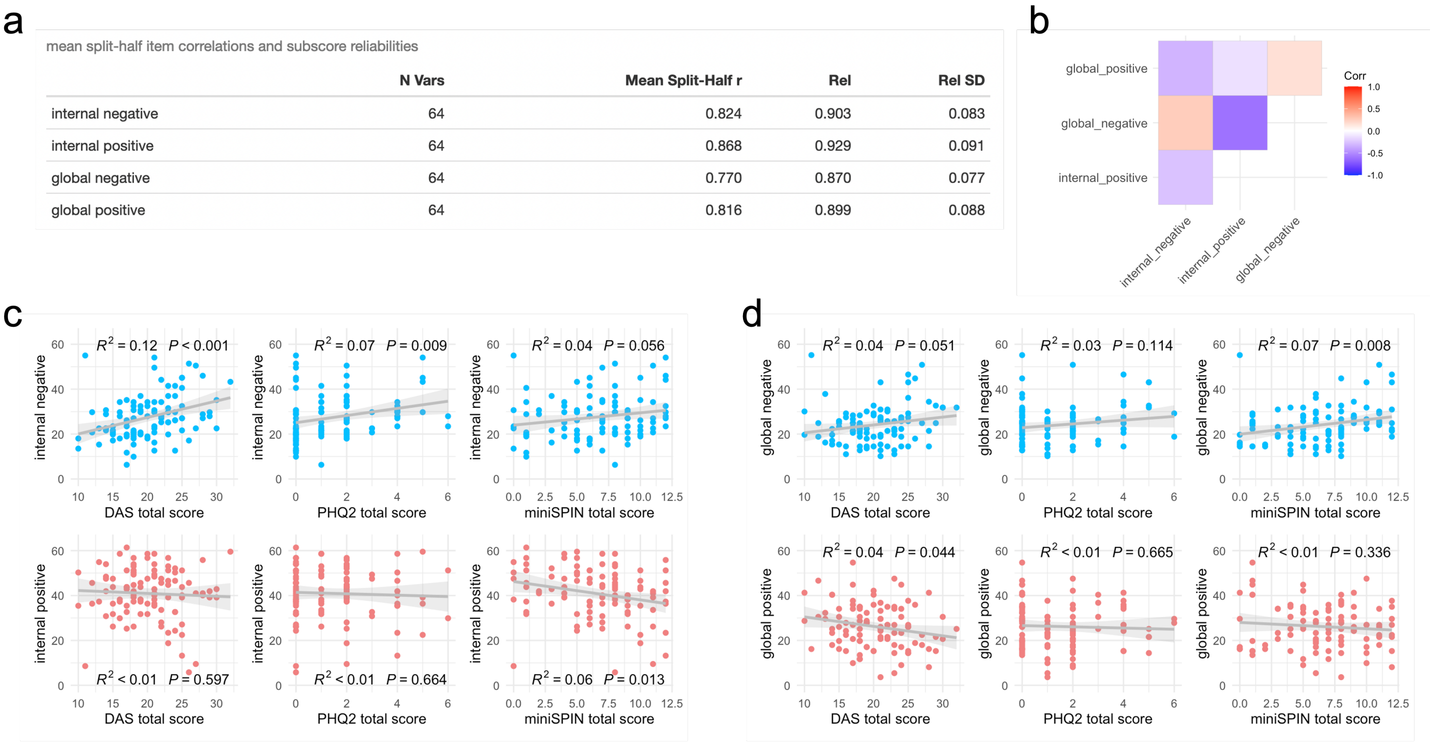


Figure S6: **Attribution choices and intervention time data for the causal attribution task.** **a** Proportionate choice of internal (vs external), and global (vs specific) attributions chosen for positive and negative events pre- and post-intervention (task no 0, task no 1), in the initial discovery sample. **b** The same data as in (a), for the replication sample. **c** Time spent on the intervention in each condition in the initial discovery sample. **d** The same data as in (c), for the replication sample participants.



mean split-half item correlations and subscore reliabilities

	N Vars	Mean Split-Half r	Rel	Rel SD
internal negative	64	0.824	0.903	0.083
internal positive	64	0.868	0.929	0.091
global negative	64	0.770	0.870	0.077
global positive	64	0.816	0.899	0.088

Figure S7: **Details of split-half reliability, within-participant correlation structure, and relationship to clinical scores of simple subscores derived from the full causal attribution task battery, during task development.** **a** Split-half reliability and internal consistency estimates for the full (128) item set in N=102 pilot study participants. *Mean Split-Half r*, the average of all estimated split-half correlations; *Rel*, the average of all split-half reliabilities (equivalent to Cronbach's alpha); *Rel SD*, the standard deviation of all split-half reliabilities. **b** Correlation matrix for within-participant variation in subscore estimates. **c** Bivariate relationships between negative and positive internality subscores, and self-reported clinical symptoms. d Bivariate relationships between negative and positive globality subscores, and self-reported clinical symptoms. *DAS*, Dysfunctional Attitudes Scale (short form). *PHQ2*, Patient Health Questionnaire 2-item measure of depressed mood, *miniSPIN*, 3-item mini Social Phobia Inventory.

SUPPLEMENTARY TABLES

		Reward-effort sample 1	**Reward-effort sample 2**	**Causal attribution sample 1**	**Causal attribution sample 2**	**Crossover study sample 1**	**Crossover study sample 2**
	N	100	102	100	100	197	208
Age (years)	mean (SD)	35.3 (11)	40.1 (11.6)	36 (9.5)	38.5 (11.4)	37.4 (12.8)	38.7 (12.3)
	range	19-60	18-64	19-60	19-63	18-65	18-65
Gender	Woman	77 (79%)	61 (60%)	66 (66%)	44 (44%)	99 (52%)	112 (54%)
	Man	20 (20%)	40 (39%)	30 (30%)	56 (56%)	92 (48%)	94 (45%)
	Non-binary or other	1 (1%)	1 (1%)	4 (4%)	0 (0%)	1 (1%)	2 (1%)
Race/ ethnicity	White	80 (82%)	87 (86%)	80 (80%)	85 (85%)	164 (85%)	165 (80%)
	Asian	5 (5%)	6 (6%)	7 (7%)	7 (7%)	15 (8%)	16 (8%)
	Black	1 (1%)	4 (4%)	3 (3%)	2 (2%)	2 (1%)	11 (5%)
	Mixed	7 (7%)	3 (3%)	4 (4%)	4 (4%)	7 (4%)	9 (4%)
	Other	5 (5%)	1 (1%)	6 (6%)	2 (2%)	4 (2%)	7 (3 %)
Employment status	Employed	64 (65%)	69 (68%)	79 (80%)	68 (68%)	142 (74%)	156 (75%)
	Unemployed	13 (13%)	11 (11%)	9 (9%)	10 (10%)	19 (10%)	18 (9%)
	Not seeking	21 (21%)	22 (22%)	11 (11%)	22 (22%)	31 (16%)	34 (16%)
Financial status	Doing okay	52 (53%)	39 (38%)	46 (46%)	49 (49%)	91 (47%)	111 (53%)
	Just about getting by	31 (32%)	43 (42%)	38 (38%)	36 (36%)	70 (37%)	74 (36%)
	Struggling	15 (15%)	20 (20%)	15 (15%)	15 (15%)	31 (16%)	23 (11%)
Housing status	Homeowner	44 (45%)	47 (46%)	42 (42%)	48 (48%)	85 (44%)	94 (45%)
	Tenant	29 (30%)	39 (38%)	48 (48%)	38 (38%)	71 (37%)	77 (37%)
	Other	25 (26%)	16 (16%)	9 (9%)	14 (14%)	36 (19%)	37 (18%)
Neurodivergence	Yes	19 (19%)	18 (18%)	15 (15%)	10 (10%)	37 (19%)	31 (15%)
	No	72 (73%)	80 (78%)	80 (81%)	87 (87%)	146 (76%)	172 (83%)
	Prefer not to say	7 (7%)	4 (4%)	5 (5%)	3 (3%)	9 (5%)	5 (2%)
Previous treatment for a mental health problem	Yes	49 (50%)	42 (41%)	52 (53%)	37 (37%)	97 (51%)	76 (37%)
	No	47 (48%)	58 (57%)	48 (48%)	55 (55%)	93 (48%)	128 (62%)
	Prefer not to say	2 (2%)	2 (2%)	0 (0%)	8 (8%)	1 (1%)	4 (2%)
If yes, type of treatment (all that apply)	Talking therapy	38 (39%)	31 (30%)	36 (36%)	26 (26%)	72 (38%)	57 (27%)
	Medication	35 (36%)	29 (28%)	35 (35%)	27 (27%)	70 (37%)	51 (25%)
	Self-guided	23 (23%)	20 (20%)	21 (21%)	18 (18%)	38 (20%)	32 (15%)
	Other	7 (7%)	2 (2%)	4 (4%)	1 (1%)	9 (5%)	6 (3%)
PHQ9 total	mean (SD)	9.1 (6.7)	7.6 (6)	7.7 (6.1)	6.9 (6.3)	7.4 (6.1)	6.7 (5.6)
AMI behaviour	mean (SD)	1.8 (0.8)	1.8 (0.8)	-	-	1.6 (0.7)	1.6 (0.8)
DAS-SF total	mean (SD)	-	-	19.1 (4.5)	19.3 (4.8)	19.4 (5)	19.1 (4.8)
miniSPIN total	mean (SD)	7.2 (3.3)	7 (3.2)	6.1 (3.3)	5.4 (3.8)	5.7 (3.3)	5.6 (3.5)

Table S1: **Self-reported demographic and clinical data for all study participants.** For reward-effort decision-making and causal attribution studies, samples 1 and 2 represent the initial discovery and replication samples, respectively. For the crossover study, sample 1 represents individuals who were randomized to the reward-effort decision-making task, and sample 2 represents individuals who were randomized to the causal attribution task. Response categories for employment, financial, and housing status were based on those described in (Buckman et al. 2022). Employment status categories were employed (including full-time and part-time employment), unemployed (job seekers and those unemployed owing to ill health), and not seeking employment (stay-at-home parents, students, and retirees). Housing status categories were homeowner (including those with a mortgage), tenant, and other (living with family or friends, homeless, or living in a hostel). Neurodivergence was defined as “a term for when someone processes or learns information in a different way to that which is considered ’typical’: common examples include autism and ADHD”. Categories for previous mental health treatment were talking therapy (including cognitive-behavioural therapies), medication, self-guided (e.g., workbooks or apps), or other. PHQ9 total, Physician’s Health Questionnaire 9-item measure of depressed mood total score (possible range 0-27). AMI: behavioural, Apathy Motivational Index behavioural amotivation subscale score (possible range 0-4, mean score across 6 items). miniSPIN total, mini Social Phobia Inventory total score (possible range 0-12). DAS-SF total, Dysfunctional Attitude Scale short-form total score (possible range 9-36). -, questionnaire not administered in this sample.

	mean	**se (mean)**	**sd**	**5%**	**95%**	N_{eff}	$\hat{R}$
Initial discovery sample							
Mean effort sensitivity at time 1	-0.9986	0.0052	0.1613	-1.2775	-0.7487	980	1.0039
Mean effort sensitivity at time 2	-0.9577	0.0048	0.1878	-1.2807	-0.6716	1517	1.0003
Mean reward sensitivity at time 1	0.0160	0.0069	0.1743	-0.2651	0.3002	640	1.0071
Mean reward sensitivity at time 2	-0.2036	0.0160	0.4663	-0.9848	0.5750	844	1.0025
Effect of goal-setting on reward sensitivity at time 2	0.2745	0.0149	0.5042	-0.5342	1.0980	1139	1.0035
Effect of goal-setting on effort sensitivity at time 2	-0.5653	0.0051	0.2114	-0.9132	-0.2260	1693	1.0069
Replication sample							
Mean effort sensitivity at time 1	-1.1009	0.0052	0.1667	-1.3882	-0.8397	1014	1.0053
Mean effort sensitivity at time 2	-1.1674	0.0052	0.1824	-1.4900	-0.8899	1221	1.0023
Mean reward sensitivity at time 1	-0.0587	0.0066	0.1523	-0.3075	0.2044	533	1.0043
Mean reward sensitivity at time 2	-0.0062	0.0124	0.3233	-0.5253	0.5475	681	1.0021
Effect of goal-setting on reward sensitivity at time 2	0.5535	0.0087	0.3414	0.0163	1.1160	1540	1.0008
Effect of goal-setting on effort sensitivity at time 2	-0.3192	0.0044	0.1972	-0.6519	-0.0044	2043	0.9993

Table S2: **Hierarchical Bayesian model results for effect of goal-setting on reward-effort decision-making.** Mean, posterior mean; se (mean), standard error of the posterior mean. 5%, 95%, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size (an estimate of the number of independent draws from the posterior distribution of the estimand of interest); $\hat{R}$, the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains (if all chains are at equilibrium, $\hat{R}$ will be 1).

	mean	**se (mean)**	**sd**	**5%**	**95%**	N_{eff}	$\hat{R}$
Initial discovery sample							
Mean θ for internal attributions of negative events at time 1	-0.2243	0.0017	0.0811	-0.3574	-0.0880	2351	1.0006
Mean θ for internal attributions of negative events at time 2	-0.4831	0.0028	0.1359	-0.7096	-0.2630	2288	1.0012
Mean θ for internal attributions of positive events at time 1	1.1002	0.0018	0.0977	0.9430	1.2627	2893	1.0015
Mean θ for internal attributions of positive events at time 2	2.2414	0.0054	0.2249	1.8854	2.6313	1762	1.0026
Mean θ for global attributions of negative events at time 1	-0.5098	0.0012	0.0694	-0.6267	-0.3965	3498	0.9994
Mean θ for global attributions of negative events at time 2	-0.6684	0.0018	0.1001	-0.8410	-0.5070	3170	0.9994
Mean θ for global attributions of positive events at time 1	-0.0141	0.0020	0.0953	-0.1745	0.1414	2288	1.0003
Mean θ for global attributions of positive events at time 2	0.7039	0.0054	0.2177	0.3506	1.0640	1650	1.0027
Effect of restructuring on θ internal-negative at time 2	-0.5632	0.0035	0.1905	-0.8725	-0.2407	2956	1.0006
Effect of restructuring on θ internal-positive at time 2	0.3848	0.0071	0.3037	-0.1143	0.8799	1836	1.0014
Effect of restructuring on θ global-negative at time 2	0.0093	0.0023	0.1364	-0.2130	0.2369	3555	0.9996
Effect of restructuring on θ global-positive at time 2	0.4723	0.0076	0.3028	-0.0352	0.9611	1572	1.0024
Replication sample							
Mean θ for internal attributions of negative events at time 1	-0.1987	0.0019	0.0854	-0.3399	-0.0562	2015	1.0024
Mean θ for internal attributions of negative events at time 2	-0.5091	0.0023	0.1139	-0.6922	-0.3218	2405	1.0012
Mean θ for internal attributions of positive events at time 1	0.9753	0.0022	0.0974	0.8163	1.1359	2041	1.0022
Mean θ for internal attributions of positive events at time 2	2.2854	0.0046	0.2184	1.9407	2.6427	2230	1.0005
Mean θ for global attributions of negative events at time 1	-0.6103	0.0014	0.0715	-0.7252	-0.4951	2675	1.0011
Mean θ for global attributions of negative events at time 2	-0.7800	0.0014	0.0833	-0.9193	-0.6440	3585	1.0001
Mean θ for global attributions of positive events at time 1	-0.0987	0.0017	0.0803	-0.2341	0.0339	2106	1.0010
Mean θ for global attributions of positive events at time 2	0.4347	0.0046	0.1995	0.1084	0.7621	1888	1.0018
Effect of restructuring on θ internal-negative at time 2	-0.3367	0.0034	0.1707	-0.6113	-0.0547	2501	1.0013
Effect of restructuring on θ internal-positive at time 2	0.1488	0.0064	0.3008	-0.3397	0.6592	2180	0.9996
Effect of restructuring on θ global-negative at time 2	0.2084	0.0022	0.1233	0.0064	0.4103	3271	1.0003
Effect of restructuring on θ global-positive at time 2	0.4352	0.0061	0.2800	-0.0293	0.8978	2100	1.0015

Table S3: **Hierarchical Bayesian model results for effect of cognitive restructuring on causal attribution.** *Continued on next page.*

Table S3: θ, parameters describing latent tendency to attribute events to different kinds of causes. Mean, posterior mean; se (mean), standard error of the posterior mean. 5%, 95%, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size; $\hat{R}$, the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains).

	mean	**se (mean)**	**sd**	**5%**	**95%**	N_{eff}	$\hat{R}$
Reward-effort decision-making task sample							
Mean effort sensitivity at time 1	-1.1340	0.0027	0.0996	-1.3037	-0.9842	1377	1.0014
Mean effort sensitivity at time 2	-1.0799	0.0033	0.1282	-1.2949	-0.8817	1537	1.0006
Mean reward sensitivity at time 1	-0.0451	0.0050	0.1189	-0.2339	0.1534	555	1.0080
Mean reward sensitivity at time 2	0.3784	0.0091	0.2631	-0.0321	0.8264	838	1.0047
Effect of goal-setting on reward sensitivity at time 2	-0.4231	0.0076	0.2704	-0.8800	0.0151	1266	1.0016
Effect of goal-setting on effort sensitivity at time 2	-0.4108	0.0036	0.1528	-0.6628	-0.1628	1849	1.0018
Causal attribution task sample							
Mean θ for internal attributions of negative events at time 1	-0.2361	0.0013	0.0540	-0.3245	-0.1475	1633	1.0025
Mean θ for internal attributions of negative events at time 2	-0.0111	0.0017	0.0750	-0.1358	0.1109	1970	1.0006
Mean θ for internal attributions of positive events at time 1	0.9401	0.0015	0.0595	0.8439	1.0371	1631	1.0014
Mean θ for internal attributions of positive events at time 2	0.9665	0.0028	0.1094	0.7857	1.1499	1476	1.0028
Mean θ for global attributions of negative events at time 1	-0.4827	0.0011	0.0523	-0.5669	-0.3968	2373	0.9999
Mean θ for global attributions of negative events at time 2	-0.4207	0.0017	0.0768	-0.5460	-0.2992	2135	1.0001
Mean θ for global attributions of positive events at time 1	-0.1165	0.0012	0.0536	-0.2032	-0.0277	2136	1.0005
Mean θ for global attributions of positive events at time 2	-0.3409	0.0020	0.0877	-0.4858	-0.1981	1888	0.9992
Effect of restructuring on θ internal-negative at time 2	-0.2756	0.0019	0.0972	-0.4326	-0.1135	2723	1.0006
Effect of restructuring on θ internal-positive at time 2	0.4595	0.0029	0.1435	0.2229	0.6963	2531	1.0008
Effect of restructuring on θ global-negative at time 2	-0.1457	0.0019	0.1000	-0.3082	0.0210	2762	1.0016
Effect of restructuring on θ global-positive at time 2	0.0650	0.0023	0.1149	-0.1266	0.2542	2519	0.9995

Table S4: **Hierarchical Bayesian model results for effects of goal-setting vs restructuring on reward-effort decision-making and causal attribution in the crossover study.** *Continued on next page.*

Table S4: θ, parameters describing latent tendency to attribute events to different kinds of causes. Mean, posterior mean; se (mean), standard error of the posterior mean. 5%, 95%, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size; $\hat{R}$, the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains.

	mean	**se (mean)**	**sd**	**5%**	**95%**	N_{eff}	$\hat{R}$
Reward-effort decision-making initial + replication samples							
Effect of goal-setting on reward sensitivity	0.4067	0.0092	0.2988	-0.0768	0.9058	1054	1.0000
Effect of goal-setting on effort sensitivity	-0.3563	0.0038	0.1336	-0.5769	-0.1384	1227	1.0014
βa, baseline effort sensitivity	0.2443	0.0042	0.1180	0.0561	0.4396	804	1.0093
βa, effect of goal-setting on effort sensitivity	-0.3655	0.0042	0.1380	-0.5971	-0.1540	1091	1.0040
Crossover study reward-effort decision-making sample							
Effect of goal-setting on reward sensitivity	-0.4498	0.0083	0.2896	-0.9331	0.0218	1232	1.0012
Effect of goal-setting on effort sensitivity	-0.3652	0.0070	0.2514	-0.7583	0.0414	1302	1.0000
βa, baseline effort sensitivity	0.1642	0.0063	0.1931	-0.1165	0.5088	953	1.0033
βn, baseline effort sensitivity	-0.1455	0.0060	0.1703	-0.4428	0.1053	801	1.0032
βa, effect of goal-setting on effort sensitivity	-0.1105	0.0088	0.3031	-0.6200	0.3687	1179	1.0016
βn, effect of goal-setting on effort sensitivity	-0.0649	0.0075	0.2428	-0.4592	0.3222	1051	1.0071
Causal attribution task initial + replication samples							
Effect of restructuring on θ internal-negative attributions	-0.4775	0.0030	0.1386	-0.7037	-0.2458	2098	1.0011
Effect of restructuring on θ internal-positive attributions	0.3615	0.0071	0.2667	-0.0831	0.8057	1431	0.9999
Effect of restructuring on θ global-negative attributions	0.0701	0.0016	0.0920	-0.0788	0.2204	3290	0.9994
Effect of restructuring on θ global-positive attributions	0.4975	0.0057	0.2415	0.1015	0.8886	1819	0.9993
βn, baseline θ internal-positive	-0.1610	0.0023	0.0780	-0.2922	-0.0352	1160	1.0023
βn, effect of restructuring on θ internal-positive	-0.2383	0.0051	0.1905	-0.5541	0.0741	1414	1.0021
Crossover study causal attribution sample							
Effect of restructuring on θ internal-negative attributions	-0.3085	0.0020	0.1015	-0.4761	-0.1464	2705	0.9997
Effect of restructuring on θ internal-positive attributions	0.6469	0.0091	0.3081	0.1667	1.1676	1143	1.0039
Effect of restructuring on θ global-negative attributions	-0.1575	0.0021	0.1059	-0.3295	0.0175	2458	1.0014
Effect of restructuring on θ global-positive attributions	0.0665	0.0028	0.1236	-0.1402	0.2674	1950	1.0007
βa, baseline θ internal-positive	-0.1788	0.0043	0.1406	-0.4306	0.0148	1066	1.0004
βn, baseline θ internal-positive	-0.1968	0.0048	0.1435	-0.4554	0.0123	910	1.0021
βa, effect of restructuring on θ internal-positive	-0.4500	0.0084	0.2902	-0.9711	-0.0382	1195	1.0014
βn, effect of restructuring on θ internal-positive	0.0787	0.0081	0.2857	-0.3659	0.5529	1241	1.0026

Table S5: **Hierarchical Bayesian model results for models taking into account individual differences in self-reported behavioural amotivation and negative cognition.** *Continued on next page.*

Table S5: βa, posterior weight for influence of behavioural amotivation on baseline or intervention-induced change in parameter estimates; βn, posterior weight for influence of negative cognition on baseline or intervention-induced change in parameter estimates; θ, parameters describing latent tendency to attribute events to different kinds of causes. Mean, posterior mean; se (mean), standard error of the posterior mean. 5%, 95%, posterior probability quantiles for parameter estimates; N_{eff}, effective sample size; $\hat{R}$, the ratio of the average variance of draws within each chain to the variance of the pooled draws across chains.