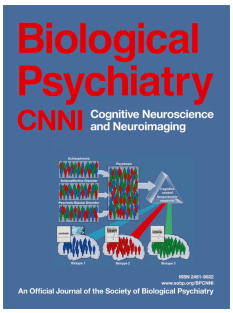


Learning training as a cognitive restructuring intervention

Agnes Norbury, Quentin Dercon, Tobias U. Hauser, Raymond J. Dolan, Quentin J.M. Huys

PII: S2451-9022(25)00136-3

DOI: https://doi.org/10.1016/j.bpsc.2025.04.008

Reference: BPSC 1421

To appear in: *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*

Received Date: 30 April 2024

Revised Date: 14 April 2025

Accepted Date: 14 April 2025

Please cite this article as: Norbury A., Dercon Q., Hauser T.U., Dolan R.J. & Huys Q.J.M., Learning training as a cognitive restructuring intervention, *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* (2025), doi: https://doi.org/10.1016/j.bpsc.2025.04.008.

Learning training as a cognitive restructuring intervention

Short Title: Learning training as cognitive restructuring

Agnes Norbury[1], Quentin Dercon[1†], Tobias U. Hauser[2,3,4,5], 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, Medical School and University Hospital, Eberhard Karls University of Tubingen, Germany
5 German Center for Mental Health (DZPG)

†Corresponding Author: quentin.dercon.22@ucl.ac.uk

Abstract

Background. A core part of cognitive therapy for low mood is learning to identify and challenge negative beliefs. However, it is currently unclear whether improved ability to recognise such beliefs, and the biased interpretations of events which may maintain them, is a mechanism of symptom change during treatment.

Methods. We investigated the effects of completing a learning task (training to identify and select self-enhancing interpretations of events) and a brief cognitive restructuring intervention (how exploring alternative explanations of events may result in improved mood) on causal attribution tendencies. Studies were conducted online using randomized-controlled experimental designs (*N*=200 & *N*=164), and data were analysed using hierarchical Bayesian models.

Results. We found that both learning training and the restructuring intervention decreased tendencies to make unhelpful attributions and increased tendencies to make self-enhancing attributions. Across two studies, changes in attribution tendencies were associated with higher learning rates during learning training, an effect specific to learning about different kinds of event attribution. Contrary to expectation, we found no evidence that faster learning was associated specifically to changes in attribution tendencies following cognitive restructuring. Since participants with higher learning rate estimates also provided explicit ratings and free-text descriptions of event causes which were closer to the ground truth, we interpret this as representing a greater benefit of learning training in individuals who were better able to understand the task state space.

Conclusions. We suggest that personalized training, in conjunction with feedback based on interpretable computational model output, may provide a useful form of augmentation or learning support tool during therapy.

INTRODUCTION

A core aspect of cognitive therapy for low mood is learning to identify negative beliefs and exploring alternative explanations for events which challenge these beliefs ('cognitive restructuring') (*1*, *2*). However, there is currently little definitive evidence as to whether learning to identify negative beliefs and application of restructuring skills are key drivers of symptom change during psychological therapy for low mood (*3*, *4*). Demonstrating this using data from traditional randomized-controlled trials involving psychotherapy treatment programs (e.g., cognitive-behavioural therapy (CBT)) is challenging, given the multiple types of interventions delivered in each program coupled with a lack of the fine-grained resolution needed to infer temporal dependencies between changes in beliefs and symptoms (*3*, *5*). There is some evidence to suggest that greater self-reported frequency and/or skill in applying cognitive strategies is associated with greater overall symptom reduction following C(B)T (*6*–*11*). That said, the degree of conceptual overlap between self-report measures of cognitive skills and symptoms themselves (the 'jangle' fallacy) makes disentangling changes in the former from overall treatment response or residual symptom burden considerably more difficult (*6*, *12*).

Behavioural measures of cognitive processes may be one way to help solve this problem, since they are less close to the target construct of interest: symptom change (*13*–*15*). Combining cognitive-behavioural measures with randomized allocation of therapy-like interventions in high-throughput testing can provide an efficient way to test whether specific components of psychological treatments may causally impact specific cognitive processes, prior to extending testing to resource-intensive clinical settings (*16*, *17*). Here, we use this approach to test whether a behavioural measure of attribution tendencies (how people tend to reason about the causes underlying events) is affected by (a) training in learning to identify different kinds of causal attributions (a learning task intervention) and (b) practice in identifying and challenging unhelpful attributions of events in their own lives (a brief cognitive restructuring intervention). Cognitive therapy can be considered a process of learning (*13*), and it has been suggested that individuals with greater capacity for learning during treatment show greater benefits (*18*). On this basis, we initially hypothesized that individual differences in learning task performance would be related to individual differences in response to a brief cognitive restructuring intervention.

Instead, across two studies, we found evidence that both the learning task training and the brief cognitive restructuring intervention affected causal attribution tendencies, shifting them away from unhelpful or 'depressogenic' patterns (e.g., lower tendency to attribute negative events to self-related or internal causes) and towards self-enhancing styles (e.g., higher tendency to attribute positive events to internal causes). In both studies, greater shifts in attribution tendencies were associated with higher learning rate estimates on the learning training task. Since we found no association between attribution change and learning rates from a matched control task (which did not concern causal attributions), we interpret this as being due to greater ability to discriminate between different kinds of attributions, or a better understanding of the learning task state space. Contrary to expectations, there was no evidence that individuals with faster learning rates showed greater responses to the cognitive restructuring intervention specifically. We discuss these findings with reference to recent proposals for augmenting psychological treatments with strategies aimed at boosting learning and memory of treatment content, and for whom this might be most effective for (*19*, *20*).

RESULTS

We report results of two cross-sectional studies with similar overall designs (Figure 1). In both studies, participants completed a task-based measure of causal attribution tendencies, before and after two types of intervention: a learning training (or control learning) task, and a brief cognitive restructuring (or control) intervention.

PARTICIPANTS

Participants for both studies were recruited from an online research participation platform (Prolific (*21*)) and are described in Table 1. In both studies, samples showed evidence of self-selection for mental health research, given 40% reporting of previous treatment for a mental health problem, and mild-to-moderate average endorsement of current low mood and social anxiety symptoms (proportion of participants above cut-off score for clinically-significant depressed mood according to the 9-item patient health questionnaire (PHQ-9)=32% & 27%; proportion of participants with significant social anxiety according to the 3-item social phobia inventory (miniSPIN)=48% & 46%; Figure S1).

SEPARATE EFFECTS OF LEARNING TRAINING AND BRIEF COGNITIVE RESTRUCTURING ON CAUSAL ATTRIBUTION TENDENCIES

We first examined whether there was evidence for separate effects of completing the learning training task and brief cognitive restructuring intervention on attribution tendencies, as measured on the causal attribution task. Specifically, we used a hierarchical Bayesian modelling approach to test whether there was evidence for additional group-level effects of having been randomized to learning training *vs.* control learning task conditions, and cognitive restructuring *vs.* control intervention conditions (see Methods).

In study 1 all participants completed the learning training task, so here we were only able to examine group-level effects of cognitive restructuring *vs.* control intervention conditions. As reported previously, we found that completion of the brief cognitive restructuring intervention resulted in decreased tendency to attribute negative events to internal causes (posterior estimate=-0.48 [90% credible interval (CrI)=(-0.70, -0.26)]), and an increased tendency to attribute positive events to general or global causes (posterior estimate=0.50 [90% CrI=(0.11, 0.90)]) (Figure 2*A*-*B*; Table S1). Of interest, the group means for each parameter showed some evidence of shifts between time-points, with participants showing slightly higher mean endorsement of internal and global attributions of positive events at the second measurement (Figure 2). These group-level shifts could represent common effects of completing the cognitive restructuring and control interventions on attribution tendency. However, as the control intervention made no reference to how interpretations of events might affect mood, or reappraisal strategies, this is unlikely. An alternative explanation is that these effects are due to completion of the learning training task by all study participants, since this directly involves learning to recognise different kinds of attributions.

We tested this idea directly in study 2. Importantly, this study included a control learning task, as well as cognitive restructuring and control intervention conditions. To formally test whether completion of the learning task resulted in group-level changes in attribution tendencies, we augmented the analysis model for these data such that post-intervention (time 2) attribution tendencies could be influenced by learning training condition, as well as restructuring intervention condition (see Methods).

Model comparison revealed that the model with additional effects for learning task condition had marginally better predictive accuracy for causal attribution task data than the model with restructuring intervention condition alone (difference in expected log pointwise predictive density (ELPD) for left-out causal attribution task data, $ELPD_{diff}$=-0.4, but of less than 5x than the standard error (SE) of the estimate: SE_{diff}=6.8), suggesting that this indeed had an additional impact on changes in attribution tendencies.

Inspection of changes in individual parameter estimates between time 1 (pre-intervention) and time 2 (post-intervention) revealed that participants who completed both the learning training task and cognitive restructuring intervention showed the greatest shifts away from depressogenic (internal, global) attributions of negative events, and towards self-enhancing attributions of positive events (Figure 2*C*). Posterior parameter estimates for group-level effects revealed that, when accounting for learning task condition, the restructuring intervention both decreased tendency to attribute negative events to internal causes (posterior estimate=-0.32 [90% CrI=(-0.57, -0.06)]), and increased tendency to attribute positive events to internal causes (posterior estimate=0.65 [90% CrI=(0.19, 1.11)]) (Figure 2*D*, Table S2). There was also evidence for separate group-level effects of completion of the learning training vs control learning task on attribution tendencies. Specifically, completion of the learning training task further decreased internal attribution of negative events, as well as increased internal and global attribution of positive events (posterior estimates=-0.51 [90% CrI=(-0.77, -0.26)], 1.24 [90% CrI=(0.77, 1.72)], 1.03 [90% CrI=(0.58, 1.47)]; Table S2).

Therefore, at the group level, both completion of the restructuring intervention and completion of learning training task impacted causal attribution tendencies for everyday events—with both intervention components resulting in a decreased tendency to choose unhelpful and increased tendency to choose self-enhancing interpretations.

Learning rates from the learning training task and changes in self-enhancing attributions

If learning is critical to the effects described above, we might reasonably expect that the effects of the learning task intervention to depend on individual differences in learning performance. We next explored whether model-based metrics of learning were related to changes in causal attribution tendencies.

Learning rates were estimated from learning training task data using a simple Rescorla-Wagner model (see Methods). Full information on model derivation via model comparison, chosen model performance, and simulation-based calibration analysis (including recovery of individual model parameters) can be found in the Methods and Supplementary Results.

Given we observed minimal variation in learning about negative events in our samples (Figure S2), we focused our analysis on learning estimates for positive events. Specifically, positive learning rates from the learning task were then compared to changes in self-enhancing attributions (internal and global interpretations of positive events) on the causal attribution task.

As a first-pass analysis, we examined relationships between point estimates (posterior parameter means) from separately modelled learning and causal attribution task data. We then carried out a formal test of association by analysing learning and causal attribution task data together in a joint hierarchical Bayesian model. This approach allows for the direction

estimation of associations between relevant parameters in the form of posterior regression weights (see Methods).

Associations between separately modelled learning and attribution task data. We observed associations between positive learning rates (α_{pos}) and changes in internal and global attributions of positive events (study 1: $R_{\alpha_{pos},\Delta internal}$=0.24, p<0.001, $R_{\alpha_{pos},\Delta global}$=0.10, p=0.15; study 2: $R_{\alpha_{pos},\Delta internal}$ =0.24, p<0.001, $R_{\alpha_{pos},\Delta global}$ =0.20, p<0.001; all correlations weighted by the posterior precision of α_{pos} estimates; Figure 3*A*,*D*; for pre- and post-intervention parameter estimates see Figure S3). These relationships were not evident for learning rates derived from the control learning task (*R*=0.14 & 0.10; Figure 3*D*).

There was no convincing evidence that the strength of these correlations differed between participants who received the cognitive restructuring compared to control interventions (for change in internal-positive attribution tendencies, study 1: *R*=0.27 & *R*=0.21, study 2: *R*=0.12 & *R*=0.22; for change in global-positive attribution tendencies, study 1: *R*=0.20 & *R*=0.04, study 2: *R*=0.25 & *R*=0.18, all p>0.9, Fisher's R-to-Z tests).

Joint hierarchical Bayesian modelling of learning and attribution task data.

Results of the first joint models provided strong evidence of positive relationships between positive learning rate (α_{pos}) estimates and changes in internal and global attributions of positive events, across intervention conditions, in study 1 participants (β_{LEARN} internal-positive=0.56 [90% CrI=(0.34, 0.88)], β_{LEARN} global-positive=0.46 [90% CrI=(0.27, 0.72)], Figure 3*B*, Table S3). These effects were replicated in study 2 data (β_{LEARN} internal-positive=0.29 [90% CrI=(0.17, 0.45)], β_{LEARN} global-positive=0.26 [90% CrI=(0.13, 0.42)]) - but were not evident for learning rates estimated from the control learning task ($\beta_{CONTROL}$ internal-positive=0.01 [90% CrI=(-0.02, 0.03)], $\beta_{CONTROL}$ global-positive=0.01 [90% CrI=(-0.01, 0.03)], Figure 3*E*, Table S4). This suggests that associations between speed of learning and subsequent change in self-enhancing attribution tendencies were specific to learning training in the domain of causal attributions.

Results of the second joint models provided some weak evidence for an additional influence of α_{pos} estimates on change in internal-positive attributions in participants who completed the restructuring intervention in study 1 ($\beta_{LEARN+CR}$ internal-positive=0.23 [90% CrI=(-0.002, 0.42)]), but there was no evidence for this effect in study 2 ($\beta_{LEARN+CR}$ internal-positive=-0.07 [90% CrI=(-0.23, 0.08)]). In neither study was there any convincing evidence for an additional influence of α_{pos} estimates on change in global-positive attributions in restructuring group participants (study 1: $\beta_{LEARN+CR}$ global-positive=0.10 [90% CrI=(-0.06, 0.28)],Figure 3*C*, Table S5, study 2: $\beta_{LEARN+CR}$ global-positive=0.09 [90% CrI=(-0.04, 0.24)], Figure 3*F*, Table S6). Therefore, we found no strong evidence in favour of a selective interaction between faster learning on the learning training task and response to the cognitive restructuring intervention.

Importantly, when the likelihood of the attribution task data was compared between the original analysis model and joint models, both joint models had superior predictive accuracy in left-out data (Table S7). This suggests that overall estimates of learning rates from the learning task were providing relevant information for inferring post-intervention causal attribution task parameter values.

LEARNING RATES FROM LEARNING TASK DATA REFLECT UNDERSTANDING OF THE TASK STATE-SPACE

We next explored relationships between learning rates estimates and other learning task data. Specifically, after each learning task scenario, participants were asked to provide explicit ratings of the kinds of causes that were thought to be 'correct', along internal-external and global-specific dimensions, and also provided free-text descriptions of each cause. Full analysis of learning task data (choice accuracy, response times, explicit-cause ratings and free-text cause descriptions) is available in the Supplementary Results.

Relationships between positive learning rates and explicit cause ratings. Posterior mean estimates of learning rates for positive events (α_{pos}) were positively associated with the explicit ratings of 'correct' causes for each task scenario. In other words, participants who learned faster to select internal-global attributions of positive events during the task were also able to better identify that correct causes were internal (self-related) and global (general), using explicit rating scales (study 1: R=0.2−0.35, p<0.005, study 2: R=0.20−0.33, p≤0.033, Figure S6). These relationships persisted in linear mixed-effects models controlling for scenario number and mean posterior inverse temperature (β) parameter values, weighted by posterior precision of α_{pos} estimates (internal-external cause ratings: study 1: $F_{1,238=}$17.2, study 2: $F_{1,114.5=}$9.5, p<0.005; global-specific cause ratings: study 1: $F_{1,235=}$13.2; p<0.001, study 2: $F_{1,112.7=}$4.1, p<0.05).

Relationships between positive learning rates and free-text cause description label probabilities. In study 1, there was strong evidence that posterior mean estimates of α_{pos} were positively correlated with classifier label probabilities for positive events in each scenario, along the internal-external dimension ([events were caused by] "myself", R=0.24−0.28, p<0.001; [events were caused by] "other people", R=0.2−0.32, p<0.001; Figure S7). These effects persisted in linear mixed-effect models controlling for scenario number and posterior mean inverse temperature (β) parameter values, weighted by posterior precision of α_{pos} estimates ($F_{1,239=}$12.1, p<0.001; $F_{1,267=}$5.6, p<0.02). In study 2, this association was only marginally evident ([events were caused by] "myself", R=0.17−0.22, p<0.07; [events were caused by] "other people", R=0.14−0.25, p<0.10), and did not survive in the controlled model ($F_{1,111=}$2.13, p=0.15, $F_{1,140=}$2.59, p=0.11). No relationships were evident between learning rates and classifier label probabilities for the free-text descriptions of positive events in the global-specific dimension in either sample, likely as this dimension was represented much more noisily in classifier output (see Supplementary Results).

In summary, in a reinforced setting, participants who learned more quickly to select self-enhancing attributions were also able to better describe the types of causes reinforced as correct during the task. This suggests that participants with higher learning rate estimates (α_{pos}) may have had a greater understanding of the ground truth dimensions along which response options (potential causal explanations) varied, allowing them to more quickly choose the 'correct' responses for a given scenario.

DISCUSSION

Theories of cognitive restructuring suggest it is a process based on learning (*13*). Individual differences in learning and memory of therapy content may be a moderator of symptom change during treatment (*18*, *19*). Inspired by recent demonstrations that clinically relevant inference processes can be reliably measured using computerised learning tasks (*22*, *23*), we sought to explore whether ability to recognise and learn about different attributions during a learning task was related to the subsequent changes in causal attribution tendencies, in the absence or presence of a brief cognitive restructuring intervention.

Contrary to our expectations, we found little evidence that individual differences in learning were specifically related to change in attribution tendencies following the restructuring intervention. Instead, we found robust evidence to support the idea that completion of the learning task had additive effects to completion of either intervention condition, particularly in boosting shifts towards self-enhancing (internal and global) attributions of positive events. Across studies, the magnitude of these effects was related to how quickly participants updated their choices according to reinforcement of an (implicit) internal-global response dimension on the learning task. Participants with faster learning rate estimates also showed greater ability to explicitly label correct responses along these ground truth dimensions, suggesting better overall understanding of the task state-space. Together, this suggests that individuals with a more intuitive understanding of these dimensions may be most likely to respond to this kind of training.

Several previous studies which have attempted to shift appraisals of everyday events using online training. For example, in non-clinically depressed participants, a single session of app-based reappraisal training was found to result in maladaptive response biases to ambiguous imagined scenarios in individuals given negative training, and adaptive biases in individuals given positive training (*24*). Similarly, three weeks of online training was found to increase self-reported reappraisal skill use—with participants who reported low levels of reappraisal use at baseline benefiting more in terms of improvement in depression symptoms (*25*). A recent meta-analysis also found evidence that cognitive bias training (where participants are typically presented with ambiguous everyday scenarios and trained to resolve them in favour of neutral or positive interpretations) reduced symptoms of anxiety and depression compared to some control conditions (*26*).

A novel aspect of the learning task described here is the use of a third-person perspective, alongside explicit reinforcement. It is possible that this is an effective strategy in helping participants learn to recognise different kinds of causal attribution tendencies, since distancing techniques are often employed during cognitive restructuring (*27*) and can alter learning (*16*). One advantage of tasks that can measure attribution biases along multiple dimensions—in conjunction with interpretable computational models—is that this information can be fed back to users over time. Future studies could explore the impact of this kind of informed training on learning speed, self-relevant attribution, and symptom change, as a form of acute psychological treatment augmentation (*20*).

The major limitations of the studies presented here are that participation was not restricted to individuals currently experiencing clinically significant levels of psychological symptoms, and that the brief restructuring intervention used here was not a real-world (i.e., clinically validated) psychological treatment component. It will be important to test in future work whether findings extend to these settings. However, measuring the impact of isolated therapy components on their proposed cognitive mechanisms in experimental settings has been

proposed to be a useful first step in understanding how and when psychotherapeutic techniques result in meaningful clinical improvement (*28*, *29*).

There are also some methodological limitations of our studies. First, for consistency with our previous work (*17*), we interpret 90% CrIs excluding zero as indicative of a meaningful contribution of a given parameter to the behavioural dimension of interest. This may be more prone to false positives than wider intervals, though we note 95% CrIs also excluded zero for all single-signed credible intervals we report in the text (see Supplementary Tables). More pertinently, although inference procedures for the learning task model were well-calibrated, we do not provide empirical data regarding the test-retest reliability of these measures. This limits our ability to infer reliable individual differences in learning between participants. We were unable to investigate individual differences in learning about negative events on the learning training task (a dimension that may be particularly relevant for depression), given our sample was mostly at ceiling for this response dimension. It is also important to note that our single-session experimental design, whilst supporting fast and high-throughput measurement in a relevant sample of individuals, may result in increased likelihood of motivational biases or demand effects influencing our primary dependent measures (i.e., participants updating their responses on the second attribution task in line with previously reinforced 'correct' responses on the learning training task, or the perceived purpose of the study). It will therefore be vital to determine in future work whether effects observed here are evident over longer timescales, and if they generalise to interpretations of the causes of events in users' own lives.

A fundamental aim of this kind of research is to help address barriers to the uptake and use of existing psychological interventions—in particular, remotely-delivered treatments where the potential for impact is large, but where initializing engagement and high attrition rates are acute problems (*30*). One factor that has been identified by users of digital mental health products is a "need... to experience a sense of 'self' in the treatment" (*31*). It is possible that using cognitive tasks with interpretable model-based output, and, critically, feeding this information back to users can help address this need. The utility of these approaches needs to be established in empirical studies, ideally with participation from all relevant stakeholders. Promisingly, e-mental health applications offer the potential to test these questions directly and at scale in an agile way, which may help substantially reduce the time between development and implementation of new treatment strategies (*32*).

METHODS

DATA AND CODE AVAILABILITY STATEMENT

Code for implementing all tasks and analyses described here, alongside anonymized study data is available on the study GitHub repository.

ETHICAL APPROVAL

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

PARTICIPANTS

Participants were recruited from an online research participation platform (Prolific (*21*)), and required to be resident in the UK, 18-65 years old, and fluent in English. Power analyses for both studies are available in the Supplementary Methods.

STUDY DESIGN

The design of each study is described in Figure 1*A*. Upon recruitment to each study, participants were assigned to a study arm using a random number generation-based procedure. All studies took place online over a single session, of approximately one hour.

MEASURES

Causal attribution task

A full description of the causal attribution task, including task development, design optimization, and measurement properties can be found in Norbury *et al.* (2024) (*17*). Of note, output parameters from the associated analysis model have excellent identifiability and test-retest reliability, and have previously been found to be associated with self-reported negative self-beliefs and current depression symptom severity.

Briefly, participants were instructed 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, if it had happened to them.

In each of two equivalent versions of the task (one pre- and one post-intervention), participants were presented with 32 event scenarios (16 positive and 16 negative events, randomly interleaved), divided into two blocks. Event scenarios differed across the two task versions. 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) (*33*).

Learning training task

The learning training task was developed as a measure of how easily participants can learn to select different kinds of causal attributions, in a reinforced setting. In contrast to the causal attribution task, the learning training task used a third person framing. Specifically, participants were told that they would be learning about how a hypothetical person in a particular mood might reason about the causes behind events. For each scenario, it was their job to learn to select the correct kinds of explanations for that person in that mood, via trial and error. Participants were provided with explicit instructions stressing the differences between the learning and attribution tasks, and required to pass a multiple-choice post-instructions quiz before proceeding (for full details, see screenshots available on the study GitHub repository). After each scenario, participants were asked to provide ratings and brief free-text description of the kinds of causes that were correct in that scenario (see Supplementary Methods).

Control learning task

The control learning task was exactly matched in trial type and reinforcement structure to the causal learning task. Participants were told that they would see a series of different coloured and shaped baskets, below which would be two different objects that could potentially belong to them. For each scenario, it was their job to learn which kinds of objects belonged in each basket, by trial and error (see Supplementary Methods). Again, participants provided explicit ratings and free-text descriptions of objects that belonged in each type of basket at the end of each scenario. All other aspects of task design were identical to the learning training task.

Brief cognitive restructuring and control interventions

The brief cognitive restructuring 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 cognitive restructuring intervention was based on cognitive therapy materials (*1*) and consisted of information about a cognitive model of mood, 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. The control intervention was based on materials from emotion-focused therapy (*34*), 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. The full content of each intervention is available on the study GitHub repository.

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 sociodemographic information (see Supplementary Methods).

Analysis

All analyses were conducted in R version 4.1.2 (R Core Team, 2021).

Initial statistical analysis of learning task data

Preliminary statistical analysis of learning task data was via mixed-effects linear regression models, as implemented in lme4 (see Supplementary Methods).

Classification of learning task free-text data

To measure how well participants were able to describe the ground-truth causes in each scenario in their own words, free-text responses were passed to a zero-shot natural language processing (NLP) classification pipeline (Facebook's BART-MNLI-LARGE transformer model (*35*)), with the non-mutually-exclusive candidate labels [”myself”, ”other people”, ”in general”, ”specific situations”]. Output probabilities for each candidate label were further analysed as above.

Hierarchical Bayesian modelling

General methods. Model parameters were estimated using Markov chain Monte Carlo (MCMC) sampling as implemented in Stan 2.21.0 (*36*), using RStan 2.21.3 (Stan Development Team, 2021). All models used generic weakly-informative priors (see Supplementary Methods). We report quantile-based 90% CrIs for consistency with results reported in our previous work on similar data (*17*), though the respective quantiles for 95% intervals can be found in the Supplementary Tables.

Hierarchical Bayesian analysis of causal attribution task data

Modelling of causal attribution task data followed the approach previously described in Norbury *et al.* (2024) (*17*), using an analysis model for which task design was previously optimised (see Supplementary Methods). Group-level parameters described potential effects of allocation to the restructuring intervention on individual-level parameter estimates at time 2, with priors for these parameters centred on 0. For study 2, this model was augmented to include potential effects of allocation to the learning training task condition.

Hierarchical Bayesian analysis of learning task data

For model-based analysis of learning task data, choices were collapsed to binary selection of internal-global and non-internal-global responses, separately for positive and negative events, to allow for repeat assessment of learning across the three task scenarios. Choice data were then modelled using a series of simple reinforcement-learning models based on the Rescorla-Wagner algorithm (see Supplementary Methods). Under this framework, values of each response option (internal-global and non-internal-global explanations) in each state (for a positively or negatively valence event) are updated on each trial using a surprise term, which is simply the difference between trial feedback (correct or incorrect) and the previously estimated value for that option in that state, multiplied by a learning rate.

Associating separately modelled causal attribution and learning task data parameters

As simple first-pass check, we examined correlations between point estimates (posterior means) of each parameter, weighted by the posterior precision (i.e., 1/standard deviation (SD)) of the predictor variable (α_{pos}). This is not an optimal way to test for associations between different estimates, since it neglects information about the individual precision of both parameter estimates.

Joint modelling of causal attribution and learning task data

To formally test for associations between parameters, we constructed a series of joint models of causal attribution and learning task data (*37*, *38*). For the first joint models, the causal attribution task analysis model (Supplementary Methods) was extended such that individual estimates for positive learning rates from the learning task (α_{pos}) were allowed to influence relevant post-intervention (time 2) causal attribution task parameter estimates ($\phi_{p,2}$) via the inclusion of β weight parameters (β_{LEARN}; see Hopkins *et al.* (2021) (*23*) and Haines *et al.* (2020) (*39*)). These β weights can interpreted similarly as in a standard regression model, with the group-level intervention effects (e.g., ϕ_{CR} for the cognitive restructuring (CR) intervention) now representing the intercept.

$$\phi_{p,1} = \phi_{\mu,1} + \tilde{\phi}_{p,1}$$

$$\phi_{p,2} = \phi_{\mu,2} + \tilde{\phi}_{p,2} + \begin{cases} \phi_{CR} + \alpha_{\text{pos}} * \beta_{\text{LEARN}} & \text{if CR intervention + learning task,} \\ \alpha_{\text{pos}} * \beta_{\text{LEARN}} & \text{if control intervention + learning task.} \end{cases}$$

(1)

For study 2 data, the first joint model included separate β weights for participants who completed the learning training *vs.* control learning tasks (β_{LEARN}, $\beta_{CONTROL}$):

$$\phi_{p,1} = \phi_{\mu,1} + \tilde{\phi}_{p,1}$$

$$\phi_{p,2} = \phi_{\mu,2} + \tilde{\phi}_{p,2} + \begin{cases} \phi_{\text{CR}} + \phi_{LEARN} + \alpha_{\text{pos}} * \beta_{\text{LEARN}} & \text{if CR intervention + learning task,} \\ \phi_{LEARN} + \alpha_{\text{pos}} * \beta_{\text{LEARN}} & \text{if control intervention + learning task,} \\ \phi_{\text{CR}} + \alpha_{\text{pos}} * \beta_{\text{CONTROL}} & \text{if CR intervention + control learning task.} \end{cases}$$

(2)

The second joint models added additional β weights for participants randomized to complete the CR intervention ($\beta_{LEARN+CR}$), to test for the presence of larger influences of learning rates on pre-post changes in attribution in participants who received both learning training and brief CR.

For study 1:

$$\phi_{p,1} = \phi_{\mu,1} + \tilde{\phi}_{p,1}$$

$$\phi_{p,2} = \phi_{\mu,2} + \tilde{\phi}_{p,2} + \begin{cases} \phi_{CR} + \alpha_{\text{pos}} * (\beta_{\text{LEARN}} + \beta_{LEARN+CR}) & \text{if CR intervention + learning task,} \\ \alpha_{\text{pos}} * \beta_{\text{LEARN}} & \text{if control intervention + learning task.} \end{cases} \quad (3)$$

For study 2:

$$\phi_{p,2} = \phi_{\mu,2} + \tilde{\phi}_{p,2} + \begin{cases} \phi_{\text{CR}} + \phi_{LEARN} + \alpha_{\text{pos}} * (\beta_{\text{LEARN}} + \beta_{LEARN+CR}) & \text{if CR intervention + learning task,} \\ \phi_{LEARN} + \alpha_{\text{pos}} * \beta_{\text{LEARN}} & \text{if control intervention + learning task,} \\ \phi_{\text{CR}} + \alpha_{\text{pos}} * \beta_{\text{CONTROL}} & \text{if CR intervention + control learning task.} \end{cases}$$

(4)

For all joint models, the priors for β effects were centred on zero (e.g., $\beta_{LEARN} \sim N(0,1)$).

ACKNOWLEDGEMENTS

This research was funded by a research grant from Koa Health to QJMH, SF and RD and a Wellcome Trust grant to QJMH (221826/Z/20/Z). We acknowledge support by the UCLH NIHR BRC. QD is supported by a Wellcome Trust PhD studentship. 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. 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.

DISCLOSURES

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. All other authors report no biomedical financial interests or potential conflicts of interest.

Supplement Description:
Supplement Methods, Results, Figures S1-S10, Tables S1-S8

Figure 1: Overview of study designs and measures. **A** Experimental designs and randomisation conditions for each study. In both studies, a cognitive-behavioural measure of causal attribution tendencies (the causal attribution task), was completed pre- and post-completion of two types of intervention. In study 1, all participants completed the learning training task and were randomly allocated to complete either brief cognitive restructuring or a control intervention. In study 2, participants were randomly assigned to complete either learning training or a control learning task, followed by either brief cognitive restructuring or a control intervention. All studies took place online, over a single experimental session (around one hour in length). **B** Representative screenshots of different study measures. The causal attribution task asks participants to choose between four different potential explanations of events, if such an event happened to them. The learning training task uses a third person framing and requires participants to learn the kinds of explanations thought to be correct for a hypothetical person in a particular mood state, given explicit feedback. The control learning task, identical in structure, requires participants to learn about the properties of objects, rather than causal explanations. The brief cognitive restructuring (and control) interventions both took the form of a series of interactive worksheets, which asked participants to learn about a particular therapy model and then apply it to recent events from their own lives. Further screenshots and demonstrations of the tasks and interventions are available on the study GitHub repository.

Figure 2: Independent effects of learning training and brief cognitive restructuring on causal attribution. **A** Posterior mean (and SD) parameter estimates for the causal attribution task for each participant at time 1 (pre-intervention) and time 2 (post-intervention) by randomisation group, in study 1 participants (*N*=200). Parameter estimates plotted here represent the probability of endorsing a given kind of attribution for positive and negative events, which are governed by the latent trait parameters (*θ*). Lines of best fit for mean time 1 *vs.* time 2 estimates for individuals in each group are plotted for illustration purposes. **B** Posterior parameter estimates for group means (over all participants/randomisation conditions) for each parameter at each time point, and the additional effect of the cognitive restructuring intervention at time 2, in study 1 participants, where P denotes probability. Thick inner lines represent 50% and thin outer lines represent 90% quantile-based CrIs (i.e., 90% of the probability density contained within the interval). For visualisation purposes, intervention effects (bold text) have been scaled by the square root of the mean posterior variance estimates for parameter values at time 2, making them roughly equivalent to standardised mean differences (SMDs). **C** The same plot as (*A*), for study 2 participants (*N*=164). **D** The same plot as (*C*), for study 2 participants; P denotes probability. Here, group-level effects on time 2 parameter estimated were modelled separately for participants who completed the restructuring *vs.* control intervention, and learning *vs.* control learning training.

		Study 1	Study 2
	N	200	164
Age (years)	Mean (SD)	37.2 (10.5)	36.9 (10.5)
	Range	19-63	20-65
Gender	Woman	110 (55%)	75 (46%)
	Man	86 (43%)	86 (52%)
	Non-binary or other	4 (2%)	3 (2%)
Race / ethnicity	White	165 (83%)	125 (78%)
	Asian	14 (7%)	13 (8%)
	Black	5 (3%)	12 (7%)
	Mixed	8 (4%)	10 (6%)
	Other	8 (4%)	3 (2%)
Employment status	Employed	147 (74%)	127 (77%)
	Unemployed	19 (10%)	13 (8%)
	Not seeking	33 (17%)	24 (15%)
Financial status	Doing okay	95 (48%)	85 (52%)
	Just about getting by	74 (37%)	61 (37%)
	Struggling	30 (15%)	18 (11%)
Housing status	Homeowner	90 (45%)	87 (53%)
	Tenant	86 (43%)	49 (30%)
	Other	23 (12%)	28 (17%)
Neurodivergence	Yes	25 (13%)	25 (15%)
	No	167 (84%)	135 (82%)
	Prefer not to say	8 (4%)	4 (2%)
Previous treatment for a mental health problem	Yes	89 (45%)	55 (34%)
	No	103 (52%)	105 (64%)
	Prefer not to say	8 (4%)	4 (2%)
If yes, type of treatment (all that apply)	Talking therapy	62 (31%)	36 (22%)
	Medication	62 (31%)	37 (23%)
	Self-guided	39 (20%)	27 (17%)
	Other	5 (3%)	4 (2%)
PHQ-9 (/27)	Mean (SD)	7.3 (6.2)	6.3 (5.8)
DAS-SF (/36)	Mean (SD)	19.2 (4.6)	18.6 (4.8)
miniSPIN (/12)	Mean (SD)	5.8 (3.6)	5.5 (3.4)

Table 1: Self-reported demographic and clinical data for all study participants. Self-reported race/ethnicity was based on information provided by Prolific. All other information was recorded via our custom demographic questionnaire (see Methods). 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 explained 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 attention-deficit/hyperactivity disorder (ADHD)*'. Categories for previous mental health treatment were talking therapy (including CBT), medication, self-guided (e.g., workbooks or apps), or other. The PHQ-9 assesses depressed mood; the short-form dysfunctional attitudes scale (DAS-SF) assesses dysfunctional beliefs; and the miniSPIN assesses social anxiety.

Figure 3: Changes in self-enhancing attributions were positively associated with learning rate estimates from the learning training task, but this effect was not greater in participants who completed cognitive restructuring. **A** Correlations between posterior mean estimates for positive learning rate from the learning training task (α_{pos}) and changes in mean values of parameters governing tendency to select internal and global attributions of positive events in study 1 participants. Point weights represent the estimated posterior precision of α_{pos} (i.e., 1/SD). **B** Posterior estimates of group-level effects from joint models of learning and causal attribution task data. β_{LEARN}, posterior estimates for weight of α_{pos} estimates on change in internal and global attributions of positive events. For visualization purposes, β estimates have been scaled by the ratio of predictor (i.e., α_{pos}) and outcome (i.e., mean posterior parameter) SDs, making them roughly equivalent to standardized regression coefficients. Black lines represent 50 & 90% posterior CrIs, and shading represents the posterior probability density. **C** The same plot as *B*, for a joint model with additional β weights for participants who completed brief cognitive restructuring in addition to learning training ($\beta_{LEARN+CR}$). **D** The same plot as A, for study 2 participants. **E** The same plot as *B*, for study 2 participants. $\beta_{CONTROL}$, posterior estimates for weight of control learning task learning rate estimates on change in attribution tendencies. **F** The same plot as *E*, for a joint model with additional weights for participants who completed brief cognitive restructuring in addition to learning training.

References

1. A. T. Beck, A. J. Rush, B. F. Shaw, G. Emery, *Cognitive Therapy of Depression* (Guilford Press, New York, 1979).
2. D. A. Clark, Cognitive Reappraisal. *Cognitive and Behavioral Practice* **29**, 564–566, DOI (2022).
3. L. Lorenzo-Luaces, R. E. German, R. J. DeRubeis, It's complicated: The relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clinical Psychology Review*, Psychological Interventions for Depression **41**, 3–15, DOI (2015).
4. L. Lorenzo-Luaces, J. R. Keefe, R. J. DeRubeis, Cognitive-Behavioral Therapy: Nature and Relation to Non-Cognitive Behavioral Therapy. *Behavior Therapy* **47**, 785–803, DOI (2016).
5. A. E. Kazdin, Understanding how and why psychotherapy leads to change. *Psychotherapy Research* **19**, 418–428, DOI (2009).
6. N. E. Hundt, J. Mignogna, C. Underhill, J. A. Cully, The relationship between use of CBT skills and depression treatment outcome: a theoretical and methodological review of the literature. *Behavior Therapy* **44**, 12–26, DOI (2013).
7. D. R. Strunk, S. N. Hollars, A. D. Adler, L. A. Goldstein, J. D. Braun, Assessing Patients' Cognitive Therapy Skills: Initial Evaluation of the Competencies of Cognitive Therapy Scale. *Cognitive Therapy and Research* **38**, 559–569, DOI (2014).
8. L. L. Hawley *et al.*, Cognitive-Behavioral Therapy for Depression Using Mind Over Mood: CBT Skill Use and Differential Symptom Alleviation. *Behavior Therapy* **48**, 29–44, DOI (2017).
9. N. B. Gumport, L. Dong, J. Y. Lee, A. G. Harvey, Patient Learning of Treatment Contents in Cognitive Therapy. *Journal of Behavior Therapy and Experimental Psychiatry* **58**, 51–59, DOI (2018).
10. N. R. Forand *et al.*, Efficacy of Guided iCBT for Depression and Mediation of Change by Cognitive Skill Acquisition. *Behavior Therapy* **49**, 295–307, DOI (2018).
11. I. D. Schmidt, B. J. Pfeifer, D. R. Strunk, Putting the "cognitive" back in cognitive therapy: Sustained cognitive change as a mediator of in-session insights and depressive symptom improvement. *Journal of Consulting and Clinical Psychology* **87**, 446–456, DOI (2019).
12. L. Lorenzo-Luaces, Identifying active ingredients in cognitive-behavioral therapies: What if we didn't? *Behaviour Research and Therapy* **168**, 104365, DOI (2023).
13. M. Moutoussis, N. Shahar, T. U. Hauser, R. J. Dolan, Computation in Psychotherapy, or How Computational Psychiatry Can Aid Learning-Based Psychological Therapies. *Computational Psychiatry* **2**, 50–73, DOI (2018).
14. A. M. Reiter, N. A. Atiya, I. M. Berwian, Q. J. Huys, Neuro-cognitive processes as mediators of psychological treatment effects. *Current Opinion in Behavioral Sciences*, Computational cognitive neuroscience **38**, 103–109, DOI (2021).
15. Q. J. Huys, M. Browning, M. P. Paulus, M. J. Frank, Advances in the computational understanding of mental illness. *Neuropsychopharmacology* **46**, 3–19, DOI (2021).

16. Q. Dercon *et al.*, A core component of psychological therapy causes adaptive changes in computational learning mechanisms. *Psychological Medicine* **54**, 327–337, DOI (2023).

17. A. Norbury, T. U. Hauser, S. M. Fleming, R. J. Dolan, Q. J. M. Huys, Different components of cognitive-behavioral therapy affect specific cognitive mechanisms. *Science Advances* **10**, eadk3222, DOI (2024).

18. S. J. E. Bruijniks, R. J. DeRubeis, S. D. Hollon, M. J. H. Huibers, The Potential Role of Learning Capacity in Cognitive Behavior Therapy for Depression: A Systematic Review of the Evidence and Future Directions for Improving Therapeutic Learning. *Clinical Psychological Science* **7**, 668–692, DOI (2019).

19. A. G. Harvey *et al.*, Improving Outcome of Psychosocial Treatments by Enhancing Memory and Learning. *Perspectives on Psychological Science* **9**, 161–179, DOI (2014).

20. C. L. Nord *et al.*, A transdiagnostic meta-analysis of acute augmentations to psychological therapy. *Nature Mental Health* **1**, 389–401, DOI (2023).

21. S. Palan, C. Schitter, Prolific.Ac—A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance* **17**, 22–27, DOI (2018).

22. H. M. Dorfman, R. Bhui, B. L. Hughes, S. J. Gershman, Causal Inference About Good and Bad Outcomes. *Psychological Science* **30**, 516–525, DOI (2019).

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

24. M. L. Woud, P. Postma, E. A. Holmes, B. Mackintosh, Reducing analogue trauma symptoms by computerized reappraisal training - considering a cognitive prophylaxis? *Journal of Behavior Therapy and Experimental Psychiatry* **44**, 312–315, DOI (2013).

25. R. R. Morris, S. M. Schueller, R. W. Picard, Efficacy of a Web-based, crowdsourced peer-topeer cognitive reappraisal platform for depression: randomized controlled trial. *Journal of Medical Internet Research* **17**, e72, DOI (2015).

26. L. A. Fodor *et al.*, Efficacy of cognitive bias modification interventions in anxiety and depressive disorders: a systematic review and network meta-analysis. *The Lancet Psychiatry* **7**, 506–514, DOI (2020).

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

28. S. J. E. Bruijniks, M. Sijbrandij, C. Schlinkert, M. J. H. Huibers, Isolating therapeutic procedures to investigate mechanisms of change in cognitive behavioral therapy for depression. *Journal of Experimental Psychopathology* **9**, 2043808718800893, DOI (2018).

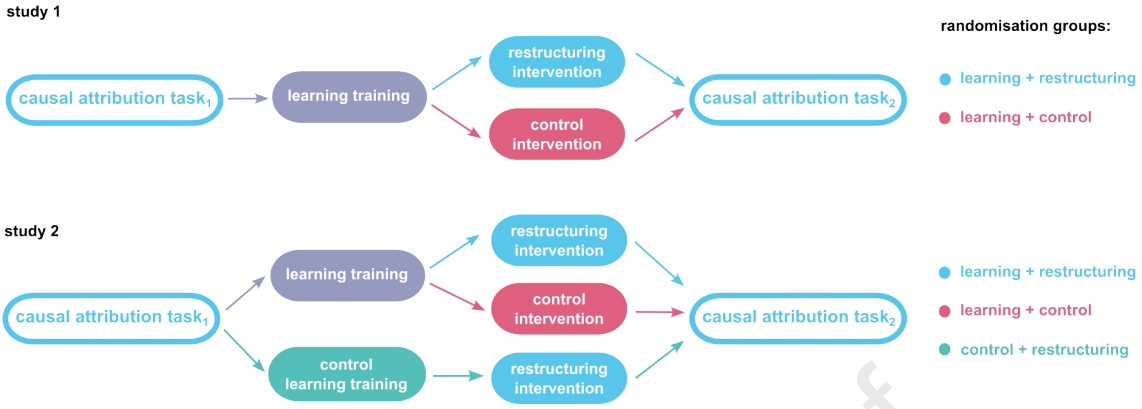
29. M. J. H. Huibers, L. Lorenzo-Luaces, P. Cuijpers, N. Kazantzis, On the Road to Personalized Psychotherapy: A Research Agenda Based on Cognitive Behavior Therapy for Depression. *Frontiers in Psychiatry* **11**, 607508, DOI (2021).

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

31. S. E. Knowles *et al.*, Qualitative meta-synthesis of user experience of computerised therapy for depression and anxiety. *PloS One* **9**, e84323, DOI (2014).

32. C. Seiferth *et al.*, How to e-mental health: a guideline for researchers and practitioners using digital technology in the context of mental health. *Nature Mental Health* **1**, 542–554, DOI (2023).
33. L. Y. Abramson, M. E. Seligman, J. D. Teasdale, Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology* **87**, 49–74, DOI (1978).
34. L. S. Greenberg, *Emotion-focused therapy: Coaching clients to work through their feelings* (American Psychological Association, Washington, DC, US, ed. 2, 2015), DOI.
35. M. Lewis *et al.*, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, 2019, Preprint on *arXiv*.
36. B. Carpenter *et al.*, Stan: A Probabilistic Programming Language. *Journal of Statistical Software* **76**, 1–32, DOI (2017).
37. B. M. Turner, B. U. Forstmann, B. C. Love, T. J. Palmeri, L. Van Maanen, Approaches to Analysis in Model-based Cognitive Neuroscience. *Journal of Mathematical Psychology* **76**, 65–79, DOI (B 2017).
38. N. Haines, PhD thesis, The Ohio State University, 2021, LINK.
39. N. Haines *et al.*, Anxiety Modulates Preference for Immediate Rewards Among Trait-Impulsive Individuals: A Hierarchical Bayesian Analysis. *Clinical Psychological Science* **8**, 1017–1036, DOI (2020).

A



B