Cognitive Therapy and Research
https://doi.org/10.1007/s10608-024-10544-3

ORIGINAL ARTICLE

Changes to Positive Self-Schemas After a Positive Imagery Training are Predicted by Participant Characteristics in a Sample with Elevated Depressive Symptoms

Amanda C. Collins[1,2,7] · **George D. Price**[1,3] · **Justin Dainer-Best**[4] · **Dawson Haddox**[1] · **Christopher G. Beevers**[5] · **Nicholas C. Jacobson**[1,2,3,6]

Accepted: 23 October 2024

Abstract

Background Depressed individuals have both heightened negative self-views and reduced positive self-views. The self-referential encoding task (SRET) can capture depressed individuals' self-schemas by asking them to endorse whether a word describes them or not. Digital interventions that target positive biases in depression can help improve positive self-schemas; however, it is important to determine who may respond best to these interventions. In the current study, we used a machine learning approach to predict changes in positive self-schemas on the SRET after a digital intervention.

Methods Participants were randomized to a digital imagery training that was either positive ($n=39$) or neutral ($n=38$) and completed the intervention every other day for 2 weeks. Participants also completed the SRET and self-report measures at pre-, mid-, and post-intervention to measure their self-schemas and psychopathology symptoms.

Results Results indicate the models were able to moderately predict changes in the number of self-referential positive words endorsed on the SRET, solely using participants' baseline characteristics ($r_{Test} = 0.33$).

Conclusions These findings suggest that certain characteristics may predict response to a digital intervention focused on improving positive biases, and current findings emphasize the use of machine learning to improve treatment match and triage persons to treatments that may work best.

Keywords Depression · Digital intervention · Machine learning · Self-schemas

✉ Amanda C. Collins
Amanda.C.Collins@dartmouth.edu

George D. Price
gdprice10@gmail.com

Justin Dainer-Best
jdainerbest@bard.edu

Dawson Haddox
dawson.d.haddox.25@dartmouth.edu

Christopher G. Beevers
beevers@utexas.edu

Nicholas C. Jacobson
nicholas.c.jacobson@dartmouth.edu

[1] Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA

[2] Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA

[3] Quantitative Biomedical Sciences Program, Dartmouth College, Lebanon, NH, USA

[4] Psychology Program, Bard College, Annandale-on-Hudson, NY, USA

[5] Department of Psychology and Institute for Mental Health Research, University of Texas atAustin, Austin, TX, USA

[6] Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Lebanon, NH, USA

[7] Center for Technology and Behavioral Health, Dartmouth College, 46Centerra Parkway Suite 300, Lebanon, NH 03766, USA

Published online: 09 November 2024

Introduction

Beck's cognitive theory posits that depressed individuals hold negative schemas, or core beliefs, about themselves, the world, and the future (Beck & Bredemeier, 2016). Negative schemas may contribute to the development and maintenance of depression (Gotlib & Joormann, 2010), and self-schemas (i.e., views about themselves) may be particularly important for predicting a relapse in depressive symptoms (LeMoult et al., 2017). Indeed, self-schemas have been viewed as the "core structure" of depression and may be a valuable target for the treatment of depression (Garratt et al., 2007). Cognitive therapies have often focused on restructuring negative self-schemas, or core beliefs, about the self (e.g., "I am a failure"), and psychopharmacological interventions may also be effective in reducing negative self-schemas (Cuijpers et al., 2010; Hasler et al., 2020).

Until recently, extant research has focused on the role of negative self-schemas in depression and has theorized that depressed individuals hold fewer positive self-schemas as an artifact of their enhanced negative self-schemas (Beck & Bredemeier, 2016). However, evidence suggests that depressed individuals do not just *lack* a positive self-schema, rather they devalue or dampen positivity regarding themselves (Collins et al., 2023a; Collins & Winer, 2024; Winer & Salem, 2016). Thus, it has been suggested that depressed individuals hold both enhanced negative and dampened positive self-schemas (Collins & Winer, 2024). Given this, it is important to identify interventions that can target both of these self-schemas.

Self-Referential Encoding Task

The Self-Referential Encoding Task (SRET; Derry & Kuiper, 1981) is an experimental paradigm that is well suited to assess self-schemas. In this task, participants view adjectives on the screen and are asked to respond 'yes' or 'no' to whether a word describes them or not. Thus, the SRET can assess an individual's self-schema, including whether the extent to which they hold positive or negative self-schemas (Dainer-Best, 2023). Meta-analytic findings indicate that depressed individuals hold both enhanced negative and dampened positive self-schemas; however, a difference emerges for individuals with more severe symptoms such that they endorse *fewer* positive words than negative words to describe themselves (Collins & Winer, 2024). Moreover, depressed individuals endorsed fewer positive words to describe themselves than nondepressed individuals, but they endorsed an equal number of positive words as nondepressed individuals to describe *another* person (Collins & Winer, 2024). This evidence suggests that this devaluation of positivity is restricted to their self-views rather than their views of others. Taken together, these meta-analytic findings demonstrate a need for targeting positive self-schemas with psychological interventions, rather than only negative self-schemas, as commonly focused on in the SRET literature.

Digital Interventions

The increased use of digital interventions provide a unique opportunity to target positive self-schemas in depressed individuals. Digital interventions can help reduce the burden placed on the individuals seeking treatment given that traditional in-person treatments can be expensive, time-consuming, and inaccessible (Davies et al., 2020; Kumar et al., 2013; Wilhelm et al., 2020). Moreover, many digital interventions can be accessed via a smartphone app, tablet, or computer, providing the unique opportunity for individuals to access the intervention whenever and wherever they need to (Srivastava et al., 2020). Thus, digital interventions are a valuable tool that can be used to increase the scalability of mental health treatment to those who need it, and they are often much more cost-effective than traditional in-person treatments (Andersson & Cuijpers, 2009).

Positive Self-Reference Training

Research has found a strong correlation between self-schemas and depressive symptoms, (Dainer-Best et al., 2018a; Dobson & Shaw, 1987; Phillips et al., 2010). Cognitive-behavioral therapy (Beck, 1976) can be utilized to address maladaptive *negative* cognitions; however, it does not place as much of an emphasis on *positive* cognitions. Relatedly, cognitive bias modification (CBM) interventions target specific cognitive biases, sometimes through the use of imagery-based techniques (Hallion & Ruscio, 2011). However, many prior CBM interventions have also primarily focused on reducing negative thoughts, behaviors, and emotions. While these interventions address the increased negativity biases, they do not consistently demonstrate consistency and frequently fail to directly address the decreased positivity biases experienced by individuals with depression, indicating a need for CBM interventions that focus on positive biases. Existing literature indicates that repeatedly encoding positive self-referent information through mental imagery makes positive self-referent information more accessible, and imagery techniques that focus on the future might subvert the memory biases associated with depression (Bradley et al., 1995; Dainer-Best et al., 2018b; Hamami et al., 2011; Klein, 2013), providing a framework for an intervention focused on positive biases. Additionally, an ideal intervention would be digital to diminish the burden on individuals seeking treatment (Davies et al., 2020; Kumar et al., 2013; Wilhelm et al., 2020).

The Positive Self-Reference Training (PSRT) aimed to address this need as a digital and imagery-based intervention focused on improving positive self-schemas in depressed individuals (Dainer-Best et al., 2018b). Before starting the training, participants were informed that the training aimed to "target unhelpful ways of thinking" and watched a 2-min video explaining the training. During the intervention, participants received positive cues twice on each training day. Examples of these cues include going to a café, being extremely interested in something, and receiving a gift. In response to the cues, participants created 3–5 min audio recordings in which they imagined and described future events that were specific, positive or fun, and involved themselves. These training sessions were completed every other day for 2 weeks during the intervention.

To investigate how positive self-schemas changed in response to the PSRT, Dainer-Best et al. (2018b) conducted a randomized controlled trial. Participants were randomized to either the PSRT or a no-training control condition (NTC). Participants in the NTC completed training sessions that were comparable to the PSRT and were likewise informed that their training aimed to "target unhelpful ways of thinking." However, the NTC removed active ingredients of the PSRT. First, the NTC cues only included places (not positive experiences), such as a clothing store, a bedroom, and an office. Second, NTC participants were instructed to respond to the cues by imagining present locations that are generic, neutral, and involved objects and spaces, rather than future and self-oriented experiences.

The SRET was administered online to both conditions at baseline, 1 week (i.e., midway), and 2 weeks (i.e., completion) of the training to assess for changes in self-schemas. Findings revealed that participants in the PSRT group demonstrated a significant increase in the number of positive words they endorsed as self-referential compared to those in the NTC group. In addition, both groups exhibited decreases in their endorsement of negative self-referential words. Nevertheless, individuals in the PSRT condition did not show a significantly greater decrease in depression symptoms compared to those in the NTC condition. Instead, both groups demonstrated a reduction in depression symptoms throughout the study. Taken together, these findings suggest that an online imagery training can improve depressive symptoms and negative self-schemas; however, an online imagery training focused on *positive* imagery may be warranted for those depressed persons who show a specific devaluative positive self-schema (Collins & Winer, 2024).

Despite the promise of these findings, it remains unclear exactly who would benefit the most from this digital intervention to improve their self-schemas. Initial findings indicate that, whereas many individuals in the PSRT group demonstrated an increase in their endorsement of positive self-referential words, there were some individuals who demonstrated no change, or worsening of their positive self-schemas (Dainer-Best et al., 2018b). Thus, there is likely heterogeneity in participants that impacts who majorly benefited from the PSRT, and who received little benefit, or no benefit, from the PSRT. A further investigation into who is likely (or not likely) to benefit from the PSRT specifically is warranted to increase treatment match in the future.

Rationale

The purpose of the current study is to conduct a secondary analysis (Dainer-Best et al., 2018b) and investigate whether machine learning is capable of predicting changes in an individual's self-schema after the PSRT. Specifically, we aim to determine whether participants' characteristics (e.g., depressive symptoms, demographics, etc.) before starting the intervention can predict improvement *after* the intervention targeting a depressed person's positive self-schema. We hypothesized that our models would be able to predict changes in the endorsement of self-referential positive words from baseline features with a moderate performance. We also investigated the most influential features in the improvement of self-referential processing throughout the intervention as an exploratory analysis to better determine who would benefit from the current digital intervention of a positive imagery training.

Methods

Participants

Participants were recruited to participate in a mood study using a variety of methods across the country. These included posts on Craigslist, ResearchMatch, and online postings through the University of Texas at Austin. Participants who provided online informed consent, their age, and completed the Center for Epidemiologic Studies - Depression scale (CES-D) were considered for eligibility. The eligibility criteria for enrollment into the study included (1) being between the ages of 18–45, (2) meeting criteria for at least mild depressive symptoms (CES-D > 13; Radloff, 1977), and being fluent in English. Participants received compensation of $25 for their participation in the study; however, participants were able to earn up to $10 extra (i.e., $35 total) if they completed extra training sessions.

Materials and Measures

Self-Referential Encoding Task

The SRET is a behavioral task that can be used to investigate self-schemas. In the current study, participants completed the SRET on their own computer at home via Inquisit software. Participants view words on the screen one at a time and are asked to quickly decide if a word describes them. The words remained on the screen until a selection was made, and participants were instructed to press the Q key to indicate that the word shown described them, and they were instructed to press the P key to indicate that the word shown did not describe them. Participants first completed a practice block of five words to become oriented with the task. Then, they completed the experimental block with 90 words (45 positive and 45 negative) that were randomized, with each word being followed by a 1500 ms intertrial interval. The SRET was given at all three timepoints, and the same 80 words were shown each time; however, 10 additional words (5 positive and 5 negative) were added at each timepoint, which were different at each time, resulting in 90 words total shown at each timepoint.

The SRET has demonstrated good internal consistency ($\alpha = 0.93$–0.97) and 1-week test-retest reliability ($r = .87$) for endorsement of positive words (Dainer-Best et al. 2018a). The endorsement of positive words showed good internal consistency in the current study across all three timepoints ($\alpha > .92$). Our primary variable of interest was changes in self-referential positive endorsement (i.e., the number of positive words that participants responded 'yes' to), so an outcome variable was calculated by summing the number of endorsed words at each time point. The changes score of endorsement of positive words also showed good internal consistency ($\alpha = .86$, 95% CI [.81, .89]). For further detail into the development of the current version of the SRET, including the word list, we direct readers to Dainer-Best et al. (2018b).

Center for Epidemiologic Studies-Depression (CES-D)

The CES-D is a 20-item self-report that assesses depressive symptoms over the past week (Radloff, 1977). Items are scored on a four-point Likert scale from 0 (rarely or none of the time) to 3 (most or all of the time). A sum score can be created by summing all items and can range from 0 to 60, with greater scores representing greater depressive symptoms. Prior work has also identified four subscales (Shafer, 2006), allowing for the examination of overall depressive symptoms and a more nuanced examination of individual factors of depression. The depressed affect subscale includes seven items that represent a primarily depressed mood (e.g., "felt depressed" and "felt sad"). The positive affect subscale includes four items that represent a lack of positive affect; however, items are reverse coded (e.g., "felt happy" and "hopeful about the future"). The somatic subscale includes seven items that primarily represent appetite, sleep, and concentration difficulties (e.g., "appetite poor" and "restless sleep"). The interpersonal problems subscale includes two items that represent difficulties with social relationships (e.g., "people dislike me" and "people were unfriendly"). The CES-D has demonstrated adequate sensitivity (median = 0.85) and specificity (median = 0.72; Vilagut et al., 2016).

Mood and Anxiety Symptom Questionnaire—Short Form (MASQ)

The Mood and Anxiety Symptom Questionnaire—Short Form (MASQ) is a 30-item self-report measure that assesses symptoms of general distress, including anxiety and depressive symptoms, over the past week. Items are scored on a five-point Likert scale from 1 (not at all) to 5 (extremely). A sum score can be created by summing all items and can range from 30 to 150. Prior work has identified three subscales, which contain ten items each (De Beurs et al., 2007; Wardenaar et al., 2010). The general distress subscale represents non-specific general distress or general negative affect (e.g., "felt hopeless" and "worried a lot about things"). The anhedonic depression subscale represents a lack of or reduced positive affect; however, items are reverse coded (e.g., "felt really happy" and "felt that I had a lot to look forward to"). The anxious arousal subscale represents somatic symptoms of anxiety (e.g., "was short of breath" and "felt dizzy or light-headed"). The MASQ and its subscales have demonstrated good internal consistency ($\alpha = 0.87$–0.93), construct validity, and convergent validity with other psychopathology measures (Wardenaar et al., 2010).

Procedure

The current study was approved by the Institutional Review Board at the University of Texas at Austin. The trial was pre-registered prior to beginning data collection at ClinicalTrials.gov (NCT03056963; Dainer-Best, 2017). Prior to beginning the study, participants provided signed consent online via the Research Electronic Data Capture (REDCap; Harris et al., 2009) and then completed the CES-D. Participants who scored a 13 or greater on the CES-D, which represents mild severity or greater of depressive symptoms, completed other self-report measures, including the MASQ and demographics. Participants were also asked to generate a personal, non-identifiable ID that could be used to link their survey responses to their other data throughout the survey. After completing the survey on REDCap, participants were then given instructions to download the Inquisit software to complete the SRET, which they completed at that time.

Participants who completed the REDCap survey and the SRET were randomized to either the PSRT or NTC condition. They were given background about the training that they were participating in (as described above) and received information about the training schedule and compensation.

To orient participants to the training, they completed practice cues that were neutral for all participants on the first day. For the intervention, they watched the imagery video on the first day and then completed two respective cue training sessions every other day for 2 weeks. During the cue training, participants recorded themselves on their personal phones or computers using their microphones, and these recordings were automatically uploaded through Telegram or the website. Participants were contacted via email as a reminder to complete at least four training sessions every other day.

Participants were asked to complete the CES-D, MASQ, and SRET on the fourth day of training after baseline and at the end of the intervention. If participants completed all three assessments (baseline, timepoint 1, timepoint 2) and at least 10 cue recordings, then they were compensated $25. If participants completed more than 10 cue recordings, they were paid an additional compensation of $10 maximum, which was scaled based on the number of extra recordings they completed. Eighty-seven participants completed the entire study, including the intervention and the baseline and timepoint 2 assessments [see Fig. 2 for the CONSORT flow chart in Dainer-Best et al. (2018b)].

Data Analytic Plan

Although the trial was pre-registered prior to data collection, the current data analytic plan was not pre-registered. Our outcome variable of interest was changes in the number of self-referential positive words endorsed on the SRET at the timepoint 1 and at timepoint 2 (# of positive words endorsed$_{\mathrm{T1\ or\ T2}}$ - # of positive words endorsed$_{\mathrm{T0}}$), with greater scores indicating an increase in self-referential endorsement of positive words. We excluded participants from our analyses who were missing SRET data at any timepoint, resulting in a final sample size of 77 participants (PSRT: 39 participants; NTC: 38 participants). Table 1 demonstrates the demographics for participants by group.

We included the following baseline variables as features in the model to predict changes in endorsement of positive words: treatment condition (PSRT or NTC), CES-D sum score and subscale scores (somatic, depressed, positive, interpersonal), MASQ sum score and subscale scores (general distress, anhedonic depression, anxious arousal), number of negative words endorsed on the SRET, age, sex, ethnicity, race, type of residence, length of living at current residence (years), length of living at current residence (months), marital status, total number of people in household, years of school completed, highest degree obtained, current student status, employment status, annual household income, current MDD episode, past MDD episode, age of first MDD episode, number of lifetime MDD episodes, prior diagnosis of an anxiety disorder, currently attending therapy, prior psychiatric hospitalization, and number of recordings completed during intervention. Prior to modeling, we scaled these features from [0, 1], which has been shown to increase model efficiency and performance (Shahriyari, 2019). We also dummy recoded categorical variables (e.g., race) to binary variables, given that some machine learning models cannot handle categorical variables, with a selection of a category (e.g., 'Asian' for the race variable) as 1 and a non-selection of a category as 0.

Table 1 Participant baseline demographics

Demographics	PSRT group ($n=39$)	NTC group ($n=38$)
Age	25.14(6.16)	27.82(7.99)
Gender	29 Women 7 Men 1 Other ("Gender-queer")	30 Women 5 Men
Ethnicity	34 Non-Hispanic 3 Hispanic	31 Non-Hispanic 4 Hispanic
Race	26 White 4 Asian 4 Black 3 Other	26 White 6 Asian 2 Black 1 Other
CES-D	25.67(9.82)	29.45(10.43)
MASQ	85.67(17.83)	88.89(15.70)
Current MDD episode	26 No 11 Yes	25 No 10 Yes

Five participants were not given demographics, and thus total counts do not add up to 77 (PSRT = 39; NTC = 38).

After data preprocessing, we ran our ML model. Given the smaller sample size for the current study, we first evaluated the performance of four commonly-used algorithms in ML on our training set to determine which ML algorithm would be best suited for our data. These algorithms included ridge, extreme gradient boosted tree, random forest, and k-nearest neighbors. Importantly, these algorithms were only trained on the training set and not the test set to avoid overfitting. Results from these evaluations indicated that k-nearest neighbors performed the best (algorithm = auto, n-neighbors = 2, weights = uniform), so we included these parameters in our final model with a nested CV. Specifically, we utilized a nested cross-validation (CV) approach with 10 splits in Python (v3.9) and the k-nearest neighbors algorithm (algorithm = ball_tree, n-neighbors = 3, weights = uniform). Traditional approaches often use k-fold CV to split the dataset (e.g., 5 splits) into training and test sets; however, this can be problematic and lead to biased and inaccurate results with smaller sample sizes (Vabalas et al., 2019). Rather, nested CV uses an iterative process to include each participant as an individual test set, providing unbiased results. Using this approach, the dataset is split into two loops: an outer loop and inner loop. In the outer loop, we split the data further into a training and test set such that each participant was held out as a test set once, with all other participants being used in the training set ($n = 76$). Each training set was then evaluated with the inner loop where it is again split into a training and validation set. Within the inner loop, we used the k-nearest neighbors algorithm on the training set and then evaluated the model on the validation set. Then, we used the resulting parameters on the test set ($n = 1$). Compared to more commonly used approaches, nested CV is more suitable to use with smaller sample sizes (Vabalas et al., 2019).

Models that demonstrated a moderate correlation or greater ($r < 0.30$) were further interpreted, as done in prior work (Collins et al., 2023b; Lekkas et al., 2022). Specifically, we utilized SHapley Additive exPlanations (SHAP) to provide a better understanding of which features were most important in influencing the model's predictions of improvements in endorsement of self-referential positive words. This approach, based on the Shapley values of game theory, calculates the relative importance of each feature in the model's prediction of the outcome variable (Lundberg & Lee, 2017). SHAP has previously been used to investigate which features may be the most influential in predicting changes in symptoms following digital interventions (Collins et al., 2023b; Price et al., 2022). As commonly done, we extracted the top five features that were the most influential to the model's predictions.

Results

Our modeling approach indicated a moderate correlation ($r_{Test} = 0.33$, $r_{Validation} = 0.41 \pm 0.36$) for predicting changes in the number of self-referential positive words endorsed on the SRET at timepoint 2 (i.e., 2 weeks from baseline) after the positive imagery training (see Table 2). We found a weak correlation ($r_{Test} = 0.09$, $r_{Validation} = 0.05 \pm 0.43$) at timepoint 1 (i.e., 1 week from baseline) and thus did not follow up with SHAP for timepoint 1 (see Fig. 1 for a scatter plot of the actual vs predicted values at each timepoint).

Participants' endorsement of self-referential positive words increased, on average, from 19.57 to 23.71. Further investigation with a t-test revealed that there were no significant differences between the PSRT group and the NTC group in changes in words endorsed ($t = 0.93$, $p = 0.36$); however, prior work using odd ratios found that participants in the PSRT demonstrated more of a clinically meaningful change than participants in the NTC (Dainer-Best et al., 2018b).

As noted above, we followed up our moderate model performance via SHAP for timepoint 2 only due to weak modeling performance at timepoint 1. We investigated the top 5 most important features in the model's predictions at timepoint 2, and results indicate that the following features emerged as the most influential: treatment condition (i.e., NTC or PSRT), history of a prior diagnosis of an anxiety disorder, receiving a terminal's bachelor degree, having a household income less than \$25,000, and having a household income between \$50,000 and \$74,999 (see Fig. 2). Specifically, (1) being in the PSRT condition (compared to the NTC condition), (2) not having a prior diagnosis of an anxiety disorder, (3) having a terminal bachelor's degree, and (4) having a household income less than \$25,000 were predictive of an increase in the number of words endorsed as positive on the SRET. Interestingly, having a household income between \$50,000 and \$74,999 was predictive of a decrease in the number of words endorsed as positive on the SRET.

The only feature related to mental health that emerged was having a prior anxiety disorder diagnosis (i.e., no depression-related features). Thus, as an exploratory

Fig. 1 Scatter plot of actual vs predicted values. *Note* Left figure represents the model's performance at timepoint 1; right figure represents the model's performance at timepoint 2

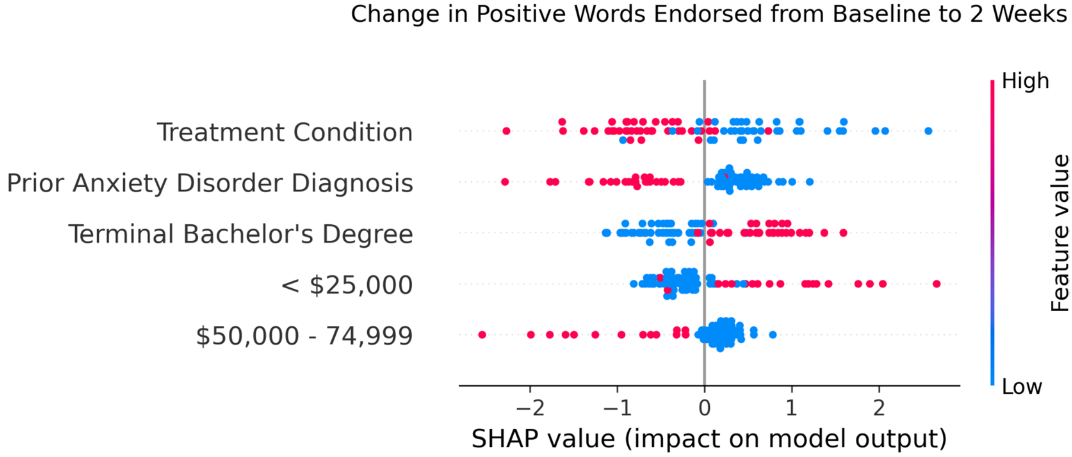

Fig. 2 Top influential features in predicting changes in self-referential positive words endorsed from baseline to 2 weeks. *Note* "< $25,000" and "$50,000–74,999" reflect self-reported Annual Household Income. A high value (i.e., red) indicates a high value of the feature, and a low value (i.e., blue) indicates a low value of the feature. Given that all variables that emerged as most influential in this figure were binarized (e.g., yes or no), there are only high and low values in the SHAP plot (e.g., no moderate). Negative SHAP values indicate a decrease in the number of positive words endorsed (timepoint 2—baseline), and positive SHAP values indicate an increase in the number of positive words endorsed (Color figure online)

Table 2 Modeling results

Model outcome	Test set	Validation set(s)	
	Correlation (*r*)	Correlation (*r*)	Standard deviation (*Z* score)
Timepoint 1	0.09	0.05	0.43
Timepoint 2	0.33	0.41	0.36

analysis, we investigated whether any other features related to mental health emerged in the top 20 features. Findings revealed that having a past MDD diagnosis (#6), history of attending therapy (#11), having a current diagnosis of MDD (#17), and a previous psychiatric hospitalization (#20) were all within the top 20 features. Measures of symptoms (i.e., scores on the CES-D or MASQ) were not included in the 20 features.

Discussion

The aim of the current study was to determine whether or not changes in positive self-schemas after a digital intervention could be predicted by baseline features to enhance treatment match. To accomplish this aim, we utilized a machine learning approach to predict changes in the endorsement of self-referential positive words from pre- to post-intervention (i.e., baseline to timepoint 2) of a positive imagery training intervention. Our findings indicate that our model was capable of predicting changes in positive self-schemas after the intervention with a moderate correlation, solely using participants' baseline features.

Interestingly, our model was not able to predict changes in positive self-schemas at timepoint 1 (i.e., 1 week after starting the intervention). This may indicate that a meaningful change in self-schemas may not occur after a short period of time (i.e., 1 week or three intervention sessions), supporting the initial findings of Dainer-Best et al. (2018b). Little research has investigated the time frame in which self-schemas may change, so the lack of findings after 1 week may indicate that a longer time frame is needed for one to develop the skills necessary for altering their self-schemas. Prior work has investigated how self-schemas change in response to cognitive-behavioral therapy and pharmacotherapy, with a steady improvement of positive self-schemas across 16 weeks of receiving either treatment (Quilty et al., 2014). However, whether self-schemas can be altered in a shorter time frame in response to an intervention had not yet been investigated until the current research (Dainer-Best et al., 2018b). Future work can investigate the more precise time frame at which self-schemas change to further refine current interventions, including digital interventions such as the PSRT.

Investigation of the most influential features in the model's predictions after the intervention indicated that the condition in which participants were assigned was the most influential such that those in the PSRT were more likely to demonstrate improvements in their positive self-schemas. This finding supports the aim of the development of the PSRT, given that depressed individuals experience reduced positive self-schemas (Collins & Winer, 2024;

Dainer-Best et al., 2018b); however, it may also support the claim that including a vague, neutral imagery training is beneficial but not sufficient enough to improve self-schemas. Taken together with the initial findings of Dainer-Best et al. (2018b), which indicated that individuals in the PSRT experienced a greater improvement in both their positive and negative self-schemas compared to the NTC group, the current findings reiterate the general ability of a positive imagery digital intervention to improve self-schemas.

The findings from our SHAP plot also indicate that not having a history of an anxiety diagnosis was influential in the model's prediction of improvements in positive self-schemas. Depression and anxiety are highly comorbid with one another, and comorbid psychopathology is predictive of poorer treatment outcomes (Choi et al., 2020; Pearson et al., 2019). Thus, it is reasonable that individuals in the current study who self-reported a previous or current anxiety disorder did not respond as well to this digital intervention. Prior work has indicated that depressed individuals with anxiety may take longer to respond to treatments than depressed individuals without anxiety However, features directly related to depression (e.g., history of MDD) did not emerge within the top five most influential features. Given that SHAP evaluates *relative* feature importance, it may be that the top features (e.g., treatment condition and history of an anxiety disorder) contribute to a greater explainability than the depression-related features. Importantly, the correlations between depression-related features at baseline and changes in positive words endorsed were all non-significant. In addition, the sample recruited in the current study had to endorse at least mild depressive symptoms, which may indicate that their depression-related features were truncated given that these were measured at baseline (i.e., a skewed rather than normal distribution).

Lastly, several demographic features emerged as influential in the model's predictions, including having an income of less than $25,000, having a terminal Bachelor's degree, and having an income between $50,000 and $74,999. Follow-up analyses revealed that these features were not significantly related to engagement in the intervention (e.g., completing more recordings). Although employment status did not emerge as an important predictor, it is closely related to income. Thus, our findings do provide some support for prior research showing that employment status is an important predictor of treatment response (DeRubeis et al., 2014; Huibers et al., 2015). Future research can aim to investigate how individual patient characteristics, including household income and education level, may impact symptom outcomes after a digital intervention to better increase treatment match (Moshe et al., 2021).

Strengths and Limitations

There are several strengths to the current study. First, participants were recruited across the United States and self-reported elevated depressive symptoms, which may make these findings more generalizable to a clinical population (i.e., those diagnosed with MDD); however, replication using a clinical sample is warranted to better investigate these findings in individuals with an MDD diagnosis. Second, given the nature of nested CV, we were also able to test and train the model on every single participant in the dataset, which can help to prevent overestimation of the model's parameters. Third, the modeling approach utilized a longitudinal design such that we used baseline features to predict future symptom change, and we used SHAP to provide further detail of the model's introspection. These are particularly important to establish temporal precedence and determine who will benefit the most from the current digital intervention.

Although there are several notable strengths discussed above, there are also a few limitations to discuss. First, as noted above, the trial was pre-registered but the current data analytic plan and hypotheses were not pre-registered. This study was largely exploratory in nature to determine if ML models could predict changes in self-referential biases; however, we did not have any specific hypotheses regarding which features would emerge as the most important in the model's predictions. In addition, we did not have any a priori decisions regarding the specific ML algorithm and parameters given that prior work predicting symptom changes after an intervention have utilized different algorithms, including, but not limited to, random forest (Hornstein et al., 2021), ensemble models (Gyorda et al., 2023; Jacobson & Nemesure, 2021), and extreme gradient boosted machine learning models (Collins et al., 2023b; Lekkas et al., 2021). Thus, given the exploratory nature and lack of a priori methodological decisions, we did not pre-register the current study but do agree that pre-registration should be done whenever possible, including when there are clear hypotheses and methodological decisions. As such, the lack of pre-registration for the current study is a large limitation, and future work wanting to investigate whether these findings replicate in a different sample should pre-register their hypotheses and analytic plan. Second, a large number of participants were lost to attrition in the original sample Dainer-Best et al., 2018b), resulting in a smaller sample size for the current analysis. Although this limitation was circumvented by using nested CV, recent research has noted limitations with machine learning when generalizing to other samples (Chekroud et al., 2024), so it is important for future work to further investigate generalization in other samples, including within persons who meet diagnostic criteria for MDD, and investigate the factors that may contribute to participant dropout to enhance completion, including symptom severity and demographics. Third, although endorsement on the SRET is an important metric for investigating self-schemas, prior work has shown that recall may be as important to best capture biases, as demonstrated by larger meta-analytic effect sizes (Collins & Winer, 2024). Thus, future work can investigate how recall of self-referential words changes after a digital intervention. Fourth, as noted in the original trial paper, positive self-schemas improved more in the PSRT condition than the NTC group, but the two groups did not differ on their improvement in depressive symptoms. Thus, we focused on whether or not ML could predict changes in positive self-schemas, rather than depressive symptoms. However, it is important to also investigate who would benefit from a digital intervention like the PSRT to improve their depressive symptoms. Thus, future work with larger sample sizes and persons diagnosed with MDD could evaluate how ML models predict changes in self-schemas and depressive symptoms, as well as which participant characteristics may influence treatment outcomes for both of these constructs.

Conclusion

The current study highlights the utility of using machine learning to predict treatment outcomes following a digital intervention focused on enhancing positive biases. Our findings provide insight into the participant characteristics that may be the most influential in predicting treatment outcomes. Moreover, the current work provides important implications into who may benefit the most, based on their baseline features, from the current digital intervention to improve positive self-schemas in depression. Future work can aim to replicate these findings in a larger sample with more severe depressive symptoms.

Author Contributions ACC: conceptualization, data curation, formal analysis, investigation, methodology, writing—original draft, writing—reviewing and editing; GDP: data curation, formal analysis, methodology, visualization, writing—original draft, writing—reviewing and editing; JDB: conceptualization, data curation, investigation, methodology, writing—reviewing and editing; DDH: writing—original draft, writing—reviewing and editing; CGB: funding acquisition, investigation, methodology, resources, writing—original draft; NCJ: funding acquisition, writing—reviewing and editing.

Funding N.C.J. received funding from the National Institute of Mental Health (NIMH) and the National Institute Of General Medical Sciences (NIGMS) under R01 MH123482, and C.G.B. received funding from the NIMH under MH108650, MH110758, MH109600 and the Wayne H. Holtzman Regents Chair in Psychology from the University of Texas at Austin. This research was also supported by the National Institute on Drug Abuse Grant T32DA037202 to fund the fellowship of A.C.C.

Data Availability Data and analytic code is publicly available on OSF: https://osf.io/jhnsm/.

Declarations

Competing Interests Amanda C. Collins, George D. Price, Justin Dainer-Best, Dawson Haddox, Christopher G. Beevers, and Nicholas C. Jacobson declare that they have no competing interests.

Informed Consent The Institutional Review Board at the University of Texas at Austin approved all procedures. Participants provided signed informed consent online using REDCap electronic data capture tools hosted at UT-Austin.

Animal rights No animal studies were carried out by the authors for this article.

References

Andersson, G., & Cuijpers, P. (2009). Internet-based and other computerized psychological treatments for adult depression: A meta-analysis. *Cognitive Behaviour Therapy, 38*(4), 196–205. https://doi.org/10.1080/16506070903318960

Beck, A. T. (1976). *Cognitive therapy and the emotional disorders* (p. 356). International Universities Press.

Beck, A. T., & Bredemeier, K. (2016). A unified model of depression: Integrating clinical, cognitive, biological, and evolutionary perspectives. *Clinical Psychological Science, 4*(4), 596–619. https://doi.org/10.1177/2167702616628523

Bradley, B. P., Mogg, K., & Williams, R. (1995). Implicit and explicit memory for emotion-congruent information in clinical depression and anxiety. *Behaviour Research and Therapy, 33*(7), 755–770. https://doi.org/10.1016/0005-7967(95)00029-w

Chekroud, A. M., Hawrilenko, M., Loho, H., Bondar, J., Gueorguieva, R., Hasan, A., Kambeitz, J., Corlett, P. R., Koutsouleris, N., Krumholz, H. M., Krystal, J. H., & Paulus, M. (2024). Illusory generalizability of clinical prediction models. *Science, 383*(6679), 164–167. https://doi.org/10.1126/science.adg8538

Choi, K. W., Kim, Y. K., & Jeon, H. J. (2020). Comorbid anxiety and depression: Clinical and conceptual consideration and transdiagnostic treatment. In Y.-K. Kim (Ed.), *Anxiety disorders* (Vol. 1191, pp. 219–235). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-32-9705-0_14

Collins, A. C., Lass, A. N. S., & Winer, E. S. (2023a). Negative self-schemas and devaluation of positivity in depressed individuals: A moderated network analysis. *Current Psychology*. https://doi.org/10.1007/s12144-023-04262-4

Collins, A. C., Price, G. D., Woodworth, R. J., & Jacobson, N. C. (2023b). Predicting individual response to a web-based positive psychology intervention: A machine learning approach. *The Journal of Positive Psychology*. https://doi.org/10.1080/17439760.2023.2254743

Collins, A. C., & Winer, E. S. (2024). Self-referential processing and depression: A systematic review and meta-analysis. *Clinical Psychological Science, 12*(4), 21677026231190390. https://doi.org/10.1177/21677026231190390

Cuijpers, P., Smit, F., Bohlmeijer, E., Hollon, S. D., & Andersson, G. (2010). Efficacy of cognitive–behavioural therapy and other psychological treatments for adult depression: Meta-analytic study of publication bias. *British Journal of Psychiatry, 196*(3), 173–178. https://doi.org/10.1192/bjp.bp.109.066001

Dainer-Best. (2017). Guided imagery training and self-referential processing. https://clinicaltrials.gov/study/NCT03056963

Dainer-Best, J. (2023). Trends in research on self-reference, depression, and the self-referent encoding task: A critical review [Preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/2va7g

Dainer-Best, J., Lee, H. Y., Shumake, J. D., Yeager, D. S., & Beevers, C. G. (2018a). Determining optimal parameters of the self referent encoding task: A large-scale examination of self-referent cognition and depression. *Psychological Assessment, 30*(11), 1527–1540. https://doi.org/10.1037/pas0000602

Dainer-Best, J., Shumake, J. D., & Beevers, C. G. (2018b). Positive imagery training increases positive self-referent cognition in depression. *Behaviour Research and Therapy*, *111*, 72–83. https://doi.org/10.1016/j.brat.2018.09.010

Davies, F., Shepherd, H. L., Beatty, L., Clark, B., Butow, P., & Shaw, J. (2020). Implementing web-based therapy in routine mental health care: Systematic review of health professionals' perspectives. *Journal of Medical Internet Research*, *22*(7), e17362. https://doi.org/10.2196/17362

De Beurs, E., Den Hollander-Gijsman, M. E., Helmich, S., & Zitman, F. G. (2007). The tripartite model for assessing symptoms of anxiety and depression: Psychometrics of the Dutch version of the mood and anxiety symptoms questionnaire. *Behaviour Research and Therapy*, *45*(7), 1609–1617. https://doi.org/10.1016/j.brat.2006.07.004

Derry, P. A., & Kuiper, N. A. (1981). Schematic processing and self-reference in clinical depression. *Journal of Abnormal Psychology*, *90*(4), 286–297. https://doi.org/10.1037/0021-843X.90.4.286

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS ONE, 9*(1), e83875. https://doi.org/10.1371/journal.pone.0083875

Dobson, K. S., & Shaw, B. F. (1987). Specificity and stability of self-referent encoding in clinical depression. *Journal of Abnormal Psychology*, *96*(1), 34–40. https://doi.org/10.1037/0021-843X.96.1.34

Garratt, G., Ingram, R. E., Rand, K. L., & Sawalani, G. (2007). Cognitive processes in cognitive therapy: Evaluation of the mechanisms of change in the treatment of depression. *Clinical Psychology: Science and Practice*, *14*(3), 224–239. https://doi.org/10.1111/j.1468-2850.2007.00081.x

Gotlib, I. H., & Joormann, J. (2010). Cognition and depression: Current status and future directions. *Annual Review of Clinical Psychology*, *6*(1), 285–312. https://doi.org/10.1146/annurev.clinpsy.121208.131305

Gyorda, J. A., Nemesure, M. D., Price, G., & Jacobson, N. C. (2023). Applying ensemble machine learning models to predict individual response to a digitally delivered worry postponement intervention. *Journal of Affective Disorders*, *320*, 201–210. https://doi.org/10.1016/j.jad.2022.09.112

Hallion, L. S., & Ruscio, A. M. (2011). A meta-analysis of the effect of cognitive bias modification on anxiety and depression. *Psychological Bulletin*, *137*(6), 940–958. https://doi.org/10.1037/a0024355

Hamami, A., Serbun, S. J., & Gutchess, A. H. (2011). Self-referencing enhances memory specificity with age. *Psychology and Aging, 26*(3), 636–646. https://doi.org/10.1037/a0022626

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Hasler, G., Suker, S., Schoretsanitis, G., & Mihov, Y. (2020). Sustained improvement of negative self-schema after a single ketamine infusion: An open-label study. *Frontiers in Neuroscience, 14*, 687. https://doi.org/10.3389/fnins.2020.00687

Hornstein, S., Forman-Hoffman, V., Nazander, A., Ranta, K., & Hilbert, K. (2021). Predicting therapy outcome in a digital mental health intervention for depression and anxiety: A machine learning approach. *Digital Health, 7*, 205520762110606. https://doi.org/10.1177/20552076211060659

Huibers, M. J. H., Cohen, Z. D., Lemmens, L. H. J. M., Arntz, A., Peeters, F. P. M. L., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the personalized advantage index approach. *PLoS ONE, 10*(11), e0140771. https://doi.org/10.1371/journal.pone.0140771

Jacobson, N. C., & Nemesure, M. D. (2021). Using artificial intelligence to predict change in depression and anxiety symptoms in a digital intervention: Evidence from a transdiagnostic randomized controlled trial. *Psychiatry Research, 295*, 113618. https://doi.org/10.1016/j.psychres.2020.113618

Klein, S. B. (2013). The complex act of projecting oneself into the future. *WIREs Cognitive Science*, *4*(1), 63–79. https://doi.org/10.1002/wcs.1210

Kumar, S., Nilsen, W. J., Abernethy, A., Atienza, A., Patrick, K., Pavel, M., Riley, W. T., Shar, A., Spring, B., Spruijt-Metz, D., Hedeker, D., Honavar, V., Kravitz, R., Lefebvre, C., Mohr, R., Murphy, D. C., Quinn, S. A., Shusterman, C., & Swendeman, D. (2013). Mobile health technology evaluation. *American Journal of Preventive Medicine, 45*(2), 228–236. https://doi.org/10.1016/j.amepre.2013.03.017

Lekkas, D., Price, G. D., & Jacobson, N. C. (2022). Using smartphone app use and lagged-ensemble machine learning for the prediction of work fatigue and boredom. *Computers in Human Behavior*, *127*, 107029. https://doi.org/10.1016/j.chb.2021.107029

Lekkas, D., Price, G., McFadden, J., & Jacobson, N. C. (2021). The application of machine learning to online mindfulness intervention data: A primer and empirical example in compliance assessment. *Mindfulness, 12*(10), 2519–2534. https://doi.org/10.1007/s12671-021-01723-4

LeMoult, J., Kircanski, K., Prasad, G., & Gotlib, I. H. (2017). Negative self-referential processing predicts the recurrence of major depressive episodes. *Clinical Psychological Science, 5*(1), 174–181. https://doi.org/10.1177/2167702616654898

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems* (pp. 4765–4774).

Moshe, I., Terhorst, Y., Philippi, P., Domhardt, M., Cuijpers, P., Cristea, I., Pulkki-Råback, L., Baumeister, H., & Sander, L. B. (2021). Digital interventions for the treatment of depression: A meta-analytic review. *Psychological Bulletin*, *147*(8), 749–786. https://doi.org/10.1037/bul0000334

Pearson, R., Pisner, D., Meyer, B., Shumake, J., & Beevers, C. G. (2019). A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. *Psychological Medicine*, *49*(14), 2330–2341. https://doi.org/10.1017/S003329171800315X

Phillips, W. J., Hine, D. W., & Thorsteinsson, E. B. (2010). Implicit cognition and depression: A meta-analysis. *Clinical Psychology Review*, *30*(6), 691–709. https://doi.org/10.1016/j.cpr.2010.05.002

Price, G. D., Heinz, M. V., Nemesure, M. D., McFadden, J., & Jacobson, N. C. (2022). Predicting symptom response and engagement in a digital intervention among individuals with schizophrenia and related psychoses. *Frontiers in Psychiatry*, *13*, 807116. https://doi.org/10.3389/fpsyt.2022.807116

Quilty, L. C., Dozois, D. J. A., Lobo, D. S. S., Ravindran, L. N., & Bagby, R. M. (2014). Cognitive structure and processing during cognitive behavioral therapy vs. pharmacotherapy for depression. *International Journal of Cognitive Therapy, 7*(3), 235–250. https://doi.org/10.1521/ijct.2014.7.3.235

Radloff, L. S. (1977). *The CES-D Scale: A self-report depression scale for research in the general population* (Vol. 1, pp. 385–401). https://doi.org/10.1177/014662167700100306

Shafer, A. B. (2006). Meta-analysis of the factor structures of four depression questionnaires: Beck, CES-D, Hamilton, and Zung. *Journal of Clinical Psychology*, *62*(1), 123–146. https://doi.org/10.1002/jclp.20213

Shahriyari, L. (2019). Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. *Briefings in Bioinformatics*, *20*(3), 985–994. https://doi.org/10.1093/bib/bbx153

Srivastava, K., Chaudhury, S., Dhamija, S., Prakash, J., & Chatterjee, K. (2020). Digital technological interventions in mental health care. *Industrial Psychiatry Journal*, *29*(2), 181. https://doi.org/10.4103/ipj.ipj_32_21

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLoS ONE, 14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Vilagut, G., Forero, C. G., Barbaglia, G., & Alonso, J. (2016). Screening for depression in the general population with the Center for Epidemiologic Studies Depression (CES-D): A systematic review with meta-analysis. *PLoS ONE, 11*(5), e0155431. https://doi.org/10.1371/journal.pone.0155431

Wardenaar, K. J., Van Veen, T., Giltay, E. J., De Beurs, E., Penninx, B. W. J. H., & Zitman, F. G. (2010). Development and validation of a 30-item short adaptation of the Mood and Anxiety Symptoms Questionnaire (MASQ). *Psychiatry Research*, *179*(1), 101–106. https://doi.org/10.1016/j.psychres.2009.03.005

Wilhelm, S., Weingarden, H., Ladis, I., Braddick, V., Shin, J., & Jacobson, N. C. (2020). Cognitive-behavioral therapy in the digital age: Presidential address. *Behavior Therapy*, *51*(1), 1–14. https://doi.org/10.1016/j.beth.2019.08.001

Winer, E. S., & Salem, T. (2016). Reward devaluation: Dot-probe meta-analytic evidence of avoidance of positive information in depressed persons. *Psychological Bulletin*, *142*(1), 18–78. https://doi.org/10.1037/bul0000022