The Dissertation Committee for Justin Lee Dainer-Best
certifies that this is the approved version of the following dissertation:

Identifying and Modifying Negative Self-Referent Cognition in Individuals with Depressive Symptoms

Committee:

Christopher G. Beevers, Supervisor

Sarah Kate Bearman

Bharath Chandrasekaran

David M. Schnyer

Jasper A. J. Smits

Identifying and Modifying Negative Self-Referent Cognition in Individuals with Depressive Symptoms

by

Justin Lee Dainer-Best

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2018

Identifying and Modifying Negative Self-Referent Cognition in Individuals with Depressive Symptoms

Justin Lee Dainer-Best, Ph.D.
The University of Texas at Austin, 2018

Supervisor: Christopher G. Beevers

Major Depressive Disorder (MDD) and other depressive disorders are associated with serious impairment. A greater understanding of the factors and biases that help maintain current depressive symptoms is important in helping to ameliorate the effects of these symptoms. Beck (1967) has argued that biases in self-relevant information processing play an especially important role in the maintenance of depression. Negative self-referent beliefs are theorized to lead to negative cognitive biases that amplify the effects of negative life stress, culminating in the symptoms of depression.

This dissertation focuses on negatively biased self-reference. It provides evidence for the strong linkage between self-reference and depressive symptoms (Study 1), shows that electrocortical measures of greater elaborative processing in response to negative information are associated with increased self-referential decision-making (Study 2), and demonstrates a method of modifying negatively biased self-referent processing (Study 3).

Study 1 investigated the psychometric properties of a task used to measure negative self-referent processing, the self-referent encoding task (SRET; Derry &

Kuiper, 1981). Study 1 found that the number of negative and positive words endorsed as self-referential, and the rate of accumulation of information to make the decision about whether each word was self-referential, were robustly predictive of depressive symptoms. Evidence also indicated that these indices of the SRET were psychometrically sound and had strong test-retest reliability.

Study 2 used event-related potentials (ERPs) in conjunction with the SRET to show that the endorsement of self-referential words occurred in a later, elaborative stage of neurocognitive processing and not in the early stages of information processing that are associated with the perception of stimuli. Participants diagnosed with MDD showed greater late, posterior ERP waveforms to negative vs. positive words when compared to healthy controls.

Study 3 shifted behavioral indices of negative self-referent processing on the SRET by using a future-oriented, guided positive self-reference training. Participants in the active condition had greater increases in their positive self-referent processing over two weeks than those who received a neutral control training.

The results of these studies bring together three levels of analysis in order to better measure, understand, and ultimately change negatively biased self-referent information processing associated with depression.

Table of Contents

List of Tables

List of Figures

General Introduction

Major Depressive Disorder (MDD) and other depressive disorders affect millions of Americans each year, and many more experience depressive symptoms that don't reach syndrome levels (Kessler et al., 2005). Even given an understanding of the impact of MDD, there remains much that we don't know about how changing thought processes may cause and interact with depressive thinking. The question of how individuals with MDD think differently motivates the studies contained in this dissertation. That is, this dissertation contains three studies that attempt to provide a basis for investigation of how those experiencing depressive symptoms think about themselves.

Besides the physical symptoms (e.g., changed appetite and sleep habits), individuals suffering from depression may present with anhedonia (diminished interest or pleasure in activities), difficulty concentrating, and negative thoughts and beliefs including feelings of worthlessness and thoughts of death (American Psychiatric Association, 2013). From another lens, latent class analyses have shown that depressive psychopathology consists of patterns relating to anhedonia, suicidality, psychomotor dysregulation, and severe depression (Chen, Eaton, Gallo, & Nestadt, 2000). Depressive symptoms may be brief or, as in the case of MDD, may persist over the course of weeks, months, or even years. Past researchers have attempted to investigate what causes these depressotypic beliefs, how they are maintained, and how to ameliorate them. Regardless of degree, the depressive disorders are debilitating illnesses, with impacts both personal and financial (Berto, D'Ilario, Ruffo, Virgilio, & Rizzo, 2000; P. E. Greenberg, Fournier, Sisitsky, Pike, & Kessler, 2015).

In understanding these difficulties, a theoretical framework is useful for asking the question of how MDD alters the cognition of depressed individuals. Various models of how MDD develops have been postulated throughout its history, including diathesis-stress models (Monroe & Simons, 1991; Caspi et al., 2003), evolutionary models (Price, Sloman, Gardner, Gilbert, & Rohde, 1994), and more behavioral models involving self-control (Rehm, 1977) or learned helplessness (Overmier & Seligman, 1967; Abramson, Seligman, & Teasdale, 1978). Although some—or even many—of these theories may be able to coexist in explaining the development of depression, the cognitive model of depression (Beck, 1967, 1974, 2008) is best suited for investigation of ways of thinking, especially given its focus on both development and treatment of depressive symptoms.

The Cognitive Model of Depression

Beck's (1967) cognitive model of depression posits that individuals' dysfunctional views about themselves, their future, and their environments work together to lay the groundwork for MDD. This tripartite construct is often called the depressive cognitive triad (Beck, 1967) and has been extensively studied (Beckham, Leber, Watkins, Boyer, & Cook, 1986).

The depressive cognitive triad is integrally tied with a second central construct of the cognitive model: the self-schema. The self-schema is synonymous with how individuals view themselves—their existing memory representation of themselves—which then influences how they process incoming information from the world around them. Theoretically, this self-schema provides a lens through which an individual interprets incoming information. That is, the schemas relating to the self are made up of past experiences that guide subsequent perception and new information is thus compared to these self-schemas (Segal, 1988; Kuiper, Olinger, MacDonald, & Shaw,

1985). For example, a history of being rejected in romantic relationships may lead to a negative self-schema that includes themes of being unlovable and/or worthless, through which ambiguous cues are interpreted as rejection. This cognitive bias, in turn, amplifies the impact of life events, particularly loss-related events in this case, and leads to the development of MDD (Beck, 1967; Disner, Beevers, Haigh, & Beck, 2011; Gotlib & Joormann, 2010). Once a negative self-schema is active, it continues to negatively bias information processing, leading to a prolonged negative mood state and may culminate in a depressive episode. In this fashion, negative self-referent processing is, therefore, implicated in the maintenance of depression (Padesky, 1994; Disner et al., 2011).

Beck (2008, 1967) has conjectured that a negative self-schema may result from adverse developmental experiences. These adverse experiences include a variety of types of loss, which Beck interprets as feeding into a self-schema that conceptualizes oneself as worthless and rejected. Such losses provide a basis for the development of self-schemas that are predictive of MDD. These dysfunctional beliefs about the self can lead to biased processing of incoming stimuli from the environment (Beck, 2008). Such biased processing is an important feature of depression: it leads to a closed loop in which negative self-schemas result in negative judgments about the world, the self, and the future—which in turn lead into the symptoms that make up MDD. This makes sense, given that especially when one is depressed, mood-congruent stimuli that also fit with a negative self-schema may be especially salient (P. C. Watkins, Vache, Verney, & Mathews, 1996; P. C. Watkins, Mathews, Williamson, & Fuller, 1992). And indeed, the cognitive theory predicts that biased processing of emotional material is integral to the maintenance of depression (Beck, 1967, 2008; Dozois & Beck, 2008; Disner et al., 2011). There has been substantial psychopathology research directed towards investigating the cognitive model of de-

pression (Haaga, Dyck, & Ernst, 1991). Research has shown that individuals who are depressed have more negative cognitions (Alloy et al., 1999; Blackburn, Jones, & Lewin, 1986; Dobson & Shaw, 1987), they are more likely to endorse negative adjectives as describing themselves (Derry & Kuiper, 1981; Dobson & Shaw, 1987; B. P. Bradley & Mathews, 1988), and demonstrate a negative self-concept (Tarlow & Haaga, 1996; Beck, Steer, Epstein, & Brown, 1990; McCauley, Mitchell, Burke, & Moss, 1988; Asarnow, Carlson, & Guthrie, 1987; Alloy et al., 1999). Although there are similarities between anxiety and depression in cognitive schemas (Segal, 1988; Barlow, Allen, & Choate, 2004), it seems clear that there are unique cognitive features in MDD, especially lack of positive affect (Clark & Watson, 1991), which are highlighted by the cognitive model.

Similarly, there has been a substantial degree of research directed towards evaluating the therapeutic outcome of this cognitive model. Cognitive therapy (Beck, 1979) developed out of the cognitive model of emotional disorders, and was followed by cognitive-behavioral therapy (CBT). Indeed, many of the symptom clusters involved in MDD, as well as the broader mechanisms that are its foundations, can be ameliorated or diminished by appropriate treatment. The cognitive model is descriptive of depression in many ways, but whether its central mechanisms play a key role in the maintenance of depression has been difficult to demonstrate. The efficacy of therapies based on this cognitive model has been studied extensively (Dobson, 1989; Hofmann, Asnaani, Vonk, Sawyer, & Fang, 2012), but it has been somewhat difficult to establish the active ingredients that lead to depression symptom change in a multifaceted treatment such as cognitive therapy or cognitive-behavioral therapy (Jacobson et al., 1996; Dimidjian et al., 2006).

Identifying the features of therapy that work best for given phenotypes is an important direction for future research. Further, developing models that identify the

features that maintain depressive symptomatology would help in identifying those active ingredients, and in developing new treatments that could be more effective inside or outside of a clinic. Identifying individual mechanisms which might change during psychotherapeutic intervention—and which are theorized to be important in the cognitive model of depression—is thus an effective method of exploring these treatments while developing new ideas for future intervention research.

Schema and Self-Reference

One promising and specific target for such interventions involves self-referent processing, specifically of negative and positive incoming information. As described above, the self-schema is integrally related to how individuals view information about themselves, influencing what they attend to, how they interpret ambiguous information, and what they subsequently recall. When individuals make decisions about whether or not information applies to them, these are considered decisions of reference.

A variety of researchers have attempted to conceptualize the process of making self-referent decisions about negative material, and how self-schema may work in maintaining depression. A negative self-schema has been assessed using the dysfunctional attitudes scale (DAS; Weissman & Beck, 1978; Oliver & Baumgart, 1985; Beck, Brown, Steer, & Weissman, 1991). The DAS has been shown to measure depressogenic beliefs relating to oneself in depressed (Beck et al., 1991) and unselected (Oliver & Baumgart, 1985) populations. When completing the DAS, negative beliefs about the self are identified by self-report. Items include, e.g., "I should be happy all the time" (Beck et al., 1991). Similarly, a schema questionnaire (Young, 1994; Schmidt, Joiner, Young, & Telch, 1995) identifies maladaptive schema processes which are present through a variety of emotional disorders. These questionnaires are

explicit in their identification of statements that reflect negative self-schema, e.g., "I am inherently flawed and defective" (Schmidt et al., 1995).

Behavioral tasks have also been employed in an attempt to measure self-schema, and more specifically self-reference. One task in particular that measures whether incoming information is self-referential is the self-referent encoding task (SRET; Derry & Kuiper, 1981). This task allows participants to identify whether positive and negative individual adjectives—e.g., lonely, happy, or upset—describe them. As such, it accesses the schema in that participants are constantly using their self-schema to identify whether individual words pertain to them (P. C. Watkins et al., 1992). Responses and reaction times (RTs) are traditionally collected; many studies have also collected either free recall or recognition data for the stimuli presented during the SRET. The SRET has been shown to be specific to depression (Dobson & Shaw, 1987; M. S. Greenberg & Alloy, 1989; Dozois & Dobson, 2001), distinguishing anxious participants from those with elevated depressive symptoms.

The SRET has been used extensively over the past forty years. It began primarily as a measure of memory encoding (Derry & Kuiper, 1981; Kuiper & Derry, 1982), and has also been used extensively as a measure of memory bias (Mineka & Nugent, 1995; B. P. Bradley & Mathews, 1988; Wisco, 2009; S. B. Klein & Loftus, 1988). It has also from the first been a measure of self-schema (Kuiper & Derry, 1982; Dobson & Shaw, 1987) and of negative affect and self-image (Clifford & Hemsley, 1987). The SRET was developed as a means of linking an individual's representation of themselves with a quantifiable behavior, the endorsement of adjectives. Thus the SRET is conceptualized "as a cognitive schema involved in the processing of personal and social information about one's self" (Derry & Kuiper, 1981, p. 286). It is useful in operationalizing negative self-perception and how self-referential stimuli become interpreted as negative; its outcome variables are discussed in Study 1.

The task was originally delivered via cards presented in a tachistoscope (Derry & Kuiper, 1981), which presented stimuli in front of participants for specified periods of time, although from shortly thereafter the task was administered using a computer in order to more accurately collect RTs (S. B. Klein, Loftus, & Burton, 1989), although the utility or RT data is unclear. It is almost universally administered via computer today.

Throughout the past decades, the SRET has been used to demonstrate the self-reference effect (for a review, see Symons & Johnson, 1997), which suggests that self-referent material is easier to recall than material which is irrelevant to the self. It takes no great leap to understand how this effect should relate to depression. If a negative self-schema makes negative information more accessible, then it follows that negative information should be easier to recall (Wisco, 2009). Indeed, the SRET has been used as a marker of negative self-schema and negative memory bias (Wisco, 2009; Kuiper & Derry, 1982; B. P. Bradley & Mathews, 1988; Sanz, 1996; Ingram, Kendall, Smith, Donnell, & Ronan, 1987). It has also more recently been used to measure negative self-referent processing more broadly (Connolly, Abramson, & Alloy, 2015; Goldstein, Hayden, & Klein, 2015; LeMoult, Kircanski, Prasad, & Gotlib, 2017).

Other recent studies have also shown that biased self-referential processing is related to neural processing. The SRET has been employed with neuroimaging techniques (fMRI; e.g., Kelley et al., 2002; Yoshimura et al., 2009; Schmitz & Johnson, 2006) and with electroencephalography (EEG; e.g., Auerbach, Stanton, Proudfit, & Pizzagalli, 2015; Auerbach et al., 2016; Speed, Nelson, Auerbach, Klein, & Hajcak, 2016; Kiang et al., 2017. Imaging studies showed that self-referent processing resulted in increased activation in the cingulate gyrus and the precuneus (Kelley et al., 2002; Yoshimura et al., 2009) and the anterior medial prefrontal cortex (Schmitz

& Johnson, 2006), areas which they posit may be involved in more automatic processing. EEG studies have specifically studied self-referent processing in depression, identifying specific time-periods during which evaluative processing of self-referent stimuli may be ongoing (Auerbach et al., 2015). These results can be viewed as furthering the argument that there is top-down control relating to affective stimuli (Comblain, D'Argembeau, & Van der Linden, 2005). Such research argues that self-referent processing may be targeted for assessment or intervention.

The SRET provides a tool for understanding how an individual's schema plays out in decision-making, and further scrutiny is important. Meta-analyses have also shown that the SRET is strongly correlated with depression (Phillips, Hine, & Thorsteinsson, 2010). Thus, given the role that self-schema plays in the cognitive theory of depression, connecting self-referential encoding to depressive symptoms and identifying methods of altering it are key endeavors.

Overview of Dissertation Studies

With the above literature as a basis, further exploration of the self-schema and its relationship with behavior becomes all the more important in understanding MDD and depressive symptoms. The research contained in the three papers of this dissertation represent a beginning to this exploration, an attempt to deepen our comprehension of how cognition, behavior, and self-concept might interrelate in depression, and whether there are certain clear entry-points to ameliorating these problems.

The first study explores the SRET, which we continue to conceptualize as a measure of self-schema. While many outcomes have been developed to take advantage of the data collected during the SRET's administration, their relative strength in predicting depressive symptoms has never been analyzed. This study thus identified

which variables of the SRET best predicted depressive symptoms. Secondly, as discussed above, the SRET has been used extensively in the literature. However, it has never been sufficiently explored from a psychometric perspective despite some past studies secondarily reporting test-retest reliability (Dobson & Shaw, 1987; Hayden et al., 2013; Auerbach et al., 2016) and internal reliability (Auerbach et al., 2016). Thus, this study aimed to investigate whether the SRET was internally reliable over multiple samples, and whether its metrics were reliable over time. To accomplish these goals, we collected large samples using online methods, as well as an online field study sample.

The second study investigated the electro-cortical mechanisms at work during the decision-making phase of the SRET. It aimed to identify neurocognitive mechanisms of self-referential judgments made by both healthy control participants and those diagnosed with MDD: Were these decisions happening early or late in the chain of cognitive processing? Did groups differ in magnitude or timing of electro-cortical response? Using a nonparametric technique, point-wise permutation tests, this study analyzed event-related potential (ERP) waveforms collected as participants completed the SRET, and identified specific time frames during which neural activity associated with self-referent cognition differentiated healthy participants and those with MDD.

The third study, in turn, aimed to explore the possibility of modifying maladaptive self-schemas, using the SRET to measure change in self-schema. Participants with elevated levels of depressive symptoms were recruited to complete an online cued imagery training developed for this study. The active training used an imagery technique that focused on positive, future-oriented self-referential events. A neutral control training removed these active elements. Participants completed the SRET three times: before training, mid-way through training, and after training.

They were randomly assigned to either the neutral control training or the active, positive self-reference training. The relative success of the positive training compared to the neutral training was assessed in terms of endorsements on the SRET and depressive symptoms.

These three studies together show a succession of research. First, I evaluated the psychometric properties of a behavioral task that connects depression symptoms with self-referential decisions, and identified the best predictors of depression. Secondly, I identified neurocognitive mechanisms involved in the elaborative processing of self-referential material. Last, I created an intervention to modify such negative self-referent processing, and conducted a pilot study to test whether the intervention altered self-referent processing of negative and positive material. This dissertation thus consists of three studies designed to further understanding of the cognitive biases underlying depression, while also identifying targets for future change driven by therapy or online interventions.

Study 1

Determining optimal parameters of the Self Referent Encoding Task: A large-scale examination of self-referent cognition and depression[1]

Introduction

The cognitive model of depression posits that depressive symptoms are maintained by negatively biased cognition, particularly negative cognition about the self (Beck, 1967). In this model, the concept of the self-schema—an internal representation of the self and the world around oneself—influences what people attend to, how they interpret new information, and what they remember at a later point in time. In depression, these self-schemas tend to be negatively biased, thus prioritizing the processing of incoming negative information. Negatively biased information processing, in turn, is thought to maintain an episode of depression.

Self-schemas are often operationalized by how many positive and negative adjectives people endorse as self-descriptive. The self-referent encoding task (SRET; Derry & Kuiper, 1981) has been used extensively for this purpose (e.g., Goldstein et al., 2015; Dozois & Dobson, 2001; Alloy, Abramson, Murray, Whitehouse, & Hogan, 1997; Prieto, Cole, & Tageson, 1992; Kuiper & Derry, 1982; Davis, 1979). When completing the SRET, participants are asked to make binary decisions (yes/no) about

[1]Dainer-Best, Lee, Shumake, Yeager, & Beevers, in press

whether or not positive and negative adjectives are self-descriptive—a clear corollary of self-schema. In addition to measuring number of word endorsements for positive and negative stimuli, decision-making reaction time and recall of SRET stimuli can also be assessed, providing a variety of metrics relating to the cognitive processing of self-relevant information.

Research in depression has examined a variety of SRET responses and the association of these metrics with depression has been somewhat variable. Faster endorsement of negative adjectives as self-descriptive on the SRET has been associated with depression (e.g., Alloy et al., 1997; MacDonald & Kuiper, 1985), suggesting that reduced reaction time indicates a dominant negative self-schema. However, not all studies find this result; many have indicated no reaction time differences between low- and high-depression groups (Gotlib, Kasch, et al., 2004; Dozois & Dobson, 2001; B. Bradley & Mathews, 1983). Similarly, preferential recall for negative rather than positive words that were previously endorsed as describing the self is also commonly used to measure a negative memory bias (e.g., B. Bradley & Mathews, 1983; Gotlib, Kasch, et al., 2004). Some results suggest that differing levels of depressive symptoms may also impact on recall, endorsement, or reaction time (Derry & Kuiper, 1981; Kuiper & Derry, 1982; Timbremont & Braet, 2004; Timbremont, Braet, Bosmans, & Van Vlierberghe, 2008).

Additionally, some have used a "processing bias" scores to investigate the SRET (e.g., Prieto et al., 1992; Johnson, Joormann, & Gotlib, 2007; Hayden et al., 2013). Positive and negative processing scores are ratios which relate to the number of self-referential words of each valence that are recalled. Prieto et al. (1992) created ratios of endorsed words of each valence to the total numbers of words endorsed, while others created ratios of the number of self-referential words recalled of each valence to the number of words endorsed of that valence (Johnson et al., 2007; Hayden et al.,

2013). However, which of these outcomes from the SRET best and most consistently predicts depressive symptom severity remains unclear.

Importantly, use of a variety of SRET metrics in past research makes it difficult to compare results across studies and, perhaps more importantly, highly flexible processing and scoring of task data has been identified as a leading contributor to inconsistent literatures and non-replications (Wicherts et al., 2016). Thus, an important focus of the current paper is to provide strong evidence for SRET metrics that are most reliably associated with depression severity and could therefore be used consistently in future research. Some past research has attempted to narrow potential predictors from the SRET down to those that were most predictive of depressive symptoms (Disner, Shumake, & Beevers, 2016). In this research, the depressive symptoms of a medium-sized group of adults ($n = 57$) were best predicted by the positive and negative words they endorsed on the SRET. However, this sample size was moderate and the possible predictors were limited. Thus, this study motivates a more thorough study of the SRET.

In addition to the problem that different SRET metrics have been used in past research, the psychometric properties of this task are not well established. There is an increasing focus on the importance of establishing task reliability (Rodebaugh et al., 2016), just as there has been for self-report assessments for many decades. As the SRET has been used in various incarnations over the years, and is used to measure a construct that is central to the cognitive theory of depression, it is important to establish whether the SRET has adequate psychometric properties.

In terms of its psychometric properties, few studies have collected SRET data over multiple time-points, and this has usually been done over extended periods of time—to measure change in self-referent processing but not to determine task reliability. There are nonetheless some indications that the SRET is stable over time.

For example, several studies in relatively large samples of children (ns ¿ 200) indicate that processing bias scores for positive and negative adjectives on the SRET remained stable over time. Hayden et al. (2013) found significant correlations ranging from $r = .24$ to $r = .39$ for SRET processing scores over one year periods, and Goldstein et al. (2015) found significant correlations ranging from $r = .10$ for negative processing to $r = .24$ for positive processing over a three-year period.

Few studies have looked at the SRET longitudinally in older samples. One early study in 24 participants diagnosed with major depressive disorder found that several SRET metrics, including endorsements and reaction time, remained stable over the course of 2-3 weeks in those who remained depressed, while changing when participants remitted from depression (Dobson & Shaw, 1987). However, that study was quite small. Similarly, a small number of studies have examined psychometric properties of the SRET and found that the SRET is stable over the course of both weeks and years, specific to depression versus anxiety, and has high internal consistency (Dobson & Shaw, 1987; Dozois & Dobson, 2001; Auerbach et al., 2016). However, these studies all had relatively small samples (ns < 100) and internal reliability is not routinely assessed with the SRET. Thus, one major goal of this study was to identify the SRET parameters that are most strongly and reliably associated with depression symptom severity and determine whether these components have strong internal consistency and are reliable over time.

Although reaction time measures have been used with the SRET, an important literature has emerged indicating that simple reaction time measures for two-choice decisions may be suboptimal (White, Ratcliff, Vasey, & McKoon, 2010). Rather than relying on reaction time, the diffusion model assumes that information is continuously accumulated (i.e., as individuals process stimuli) until a threshold is hit which results in a response (Voss, Nagler, & Lerche, 2013). The diffusion model

uses reaction time and response data, and their distributions, to draw conclusions about the cognitive processes underlying decisions. For example, the drift rate (v) is estimated using the diffusion model, representing the speed of accumulation of information, i.e., how the buildup of information leading to a decision. This component, as well as others described below, putatively provides a greater level of precision about two choice decision-making than simple reaction time response and may be important for predicting depression symptom severity. Thus, in addition to obtaining a traditional reaction time metric, the current study also examined the utility of applying the diffusion model to SRET responses.

Given the large number of potential parameters that can be derived from the SRET, what is the best way to identify the most reliable predictors of depression severity? Best subsets selection, an automated procedure that evaluates the predictive performance of various combinations of predictors using cross-validation, is ideally suited for this goal. Best subsets regression introduces linear combinations of predictors into a series of regression models, comparing all possible combinations in terms of model fit and identifying which combination of predictors provide the best model fit. Cross-validation procedures aim to identify which predictors best explain variance in *new* data, thus minimizing overfitting to the sample data. To provide an additional test of how well the best cross-validated model generalizes to new samples, before beginning we randomly selected 20% of the data in each sample to serve as a completely independent validation test (i.e., data that were never used at any point during model fitting). This test data is distinct from the data used during cross-validation, which does use all of the data, albeit not all at the same time. Generalizing to new samples is the ultimate goal for most research, and thus we believe this independent validation test is an important feature and strength of the current study.

Self-referent processing has been measured across a variety of populations in past research, and thus we elected to administer the SRET and measures of depression severity to different samples. The first sample was collected online in an undergraduate student sample (the college student sample); the second was collected online using Amazon Mechanical Turk (the MTurk sample); the third sample was from a school field study which included high school adolescents (the adolescent sample). These three samples provide the opportunity to observe whether the optimal parameters are consistent across adolescence and adulthood.

For the relationship between self-schema and depressive symptom severity, we predicted a bias towards negative information would be stronger for participants with elevated depressive symptoms. Given the rigorous approach to predictor selection designed to maximize reliability of results and generalizability to new samples, we expected that similar SRET metrics would be chosen as the best predictors of depression severity across each sample. We also hypothesized that these metrics would have strong reliability, with high internal consistency and strong test re-test reliability over the course of one week.

Methods

Participants

Participants in the college student sample were eligible to participate if they were over the age of 18, were fluent in English, and provided informed consent. Participants in this sample were undergraduates recruited from the University of Texas at Austin psychology subject pool and received course credit for their participation ($n = 527$). Although not all participants disclosed age ($n = 236$ chose not to disclose age), those participants who did reported a narrow age range (18-24; $M = 19.12$, $SD = 1.14$). A slight majority of this sample was female (62.9%) and white (58%).

Participants in the second sample were recruited from Amazon Mechanical Turk (the MTurk sample; $n = 293$). Amazon Mechanical Turk has been used for collecting behavioral data for psychological experiments with positive results (Buhrmester, Kwang, & Gosling, 2011); participants have also been shown to vary in psychopathology symptoms (Shapiro, Chandler, & Mueller, 2013). Data for this study were collected through an adjunctive website, TurkPrime (Litman, Robinson, & Abberbock, 2016), which allowed us to specifically and flexibly recruit participants who were from the United States and were positively-rated respondents on Mechanical Turk. Additionally, participants in this sample were over the age of 18, were fluent in English, and provided informed consent. This sample was in their late thirties (age $M = 37.51$, $SD = 10.7$), 57.9% female, and largely white (79%).

Participants in the third sample participated in an approved program evaluation study conducted in California, requested and approved by the school district and school principal (the adolescent group, $n = 408$). Data were collected over the course of May and July of 2016. Participants were primarily in the ninth grade and on average were 14 years old ($M = 13.99$, $SD = 1.14$); the majority was female and white (see Table 1.1 for full demographic information for all three samples).

Measures

Center for Epidemiologic Studies – Depression Scale (CES-D)

The CES-D (Radloff, 1977) assesses depressive symptoms over the past week using a 20-item self-report questionnaire. Potential scores range from 0 to 60; higher scores indicate more depressive symptoms, with 16 a common cut-off for elevated symptomatology. The CES-D was used in the college student and MTurk samples. Scores on the CES-D can be accurately reflective of both low levels of depressed mood and elevated depressive symptoms.

	College Students ($n = 572$)	MTurk ($n = 293$)	Adolescents ($n = 408$)
Age, mean (SD)	19.12 (1.14)	37.51 (10.7)	13.99 (1.14)
	$n = 336$	$n = 285$	$n = 379$
Female	212 (62.9%)	165 (57.9%)	217 (57.3%)
	$n = 337$	$n = 285$	$n = 379$
Race	$n = 144$	$n = 285$	$n = 373$
White	83 (58%)	225 (79%)	213 (57%)
Black	8 (6%)	23 (8%)	4 (1%)
Indigenous/ Pacific Islander	3 (2%)	0 (0%)	34 (9%)
Asian	24 (17%)	24 (8%)	82 (22%)
Multiracial/Other	26 (18%)	5 (1%)	—
Hispanic/Latino	—	8 (3%)	65 (18%)
Depression, mean (SD)	14.19 (9.71)	13.39 (11.85)	2.56 (3.14)

Table 1.1: Characteristics and symptom profiles of participants in all three samples. As not all participants provided demographic information, sample size (n) is provided for each row. Some participants reported belonging to multiple groups. Ethnicity was not available for the college student sample. Depression scores for college students and MTurk samples come from the Center for Epidemiologic Studies – Depression Scale (CES-D), which ranges from 0 to 60; for adolescents, from the Children's Depression Inventory: Short (CDI:S), which ranges from 0 to 20.

Children's Depression Inventory: Short (CDI:S)

The Children's Depression Inventory (Kovacs, 1981, 1992) is a self-report questionnaire for measuring depression symptoms in children between the ages of 8 and 18. A short version (Kovacs, 2003; Allgaier et al., 2012) has ten items and is sensitive as well as brief (Ahlen & Ghaderi, 2017); total scores range from 0 to 20, with higher scores indicating more depressive symptoms and a recommended cut-off of 12. The CDI:S was used in the adolescent sample.

Self-Referent Encoding Task (SRET)

The SRET (Derry & Kuiper, 1981) is a computer-based task designed to assess schema-related processing. Participants make decisions about whether positive and negative adjectives are self-descriptive. Participants view the words one at a time and make rapid judgments about whether or not each word presented described themselves following word offset. Participants viewed 26 negative and 26 positive words.[2] Words were selected from a well-validated list of positive and negative self-descriptive adjectives (Doost, Moradi, Taghavi, Yule, & Dalgleish, 1999).

For the college student and MTurk samples, the SRET consisted of three blocks; in each block, all 52 words were displayed once in random order. In the adolescent sample, due to time constraints, words were presented only once. Words were displayed in white text on a black screen and remained on-screen until participants responded. Participants were told to use the Q or P keys on their personal keyboard to answer whether the word described them or not. Each trial was followed by a 1,500 ms intertrial interval.

After completing the task, participants were asked to pause for one minute and relax. Then, participants in the college and MTurk samples were given five minutes to recall as many adjectives as possible from the SRET. In the adolescent sample, three minutes were given for free recall due to time constraints.

[2]Positive words were: Happy, Good, Joyful, Proud, Brilliant, Great, Nice, Excited, Pleased, Glad, Excellent, Wonderful, Loved, Fun, Friendly, Helpful, Confident, Fantastic, Cool, Awesome, Best, Content, Free, Playful, Kind, Funny; Negative words were: Alone, Angry, Annoyed, Ashamed, Bad, Depressed, Guilty, Hateful, Horrible, Lonely, Lost, Mad, Nasty, Naughty, Sad, Scared, Silly, Sorry, Stupid, Terrible, Unhappy, Unloved, Unwanted, Upset, Wicked, Worried. In Study 3, "Silly" was replaced by "Dumb".

SRET Metrics

Although the SRET is a simple, straightforward task, it is possible to generate a number of metrics from this task. Commonly used metrics include the number of positive and negative words endorsed as self-descriptive.[3] Individuals with a higher number of negative words endorsed are thought to have a more negative schema; the reverse is true for positive endorsements. Second, the number of positive and negative words that were endorsed as self-descriptive and then recalled is also commonly used to measure self-schema. Recalling more negative words is indicative of a more negative self-schema. Third, some research has examined reaction time (RT) to indicate whether a word is self-descriptive or not. Faster reaction time for negative words endorsed as self-descriptive is thought to reflect a stronger negative self-schema.

Additionally, responses to the SRET can be examined via a computational model known as the diffusion model. Both RTs and responses were used as input for the drift diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 1998; White et al., 2010), a sequential sampling technique that decomposes responses, RTs, and their distributions into distinct components of decision-making and processing. The diffusion model has been used once before with the SRET in a separate, previous study (Disner et al., 2016); it assumes that on each trial, evidence is accumulated until one of two response criteria have been met (i.e., whether a given word is self-descriptive).

The relative ease of evidence accumulation is measured by a component referred to as the drift rate. For the SRET, drift rate (v) can be conceptualized as an index of self-schema. A very positive drift rate to negative words indicates that it is

[3]Some (e.g., Prieto et al., 1992) have used ratios to compare, e.g., the number of positive words endorsed to the total number of words endorsed. Inclusion of such ratios in regression models induces a linear dependency with the intercept, given that they sum to one. Additionally, they reduce the degrees of freedom of the model. As such, although past work has employed them, we could not include them in these models.

easy to categorize such words as self-referential; a drift rate close to zero for positive words indicates that it is difficult to categorize such words; and a strongly negative drift rate reflects that evidence accumulation often leads to rejecting a stimulus as self-referential. Thus, a highly positive drift rate to negative words can be thought of as representing a strong negative self-schema, while a highly positive drift rate to positive words can be thought of as representing a strong positive self-schema. Relative starting point (zr) refers to the degree to which a given valence (positive or negative) may be biased towards one decision (self-referent vs. non-self-referent). Both the relative starting point and the drift rate were calculated separately for positive and negative words.

The diffusion model also estimates the following components: the threshold separation (a), or amount of information needed to make a decision; the response time constant ($t0$); the difference in speed of response execution (d); and the inter-trial variability of drift (sv), inter-trial variability of starting point (szr), and inter-trial variability of non-decisional components ($st0$). The diffusion model's components were computed with the program fast-dm (Voss & Voss, 2007).

Procedure

The Institutional Review Board at the University of Texas at Austin approved all procedures across all samples. Online participants provided informed consent via an online form and did not provide identifying information. Adolescent participants participated as part of a program evaluation study; demographic information was provided from school rosters. Participants were allowed to opt out of the study.

For all three samples, questionnaires presented in Qualtrics (Provo, UT) were used to assess symptoms of depression. Following these measures, participants were automatically directed to a separate website to complete the Self-Referential Encod-

ing Task (SRET). In both the college and MTurk samples, two waves of data collection occurred. In the first wave, participants completed the SRET using Qualtrics and the QRTEngine (Barnhoorn, Haasnoot, Bocanegra, & van Steenbergen, 2015); a second wave completed the SRET using Inquisit software (Millisecond Software LLC, Seattle, WA).[4] Inquisit is a commercial platform that runs "full screen" on participant's computers as a Java applet; the QRTEngine runs in participants' internet browsers or on school computers, within the Qualtrics website. In the college student sample, 144 participants completed the study in Qualtrics and 383 participants completed the experiment in Inquisit. In the MTurk sample, 109 participants completed the experiment in Qualtrics and 184 used Inquisit.

Following the SRET, online participants had five minutes to complete free recall of the words presented during the SRET. Participants were compensated with experiment credit (college students) or $1.75 (MTurk), which is roughly the standard for studies collected on Amazon's Mechanical Turk platform (Mason & Suri, 2012; Buhrmester et al., 2011). Participants did not complete additional measures as part of this experiment.

In the adolescent sample, a self-report survey was administered in Qualtrics to assess depressive symptoms and other internalizing symptoms. Following this, participants were led to complete a series of cognitive task batteries in Inquisit. The SRET was presented at the beginning of this battery and was immediately followed by the recall assessment. Adolescent participants were not compensated.

All MTurk participants who were included following the baseline collection were invited to repeat the procedures a second time one week later; 167 participants'

[4]Analyses were initially performed to determine whether the engine by which the SRET was administered (QRTEngine or Inquisit) significantly altered results. Given that this variable did not significantly affect results, all further analyses were repeated within group but ignoring engine.

data were included at both T1 and T2. These participants were paid slightly more when they completed the second time point ($2.00) than the first time point to encourage completion.

Participant Attrition and Data Filtering

A large number of participants did not complete all aspects of the study or in other ways provided questionable data. In keeping with published guidelines for online data collection (Shapiro et al., 2013), online participants (college and MTurk samples) were excluded prior to conducting inferential statistical analyses for a variety of reasons. Participants who incorrectly answered either of two simple math problems (e.g., "Feel free to use a calculator: What is 7 + 15?") were excluded, as these participants were assumed not to be attending to study materials. Participants in these two studies also completed the Infrequency-Psychopathology Scale of the Minnesota Multiphasic Personality Inventory—2. As described in Shapiro et al. (2013), this instrument measures bizarre beliefs; those who scored more than 3 standard deviations above the group mean were presumed to be inattentive and were excluded. Participants with incomplete data or a variance of zero on the self-report measures (i.e., they responded to every item with the same response) were also excluded.[5]

On the SRET, across all samples, participants who answered more than 10% of trials in under 200 ms were excluded. Those participants with more than 15% of trials deemed outliers for any reason were excluded. Any participants for whom there was no free recall data from the SRET were excluded. Given these exclusions, *n*s reported above are the sample size of the final data; original sample sizes for those

[5]Some items on these measures are reverse-scored; thus, zero variance does not exclude people with a score of zero on the measure.

who completed the study were $n = 676$ for the college student sample, $n = 420$ for the MTurk sample, and $n = 571$ for the adolescent sample. That is, 149 participants were lost from the college student sample, 127 from the MTurk sample, and 163 for the adolescent sample. Although such exclusions are high, they followed *a priori* plans and are representative of much online research (Hauser & Schwarz, 2016; Zhou & Fishbach, 2016). We tested whether depression scores predicted that participants' data being filtered out in each sample separately, entering all participants for whom we had CES-D data into a logistic model that predicted whether data was included by CES-D score. Dropped data in the student sample was linked to CES-D score, $z(675) = -3.01, p = .003$, while this was not the case for the MTurk sample, $z(419) = -1.46, p = .14$, or the adolescent sample, $z(570) = -1.25, p = .21$.

Data from the SRET were filtered for both time-points, as described above. Trials with RTs under 200 ms were dropped, as were trials at least 3 median absolute deviations above individual participant's median RTs (Leys, Klein, Bernard, & Licata, 2013). After dropping trials based on RT, for participants who had multiple blocks of the SRET, the number of words endorsed was calculated as a rounded average across word repetitions, such that although each word was presented three times, it could only be endorsed once. (Although most words were consistently endorsed, words that were endorsed $\frac{1}{3}$ of the time were considered not to have been endorsed, whereas those that were endorsed $\frac{2}{3}$ times were considered to have been endorsed.)

Data Analytic Plan

Three primary analyses were conducted. First, we used best subsets regression with 10-fold cross-validation, repeated under 10 different randomizations, to determine the best SRET predictors of depressive symptom severity (on the CES-D or CDI:S), including word endorsements, reaction time, recall, and the components

of the diffusion model described above. Second, the features from the SRET identified as most strongly correlated with depression symptoms were examined in terms of internal reliability, including Cronbach's alpha for positive and negative items and split-half reliability. Third, repeated-measures analyses were conducted to determine test-retest reliability, using the sub-sample of MTurk participants who completed the task twice over the course of one week.

Due to non-selected samples with a high prevalence of zero or near-zero depressive symptoms, the frequency distributions for CES-D and CDI:S responses were decidedly non-Gaussian and far better characterized as count distributions with overdispersion (data distributions can be seen in Figure 1.1). Therefore, for the best subsets analyses, we modeled depressive symptom severity using negative binomial regression, a generalized linear model (GLM) that fits an additional parameter θ to account for overdispersion. For the Mechanical Turk and adolescent samples, there was also some evidence of zero inflation (more 0 values than would be expected for a negative binomial model), and we explored fitting these data with zero-inflated negative binomial (ZINB) models. While these models achieved somewhat better fits, the decrease in prediction error was not substantial, and the same set of best predictors was selected. Thus, to facilitate comparison with the college student sample (which did not show excessive zero values), we only report here the results of the negative binomial regressions.

Best subset selection using negative binomial GLMs was implemented with the *beset* package (beset; Shumake, 2016), which uses cross-validation to find the subset of predictors that minimizes the mean cross-entropy between the predicted responses $\hat{y}$ (based on models fit to within-fold samples) and the observed responses y (from the corresponding out-of-fold samples). Mean cross entropy is an information-theoretic quantity that is analogous to mean squared error, but instead of averaging

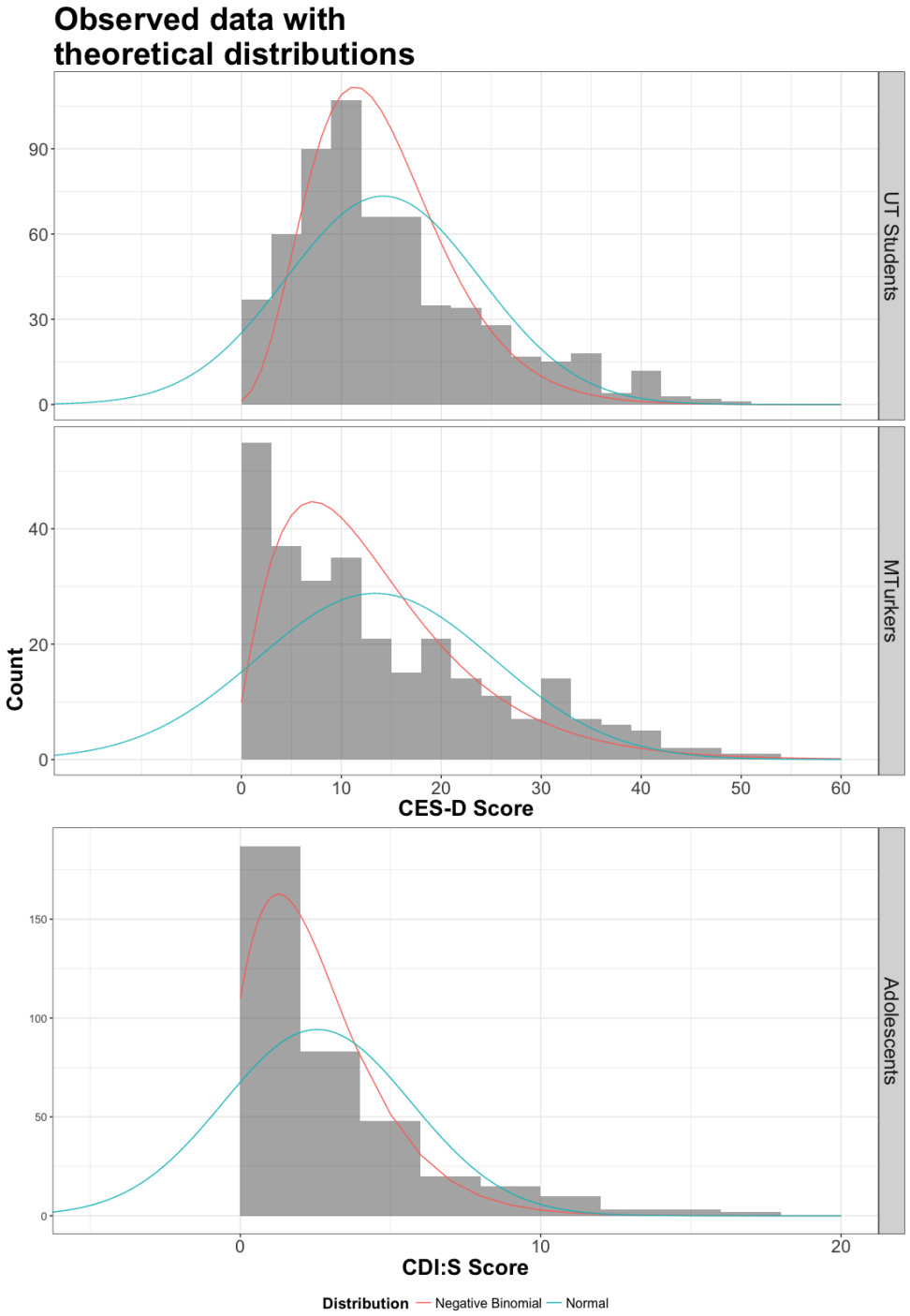

Figure 1.1: Data distributions of depression symptom severity for all three samples. Depression for college student and MTurk samples was measured using the CES-D, which ranges from 0-60; for adolescents, with the CDI:S, which ranges from 0-20. The observed data is plotted as a histogram against density curves from (in blue) the Gaussian (normal) distribution and (in red) the negative binomial distribution. The normal curve used the mean and standard deviation of each sample; the dispersion parameter for the negative binomial distribution corresponds to the estimate of θ from the negative binomial model chosen by best subsets selection; while the mean was the mean of each sample.

$(\hat{y} - y)^2$), it averages the negative logs of the probability density functions for $\hat{y}$ (given the error distribution of the GLM, in this case the negative binomial) evaluated at y. It is important to note that cross entropy simply converts prediction error into a relative likelihood parameter; if applied to the training set instead of the test set, mean cross entropy equals the negative log-likelihood of the model divided by the number of observations. (For logistic regression, this is also known as "log-loss".) Given that GLMs are fit by minimizing the negative log-likelihood (thus maximizing the likelihood), cross entropy defines a unified loss function for both fitting and cross-validating GLMs (Murphy, 2012). The end-goal of this resampling procedure is to find the most reproducible model that will generalize best to new data. Given that we had a large sample size, we randomly set aside 20% of the data from each sample prior to best subset selection to serve as an additional, independent test of how well the best model, chosen by cross-validation, performs on new data. The 20% statistic mirrors the $\frac{1}{5}$ of the data used in each fold of the cross-validation. The data, models, and corresponding analyses script, can be found in the supplementary materials (`https://doi.org/10.18738/T8/XK5PXX` and `https://jdbest.github.io/sretmodels/`).

All available metrics from the SRET, including those from the diffusion model, were included as predictors in these analyses. As such, 19 predictors were available for selection, but models were limited to a maximum of 10 predictors. For both positive and negative valence, these predictors included number of words endorsed as self-descriptive, self-referential recall, total recall, reaction time, drift rate (v), and relative starting point (zr). Additional elements from the diffusion model (a, $t0$ d, sv, szr, and $st0$), described above, were also included.

All participants included in these analyses had complete data for every variable, as well as a measure of depression (CES-D for the college student and MTurk

samples; CDI:S for the adolescent sample). The best models were chosen not for having the absolute lowest cross-validation error, but rather for having the fewest predictors while having a mean cross-validation error that remained within one standard error of the mean of the absolute lowest cross-validation error (Hastie, Tibshirani, & Friedman, 2008). This procedure results in the selection of the most parsimonious model that can achieve predictive performance comparable to that of a more complicated model. To provide a familiar index of goodness of fit and prediction, we calculated an R^2 measure appropriate for negative binomial regression. The traditional R^2 statistic is not appropriate for GLMs that utilize non-Gaussian error distributions, as its interpretation as the fraction of uncertainty explained no longer holds. Cameron and Windmeijer (1997) proposed an R^2 measure based on the ratio of Kullback-Leibler divergence (*entropy* − *cross-entropy*) between the fitted and null models that generalizes this interpretation of R^2 to non-Gaussian GLMs.

Given that entropy is equivalent to the negative log-likelihood of the saturated model, this R^2 can be calculated as $1 - \frac{dev}{nulldev}$, where dev is the deviance of the fitted model and nulldev is the deviance for the intercept-only model. We therefore abbreviate this statistic as R^2_D and refer to it as "deviance explained" to make clear the distinction between this and the "variance explained" R^2 from ordinary least squares regression. In addition to the model-fit R^2_D, which describes how well the models fit the training data, we also calculated a predictive R^2_D, using the cross-entropies computed for the withheld folds during cross-validation. This statistic indicates the fraction of uncertainty in new data that the model is expected to explain. As we also obtained 100 estimates of the predictive R^2_D from the cross-validation procedure (10 folds × 10 resamples), we further resampled these results 1,000 times to obtain a bootstrap estimate of the 95% confidence interval for the mean predictive R^2_D.

Beyond computing diffusion model components, data cleaning, simple modeling, and visualization were conducted in RStudio (version 1.0.136) running R (version 3.3.2) with the following packages: *dplyr* (Wickham, Francois, Henry, & Müller, 2015), *tidyr* (Wickham, 2015), *ggplot2* (Wickham, 2009), *lme4* (Bates, Mächler, Bolker, & Walker, 2015), and *psych* (Revelle, 2016). In addition, best subset selection with repeated cross-validation was performed using a version of R (3.2.1) compiled for high-performance computing (HPC) with the Wrangler data analysis system at the Texas Advanced Computing Center (TACC), using our own R package, *beset* (Shumake, 2016), to parallelize model resampling.

Results

Summary statistics for behavioral data and outcomes

Self-report data from the CES-D and CDI:S are presented along with basic demographic information per sample in Table 1.1. The college student ($M = 14.19$; $SD = 9.71$) and MTurk ($M = 13.39$; $SD = 11.85$) samples had, on average, mildly elevated depression symptoms on the CES-D, although symptom severity ranged mild to severe. Similarly, the adolescent sample on average had scores on the CDI:S indicating mild symptoms of depression but there were cases across the range of depression symptom severity ($M = 2.56$; $SD = 3.14$). Descriptive data for the SRET metrics are presented in Table 1.2, alongside components from the diffusion model. As a whole, participants endorsed more positive words than negative, $t(2485.5) = 62.55$, $p < .001$, Cohen's $d = 2.45$, 95% CI $[2.31, 2.61]$, and recalled more self-referential positive words, $t(2152.4) = 32.53$, $p < .001$, Cohen's $d = 1.36$, 95% CI [1.26, 1.47]. They also had a positive drift rate to positive words (ranging from 1.50 to 1.56) and a negative drift rate to negative words (ranging from -1.56 to -2.06), indicating that, on average, participants easily rated positive words as self-referent

and negative words as non-self-referent. Correlations for each sample, between all predictors of interest, are included in the supplementary materials (Tables S1-3). Many of these variables were strongly correlated with each other and depression severity.[6]

	College Students ($n = 572$)	MTurk ($n = 293$)	Adolescent ($n = 408$)
Positive Endorsements	20.38 (7.09)	20.00 (8.12)	19.64 (5.55)
Negative Endorsements	5.41 (5.36)	5.24 (6.41)	3.30 (4.88)
Positive, Self-Referential Words Recalled	7.39 (3.42)	5.23 (3.15)	5.07 (2.76)
Negative, Self-Referential Words Recalled	2.81 (2.91)	2.06 (2.55)	0.78 (1.52)
Positive Words Recalled	9.35 (3.51)	6.78 (3.47)	6.13 (2.87)
Negative Words Recalled	9.59 (3.60)	7.10 (3.50)	4.74 (3.42)
RT to Positive Words (ms)	829.14 (229.99)	708.33 (172.74)	923.16 (275.01)
RT to Negative Words (ms)	852.60 (220.56)	719.42 (171.97)	938.19 (264.99)
Drift Rate (v) to Positive Words	1.56 (1.71)	1.79 (2.11)	1.50 (1.37)
Drift Rate (v) to Negative Words	-1.56 (1.39)	-2.06 (1.82)	-1.85 (1.41)
Relative Starting Point (zr) for Positive Words	0.57 (0.10)	0.56 (0.12)	0.61 (0.14)
Relative Starting Point (zr) for Negative Words	0.47 (0.12)	0.43 (0.13)	0.44 (0.13)
Threshold Separation (a)	1.33 (0.33)	1.22 (0.34)	1.59 (0.41)
Response Time Constant ($t0$)	0.521 (0.083)	0.508 (0.091)	0.591 (0.130)
Differences in Speed of Response Execution (d)	0.011 (0.034)	0.015 (0.029)	0.012 (0.049)
Inter-Trial Variability of Starting Point (szr)	0.287 (0.118)	0.297 (0.101)	0.268 (0.112)
Inter-Trial Variability of Drift (sv)	0.572 (0.208)	0.595 (0.245)	0.547 (0.230)
Inter-Trial Variability of Non-Decisional Components ($st0$)	0.228 (0.107)	0.183 (0.089)	0.251 (0.140)
Percentage of Contaminants (p)	0.198 (0.214)	0.163 (0.170)	0.191 (0.202)

Table 1.2: Behavioral data for each sample's primary outcomes from the SRET, including outcomes from the drift diffusion model. Standard deviations included in parentheses.

[6] An interactive Shiny interface to explore the relationships between these variables can be viewed online at https://jdbest.shinyapps.io/shiny-comparisons/

Best subsets analyses

For each model, at each number of predictors, the standardized, average cross-entropy error for the train data, cross-validation data, and test data are presented in Figure 1.2. Errors are standardized for each sample by dividing by the null model's cross-entropy error for that data type, such that the standardized error for the null model is equal to 1. Across all models, standardized cross-entropy error for the training data (in gray) decreases quickly after the first two to three predictors and then continues to decrease gradually as more predictors are added to the model. For the standardized cross-validation cross-entropy error (in red), a similar pattern is observed for the college and MTurk samples, except that standardized cross-entropy error becomes relatively stable after the third predictor rather than continuing to decrease. For the adolescent model, the standardized cross-validation cross-entropy error (in red) begins to slightly increase after the fifth predictor variable. Last, for the independent test dataset (i.e., 20% of data set aside and not used to train the models), standardized cross-entropy error (in blue) mirrors the training and cross-validation error for the college student sample. For the MTurk sample, change in error for the test dataset improved initially, but started to increase for models with more than two predictors. For the adolescent sample, there was a clear improvement in test cross-entropy error for models with 1, 3, and 4 predictors, and then test cross-entropy error began to increase with more predictors in the model.

Nonetheless, prediction error improved substantially between the null model and the model with two predictors, and continued to improve with three predictors for the college student sample, and with four predictors for the MTurk and adolescent samples. Hence, three- and four-predictor models were selected as the best model for each sample, as these were the most parsimonious models that had a mean cross-validation error within one standard error of the mean of the model with the absolute

lowest cross-validation error.

The best model for the college student sample included three predictors: the number of negative words endorsed ($\beta = 0.05$, $SE = 0.005$), drift rate (v) to positive words ($\beta = -0.13$, $SE = 0.02$), and inter-trial variability of starting point (szr; $\beta = -0.46$, $SE = 0.19$). This model explained 45% and 43% of the deviance in depressive symptom severity in the training and test samples, respectively. The mean cross-validated R^2_D was 0.44, 95% CI [0.42, 0.46].

The best model for the MTurk sample included four predictors: the number of negative words endorsed ($\beta = 0.03$, $SE = 0.01$), drift rate (v) to negative words ($\beta = 0.13$, $SE = 0.07$), drift rate (v) to positive words ($\beta = -0.11$, $SE = 0.04$), and inter-trial variability of non-decisional components ($st0$; $\beta = 1.44$, $SE = 0.57$). This model explained 43% and 29% of the deviance in depressive symptom severity in the training and test samples, respectively. The mean cross-validated R^2_D was 0.41, 95% CI [0.40, 0.43].

Last, the best model for the adolescent sample included four predictors: drift rate (v) to positive words ($\beta = -0.24$, $SE = 0.06$), drift rate (v) to negative words ($\beta = 0.43$, $SE = 0.07$), relative starting point (zr) for negative words ($\beta = 1.97$, $SE = 0.47$), and threshold separation (a; $\beta = -0.52$, $SE = 0.16$). This model explained 43% and 41% of the deviance in depressive symptom severity for the training and test samples, respectively. The mean cross-validated R^2_D was 0.40, 95% CI [0.36, 0.43].

In order to compare the consistency of the predictors chosen by the best models, we compared R^2_D across samples for the best models described above. Given that R^2_D is affected by the number of parameters, we compared the best models with four predictors—the minimum required for the MTurk and adolescent samples—for all three samples; the best model with four predictors for the college students

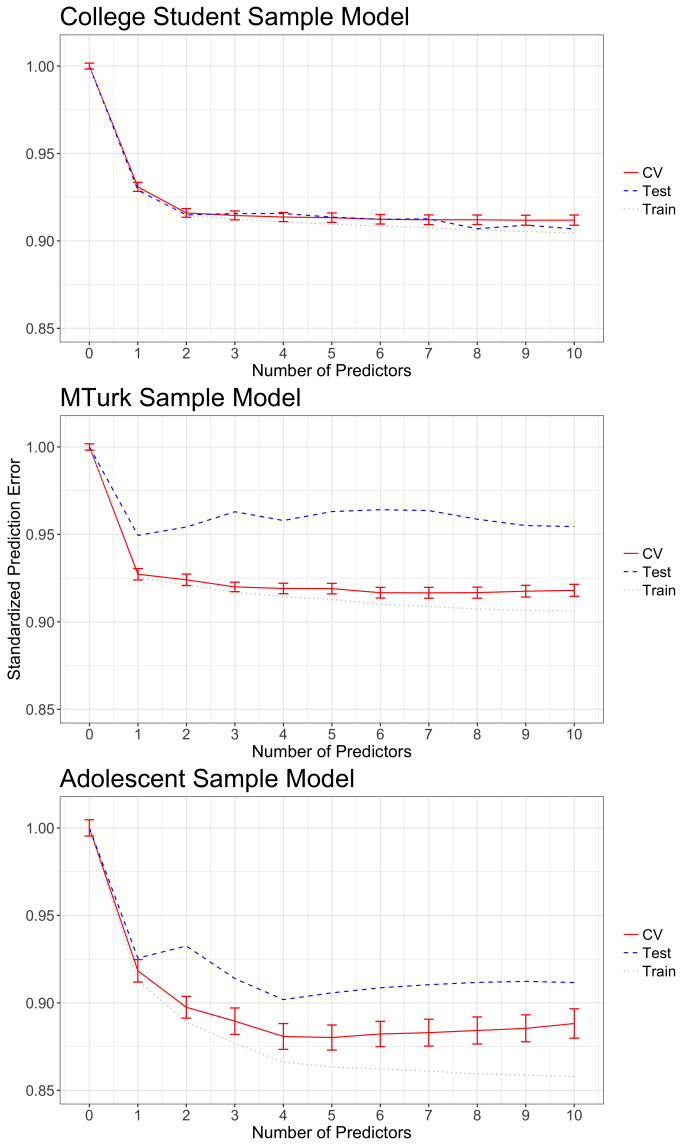

Figure 1.2: Standardized cross-entropy errors, as a function of increasing the number of predictors for each model. Cross entropy is an index of the discrepancy between model predictions and observed data. Errors are standardized to the null model (i.e., 0 predictors) for each data type, so that the error for the null model is equal to 1. The dotted gray line shows the standardized cross-entropy error for the training data ("train"), which is guaranteed to always decrease as the number of predictors increase. (However, this decrease may be due to the model picking up on random chance patterns, and does not necessarily translate into better predictions with new data.) Solid red lines show the standardized mean cross entropy across 10 repetitions of 10-fold cross-validation ("CV"), with the standard error of the mean indicated by error bars. Models were chosen based on CV error. Dashed blue lines show the standardized cross-entropy error for an independent test data set ("test"), a pseudo-randomly chosen 20% subset of each sample. For the college student sample, all three error curves are highly similar, likely because this sample has the largest sample size ($n = 527$). With the other, smaller samples (MTurk $n = 293$, adolescent $n = 408$), there are widening discrepancies between the training error vs. the cross-validation and test error as the number of predictors increases. These models indicate that prediction error improves substantially relative to the null, intercept-only model (0 predictors) with the addition of between 2 and 4 predictors, and then appears relatively constant up to 10 predictors. The 3- and 4-predictor models were chosen for each sample, as the simplest model within one standard error of the minimum of each curve.

included the three predictors described above, and the relative starting point (zr) for negative words. The predictors chosen by the MTurk model resulted in $R^2_D = 0.45$ in the college students, which is only 0.009 less than the best model. The predictors chosen by the adolescent model resulted in $R^2_D = 0.44$ in the college students, which is 0.015 less than the best model. The predictors chosen by the college student model resulted in $R^2_D = 0.41$ in the MTurk sample, which is 0.02 less than the best model. The predictors chosen by the adolescent model resulted in $R^2_D = 0.42$ in MTurk sample, which is 0.009 less than the best model. The predictors chosen by the college student model resulted in $R^2_D = 0.38$ in the adolescent sample, which is 0.05 less than the best model. The predictors chosen by the MTurk model resulted in $R^2_D = 0.38$ in adolescent sample, which is 0.05 less than the best model. That these models at worst indicated 0.05 less deviance explained is indicative of a strong degree of consistency. (The worst models chosen with four predictors had R^2_D of less than 0.001.)

Given that many models had comparable prediction errors with mean cross entropy error within the 95% confidence intervals of the best model's R^2_D, and given the high degree of correlation between many of the predictors, Figure 1.3 shows all possible two-variable models for each sample (more than two variables would have made this figure unwieldy). Squares for each comparison are shaded by degree of cross-entropy error, with lower errors receiving darker shading. The top and right of each plot has predictors relating to positive valence; the bottom and left of each plot has predictors relating to negative valence. Predictors in the middle come from the diffusion model. These plots demonstrate how well each combination of all possible two variable models predicts depression symptom severity.

Of note are the number of negative and positive words endorsed, and the drift rate to positive and negative words; these variables were strongly predictive of

depression symptom severity in any combination. Self-referential recall of negative words only—not positive words—was often a strong predictor of depression severity with a second predictor, although it was chosen in none of the best models. Also important to note is that although several non-valenced components of the diffusion model were included in the three- and four-predictor models; in these two-predictor plots, such components are only predictive of depression in conjunction with valenced predictors.

Comparing targeted models

The best subsets procedures chose endorsements and components of the diffusion model as the best predictors of depression severity and did not choose predictors relating to recall or reaction time. Given our interest in determining the best predictors that would be sufficient to predict depressive symptom severity, we targeted models to compare against one another using cross-validated R^2_D with bootstrapped 95% confidence intervals as a measure of predicted deviance explained. Thus, we were able to compare in a "head to head" fashion whether specific predictors were markedly better than others. Specifically, we were interested in comparing how well positive and negative word endorsements alone compared to the recall of positive and negative words, and to the drift rate (v) for positive and negative words. In particular, collecting recall data and reaction time data (necessary for calculating drift rate and the diffusion model) adds experimental time and burden to both researchers and participants. We were therefore interested in evaluating whether this added cost translates to a worthwhile gain in terms of predicting depression severity.

The model with just positive and negative word endorsements explained between 31% and 42% of the cross-validated deviance, cross-validated $R^2_D = .42$, 95% CI [.40, .43] for the college student sample; $R^2_D = .35$, 95% CI [.33, .37] for the

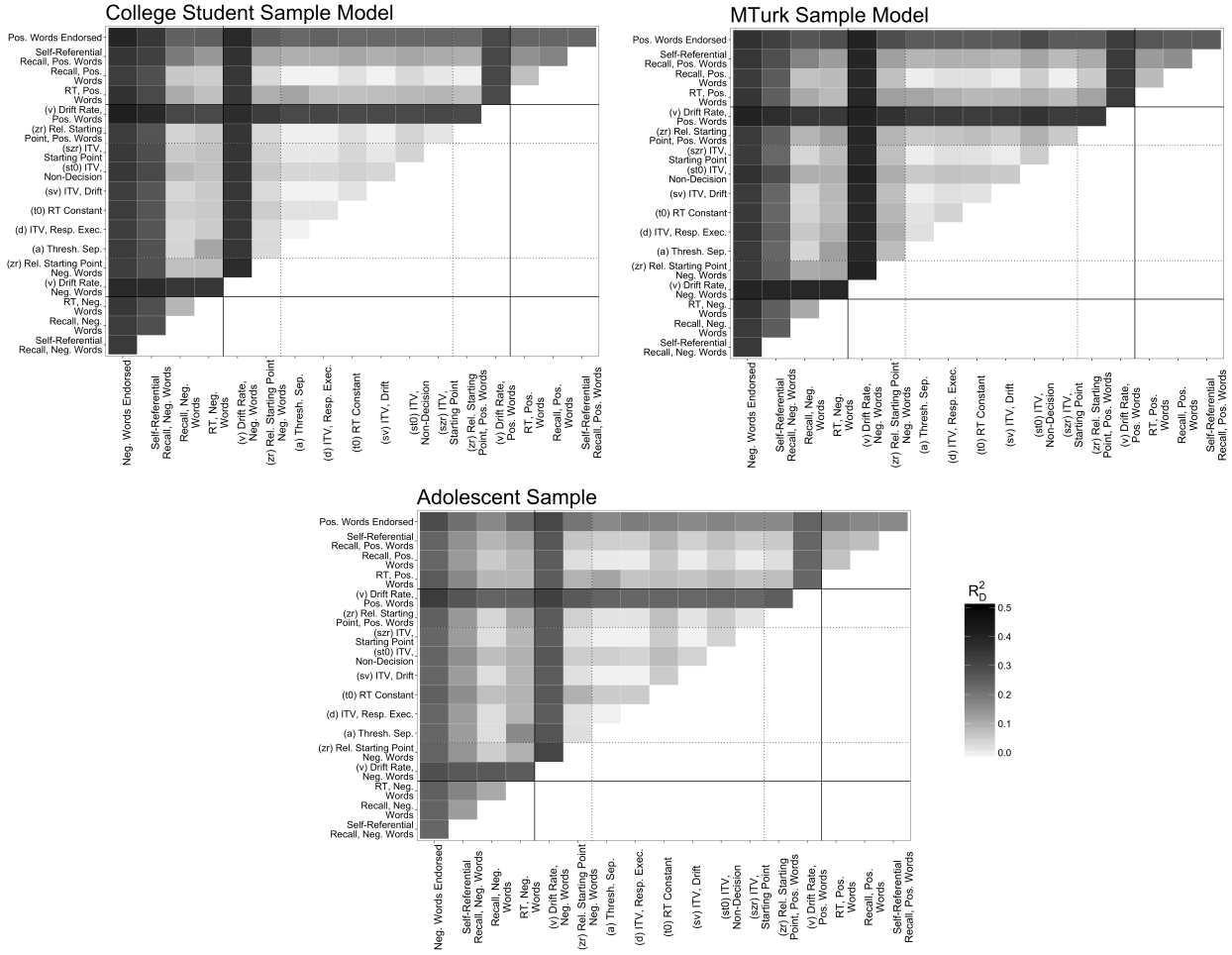

Figure 1.3: Deviance explained (R^2_D) for each possible two-predictor model predicting depression for each sample. Darker shading indicates a model that explains more of the deviance in depression symptoms, and therefore better fits the data—thus the best (highest R^2_D) two-predictor model for each sample is solid black, whereas the worst model is very light gray. Predictors are arranged such that the top and right variables are positively valenced, while the bottom and left variables are negatively valenced. Solid lines divide the elements of the drift diffusion model from the purely behavioral measures of the SRET; dotted lines separate the elements of the diffusion model that are not specific to words of one valence. Predictors that are consistently chosen for two-variable models are highlighted by virtue of having bars of solidly dark shading.

MTurk sample; and R^2_D = .31, 95% CI [.29, .33] for the adolescent sample. In comparison, the model including drift rate (v) to both positive and negative words explained between 36% and 40% of the cross-validated deviance: cross-validated R^2_D = .40, 95% CI [.38, .42] for the college student sample; R^2_D = .39, 95% CI [.37, .41] for the MTurk sample; and R^2_D = .36, 95% CI [.29, .37] for the adolescent sample. Thus, the calculation of drift rate metrics did not result in a substantial improvement in prediction: gains in predictive accuracy were marginal at best for the MTurk and adolescent samples, and non-existent for the college sample. Models using only self-referential recall explained between 16% and 35% of the cross-validated deviance: cross-validated R^2_D = .35, 95% CI [.33, .36] for the college student sample; R^2_D = .27, 95% CI [.13, .29] for the MTurk sample; and R^2_D = .16, 95% CI [.14, .18] for the adolescent sample. Thus, recall metrics did not just fail to improve predictive performance—they were markedly worse.

SRET internal consistency

For each sample (and for both time-points for the MTurk sample), Cronbach's alpha for all participants was calculated using the *psych* package in R (Revelle, 2016). Word endorsements and reaction times (RTs) were present for every participant, for every trial, and thus alpha was calculated for each of these, for positive and negative trials separately. Confidence intervals on alpha were bootstrapped using 10,000 iterations.

Cronbach's alpha across each sample for endorsements of positive words was strong and ranged from .93, 95% CI [.91, .94] to .97, 95% CI [.97, .98]. Similarly strong internal reliability was observed for endorsements of negative words, as alpha ranged from .91, 95% CI [.90, .92] to .94, 95% CI [.93, .95]. Inter-item correlations were somewhat higher for endorsements of positive items than for negative, ranging

from $r = .34$ to $r = .56$ for positive words, and from $r = .27$ to $r = .40$ for negative words. The item "silly", which had been considered a negative word, was correlated close to zero with the other items. Given the low loading, this item was replaced for the adolescent sample. (Refer to footnote 2.) However, no other item failed to correlate with the positive or negative items, indicating strong internal reliability of endorsements. Tables of inter-item correlations for each sample can be seen in the supplementary materials (`https://jdbest.github.io/sretmodels/SRET_reliability.html`).

Cronbach's alpha across all samples for RTs to positive words ranged from .95, 95% CI [.94, .95] to .98, 95% CI [.98, .99] across samples, and for RTs to negative words from .94 95% CI [.93, .95] to .98, 95% CI [.98, .99]. Inter-item reaction time correlations were roughly equivalent, and ranged from $r = .40$ to $r = .69$ for RTs to positive words, and from $r = .38$ to $r = .70$ for RTs to negative words. These high correlations and extremely high alphas indicate strong internal reliability of reaction times.

Split-half task reliability was calculated for all predictors chosen in the best subsets procedures, which included the number of positive and negative endorsed words, the drift rate (v) to positive and negative words, the relative starting point (zr) for negative words, the inter-trial variability of starting point (szr), the inter-trial variability of non-decisional components ($st0$), and the threshold separation (a). Reliability was calculated in two ways for each variable. First, we correlated the predictor of interest on even vs. odd trials across the whole task for each sample. Second, for the college student and MTurk samples, for which there were three blocks of the SRET, we correlated the predictor of interest on the first block to the predictor of interest on the third block. Because of limits on the efficacy of the diffusion model's components with too few trials, these processed scores were only calculated

for the college student and MTurk samples. Pearson's correlation coefficient (r) was calculated for each comparison and 95% confidence intervals on r were calculated by using 10,000 bootstrapped iterations.

For endorsements, correlations were high across samples for the odd-even comparison, ranging from r = .87 95% CI [.84, .90] to r = .98, 95% CI [.97, .99] for positive words and from r = .85, 95% CI [.80, .88] to r = .95, 95% CI [.93, .96] for negative words. Correlations were similarly high for the split-third comparisons, ranging from r = .87, 95% CI [.84, .89] to r = .93, 95% CI [.90, .96] for positive words and from r = .86, 95% CI [.83, .88] to r = .95, 95% CI [.91, .97] for negative words.

For drift rate to positive words, odd-even correlations ranged from r = .88, 95% CI [.86, .90] to r = .91, 95% CI [.88, .93] and split-third comparisons ranged from r = .79, 95% CI [.76, .82] to r = .86, 95% CI [.82, .90]. For drift rate to negative words, correlations were similarly high; odd-even correlations ranged from r = .80, 95% CI [.78, .83] to r = .86, 95% CI [.83, .89] and split-third comparisons ranged from r = .70, 95% CI [.65, .74] to r = .80, 95% CI [.73, .86]. For relative starting point (zr) for negative words, correlations were consistent although not as high; odd-even correlations ranged from r = .39, 95% CI [.32, .46] to r = .50, 95% CI [.38, .60] and split-third correlations ranged from r = .29, 95% CI [.22, .37] to r = .46, 95% CI [.32, .58]. That is, with the exception of zr, behavior on valenced measures was consistent between split portions of the task.

For the inter-trial variability of starting point (szr), chosen as the best predictor in the college students model, correlations were close to zero, ranging from r = .07, 95% CI [-.02, .15] to r = .12, 95% CI [-.07, .30] for odd-even comparisons and from r = .08, 95% CI [-.001, .17] to r = .21, 95% CI [.06, .35] for split thirds. For the inter-trial variability of non-decisional components ($st0$), odd-even trial correlations

ranged from $r = .40$, 95% CI [.31, .49] to $r = .53$, 95% CI [.37, .67], while for split thirds, correlations ranged from $r = .15$, 95% CI [.05, .25] to $r = .37$, 95% CI [.21, .50]. Last, for the threshold separation (a), correlations were high; odd-even correlations ranged across samples from $r = .79$, 95% CI [.75, .82] to $r = .82$, 95% CI [.68, .90] and split third correlations ranged from $r = .59$, 95% CI [.51, .67] to $r = .69$, 95% CI [.59, .77]. These metrics from the diffusion model exhibited relatively low internal reliability.

SRET test-retest reliability

Mixed-effects regression models with fixed effects of time and valence, a random intercept per participant, and a random slope per valence, were calculated for all valenced predictors chosen in the best subsets procedures. These models were run with participants from the MTurk sample who had completed two time-points, using data that passed verification checks at both time-points ($n = 167$).

We examined test-retest reliability for the number of positive and negative endorsed words, the drift rate (v) to positive and negative words, and the relative starting point (zr) for positive and negative words. None of the interactions between valence and time were significant ($|t|$s for drift rate and relative starting point < 1; t for endorsements $= -1.35$; all ps $\geq .18$); for models repeated without the interaction, time was not significantly predictive ($|t|$s for drift rate and relative starting point < 1; t for endorsements $= -1.10$, all ps $> .27$), indicating that these SRET metrics were stable (i.e., not significantly different) across time. We ran paired sample t-tests on all predictors chosen in the best subsets procedures, all $|t|$s < 0.6 and all ps $> .45$.

We next correlated the data from T1 and T2 across all predictors chosen in the best subsets procedures: the number of positive and negative endorsed words, the drift rate (v) to positive and negative words, the relative starting point (zr)

for negative words, the inter-trial variability of starting point (szr), the inter-trial variability of non-decisional components ($st0$), and the threshold separation (a). We calculated 95% confidence intervals on Pearson's r using 10,000 bootstrapped iterations.

For number of words endorsed, one-week test re-test correlations were high, with $r = .87$, 95% CI [.80, .93] for positive words and $r = .88$, 95% CI [.83, .93] for negative words. For drift rate (v), correlations were also high, with $r = .83$, 95% CI [.78, .88] for positive and $r = .80$ 95% CI [.72, .86] for negative. For relative starting point (zr) for negative words, the correlation was relatively low, $r = .37$, 95% CI [.24, .50]. For inter-trial variability of starting point (szr), the correlation was negative, $r = -.13$, 95% CI [-.29, .03]. For inter-trial variability of non-decisional components ($st0$), the correlation was moderate, $r = .41$, 95% CI [.21, .60]. Lastly, for threshold separation (a), the correlation was moderate, $r = .56$, 95% CI [.44, .65].

These analyses indicate between moderate and very strong test re-test reliability over a one-week period, with the exception of the inter-trial variability of starting point, which was poorly correlated between time points.

Discussion

These findings support a strong relationship between self-schema and depressive symptoms. A best subsets analysis shows that many of the aspects of the self-referent encoding task (SRET) are strongly linked with depressive symptom severity; this linkage was consistent across three separate samples. Analyses indicated that models with either three or four predictors, each using at minimum one positive valence and one negative valence predictor, were sufficient for explaining between 29% and 43% of the deviance in an independent test sample, and between 43% and

45% of the deviance in the training dataset; cross-validated 95% confidence intervals ranged from 36% to 43%.

Endorsements and drift rate were the most robust predictors of depressive symptom severity. (Drift rate measures ease of categorizing words of each valence as self-referential.) Throughout samples, the best models typically included at least one positive valence variable and one negative variable valence from endorsements and drift rate. Figure 1.3 clearly indicates that these variables were strongly associated with depression severity across samples. That these variables were highlighted consistently in all permutations indicates that they provided additional information beyond other variables. Importantly, even though drift rate is partially derived from whether or not a person endorses a word as self-referent, these variables explained unique variance in depression severity (or they both would not have been identified as best predictors).

There was a great deal of overlap in the predictors chosen by the best subsets regression in this study to those selected in previous work (Disner et al., 2016), and more broadly to those used in the research literature. Several findings in particular are of interest. First, the drift diffusion model (Ratcliff, 1978; Ratcliff & Rouder, 1998), which incorporates response data and reaction times, was important in these analyses. Of particular note is the drift rate, the speed of accumulation of information towards making a decision as to whether positive or negative words were self-referent. This predictor was included in all three of the best models, for all three samples, providing additional evidence for the utility of the diffusion model in clinical research (White et al., 2010). Further, the best model for the adolescent sample included only elements of the diffusion model; no other historically utilized outcomes from the SRET were included. It is possible that the brief forms of both the SRET and the depression severity index used in the adolescent sample resulted in a higher

percentage of floor responses, thus privileging a metric that takes a combination of response and reaction time into account. Importantly, traditional RT methods were very poor predictors of depression severity, indicating that while RT is not problematic *per se*, how RT is measured and processed appears to be very important.

It is, however, important to note that when we compared models including only drift rate to positive and negative words with models including only endorsements of positive and negative words, these models were comparable in their proportion of deviance explained. This indicates that the models including components of the diffusion model may not be substantially better than those without. The effort and time required to compute the components of the diffusion model should be weighed against the possible gain that results from its output. Second, although the recall of both positive and negative words previously endorsed as self-referential was considered a good predictor by the model selection procedures (refer to Figure 1.3), the number of positive or negative words recalled on their own was not strongly associated with depression severity, as recall only models explained relatively little deviance. We believe that the relative predictive power of self-referential recall may be a direct corollary of the number of positive or negative words endorsed. That is, in a sample where many participants endorsed no negative words, the only participants who can have a non-zero self-referential recall of negative words are those who already endorse negative words. With the exception of studies directly investigating memory bias, collecting recall information may not be useful when administering the SRET. It is important to note that certain predictors may in fact measure somewhat different constructs—i.e., memory bias on the SRET stems from biased self-reference but may be different from drift rate.

Third, although two-predictor models were excellent at predicting depression, best subsets analyses nonetheless chose models that included an additional,

non-valenced parameter. These parameters (in the college student sample, inter-trial variability of starting point [szr]; in the MTurk sample, inter-trial variability of non-decisional components [$st0$]; and in the adolescent sample, the threshold separation [a]) relate to inter-trial variability and to reaction times. That is, the valenced parameters already selected for each model contain sufficient information to differentiate between participants based on processing of positive and negative valence; the best method to further explain residual deviance was to use methods that measure differential reaction speeds or differential patterns of response (Iacoviello et al., 2014; but note Kruijt, Field, & Fox, 2016). In Figure 1.3, it is clear that these parameters do not have significant impact on their own, but only in conjunction with the other variables selected previously in the models. And, indeed, $st0$ is correlated with depression in the MTurk sample, and both $st0$ and a are strongly correlated with reaction times to both positive and negative words in the MTurk and adolescent samples (see supplementary correlation tables). (In the college student sample, szr was not correlated with any other variables. It additionally had low internal reliability.)

As a whole, these results were remarkably consistent across all three samples (college students, MTurk adults, and adolescents), providing a set of findings that increases confidence in the conclusions of this work. The endorsements and drift rate from the SRET also showed strong psychometric properties, including high internal consistency and high test-retest reliability. This reliability remained high in all three samples, and both online and in-person, indicating its utility in a variety of contexts. And, indeed, the predictors chosen by the best models for each sample generalized relatively well to the other two samples, as well as to a left-out percentage of each sample. That is, the predictors themselves were very consistent across samples, even as there were slight differences in the models chosen as "best". However, some metrics from the diffusion model, including relative starting point (zr) and inter-trial

variability of starting point (szr), were less consistent despite being selected in best models.

The statistical methods we used in this study also bear further discussion for their relative strengths. Our procedures, including both cross-validated regression and reserving a subset of our data for testing models, represent the most rigorous of scientific recommendations (Munafò et al., 2017; Begley & Ioannidis, 2015). By using these techniques and likewise by running iterative bootstrapped samples, we are able to provide not just estimates of confidence intervals, but also identify models that are consistent across multiple permutations of the data.

Although this study was not limited in size, it was somewhat limited in scope; participants were not screened beyond the symptom measures for other potential diagnoses. Nor were over-sampling procedures employed to make use of clinically depressed participants, although a number of participants in all three samples did endorse elevated depressive symptoms. However, many studies have utilized the SRET with nonclinical samples (e.g., Goldstein et al., 2015; Auerbach et al., 2016; Hayden et al., 2013), implying its effectiveness as a dimensional measure of depression. Additionally, measures of socioeconomic status were not collected and thus cannot be reported. However, given the three different samples (i.e., high school students, college students, and adults who participated through Amazon Mechanical Turk), it is known that there were differing levels of education. The samples were, however, largely white, female, and somewhat young in age. Future work may recruit targeted populations to increase diversity of SRET data in terms of race, sex, diagnostic information, and age. Future work may also choose to use measures of depression with multiple dimensions, including subscales beyond those evident in the CDI:S (Ahlen & Ghaderi, 2017).

Some past work has used a mood induction (Teasdale, 1988; Teasdale & Rus-

sell, 1983) before administering the SRET, based on the theory that induced sad mood may activate a latent negative self-schema. Indeed, some research has suggested that inducing sad mood may differentiate between those with a vulnerability to depression and those without (Kelvin, Goodyer, Teasdale, & Brechin, 1999; Taylor & Ingram, 1999). Such an induction has been occasionally used throughout the literature (e.g., Timbremont & Braet, 2004; Timbremont et al., 2008). This relationship between induced sad mood and depressive self-referential processing is theoretically rooted in cognitive vulnerability to depression, rather than current symptomatology; as such, it was not employed in the current studies. The inability to diagnose this sample, and the lack of information about past depressive episodes and cognitive vulnerability for depression, does limit the results. However, use of mood inductions may yet be useful in research investigating risk factors for depression.

There have been some suggestions that depressed individuals have more generally negative self-schema than others with disorders of negative affect, including anxiety (M. S. Greenberg & Beck, 1989). However, many studies have suggested that in dysphoria and depression, not only are negative schemas amplified but positive schemas are also attenuated (Walker, Skowronski, & Thompson, 2003). The findings from the SRET confirm this, with the strongly negative relationship, for example, between number of endorsed positive words and CES-D score. This "positive blockade" (Beck, 1967) in depression invites further investigation.

This study provides further evidence for the link across levels of analysis, between self-reference, self-schema, and depression symptoms. Further, it highlights the importance of measuring both positive and negative self-schema, and using them in conjunction when investigating depression, which is often thought of as a disorder focused on negative cognitive biases. Future work using the SRET should consider using variables most closely associated with depressive symptomatology: the number

of negative and positive words endorsed and the drift rate from the diffusion model. Conversely, we do not recommend the use of traditional reaction time and, with the exception of research explicitly focused on memory bias, future research may choose to drop measures of recall while still retaining sufficient information from endorsements and drift rate to accurately predict levels of depressive symptoms.

When possible or useful, given its demonstrated efficacy in our results, we do recommend the decomposition of available information possible from the diffusion model. Such a threshold may be met (a) when research collects sufficient numbers of trials per participant (i.e., 40 at a minimum, as recommended by Voss et al., 2013), (b) when there is a theoretical use for more nuanced disintegration of response behavior, and (c) when there is a theoretical utility to be provided by a predictor, i.e. the drift rate, which incorporates all available information. The choice of which outcomes from the SRET to use in future research is, of course, also dependent on the theoretical questions being asked.

We hope that this work encourages additional psychometric research with the SRET and other assessments of cognitive bias in depression and across other forms of psychopathology. Deepening our understanding of cognitive biases and their relationship with the maintenance of depression is an important next step in developing more targeted treatments. Findings from the current study indicate that how the person views themselves and the ease with which they make those determinations may be integral to understanding—and possibly changing—symptoms of depression among adolescents and adults.

Acknowledgments

The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed

to the research results reported within this paper. URL: http://www.tacc.utexas.edu. Collection of the MTurk sample was supported by the College of Liberal Arts at The University of Texas at Austin. The field study was funded by the Hope Lab.

Study 2

Sustained Engagement of Attention is Associated with Increased Negative Self-Referent Processing in Major Depressive Disorder[7]

Introduction

The way that one views oneself, one's self-concept or self-schema, is intricately tied with mood. Beck's (1967) cognitive model of depression postulates that the schema—internal beliefs and knowledge about the self, the world, and the future—influences how life events are appraised and interpreted. Schemas also prioritize the processing of incoming information, such that environmental stimuli that are consistent with one's self-schema are attended to, processed, and subsequently recalled more readily (Segal, 1988). A negative self-schema may result in biased interpretation of ambiguous stimuli, or cause elaborative processing of over-attended stimuli (Everaert, Koster, & Derakshan, 2012). This in turn has been theorized to facilitate increased recall of negative stimuli, resulting in negatively biased memory. Although other mechanisms clearly also contribute to the maintenance of depression (e.g., emotional blunting), negative self-schema may be a criticalan important mechanism that fuels many of the negative cognitive biases observed in depression and contributes to the maintenance of thought to maintain depression.

Consistent with the cognitive model, individuals with Major Depressive Dis-

[7]Dainer-Best, Trujillo, Schnyer, & Beevers, 2017

order (MDD) have been shown to display negatively biased attention, interpretation, and memory (Everaert et al., 2012). Further, depressed people often do not display protective positive cognitive biases that are observed in healthy individuals (Walker et al., 2003; Disner et al., 2011). Major depression instead privileges negative processing and, as a result, individuals with MDD are likely to view themselves as having more negative and fewer positive characteristics than non-depressed individuals.

One method of measuring self-schema is the self-referent encoding task (SRET; Derry & Kuiper, 1981). The SRET is a binary-choice, affective decision-making task combined with incidental recall of SRET stimuli. The SRET is generally a computer-based task, where positive and negative adjectives are presented one at a time to participants who determine as quickly as possible whether each word is self-descriptive or not. Following presentation of the word stimuli, participants are then asked to recall as many of the SRET stimuli as possible. Studies have shown strong correlations between endorsement of negative (but not positive) words and depressive symptoms (e.g., Disner et al., 2016); increased endorsement of negative words (and decreased endorsement of positive words) on the SRET is also predictive of depression symptom course (Disner et al., 2016; Connolly et al., 2015). Responses on the SRET also appear to be consistent over time particularly when depression symptoms remain relatively stable (Goldstein et al., 2015; Auerbach et al., 2016).

These studies provide a clear link between negative self-referent cognition, as measured by the SRET, and depressive symptoms. Understanding the neural architecture of negative self-referent bias is important, as it could provide a more comprehensive understanding of this important cognitive bias and point to translational treatment targets for neurocognitive interventions.

In the current project, we used electroencephalography (EEG) to measure the temporal characteristics of cognitive processes involved in self-appraisal. Although

the spatial resolution of EEG is not ideal, EEG is extremely effective at measuring information about the time course of cognitive phenomena (Kappenman & Luck, 2012). Thus, collecting event-related potentials (ERPs) during the SRET can provide information as to whether biased self-referent processing is occurring at an early processing level; whether it occurs at a later level of cognitive evaluation; or whether both processes contribute to this negative self-referent processing bias.

Early ERP processes include the P1 and P2, both positive deflections in an ERP waveform thought to reflect automatic processing of attentional information that may nonetheless be influenced by emotion (Hajcak, Weinberg, MacNamara, & Foti, 2012; Delplanque, Lavoie, Hot, Silvert, & Sequeira, 2004). These peaks occur between 100 and 300 ms following a stimulus. The P2 in particular may index post-perceptual selective attention, as it occurs late enough (peaking approximately 180 ms after stimulus onset) to be related to the association of new information with prior comprehension (Hajcak et al., 2012; Luck & Hillyard, 1994). As both occur early, they are understood to be related to early attentional engagement; both are typically increased when attending to emotional stimuli (Delplanque et al., 2004).

The late positive potential (LPP), conversely, is a component often considered an index of cognitive evaluation and engagement with stimuli. The LPP begins around 300ms post-stimulus and continues up to 1,500ms (Hajcak et al., 2012; Cuthbert, Schupp, Bradley, Birbaumer, & Lang, 2000). The LPP is often either posterior or central in localization (Hajcak et al., 2012). The LPP increases in positive amplitude in response to prioritization of information, indicating increased engagement, especially to negative information. Schupp and colleagues (2004) demonstrated that the LPP is greater in response to unpleasant or negative images as compared to neutral or positive images, regardless of mood state; others have shown that attending to non-arousing images reduces the LPP (Hajcak, MacNamara, Foti, Ferri, &

Keil, 2013). Moreover, some studies have shown a generally diminished LPP in participants with MDD (Blackburn, Roxborough, Muir, Glabus, & Blackwood, 1990; Proudfit, Bress, Foti, Kujawa, & Klein, 2015; Weinberg, Perlman, Kotov, & Hajcak, 2016).

Several studies have attempted to use EEG to identify the key ERP components that contribute to negative self-referent processing during the SRET. A prior study found that both current and remitted MDD groups had increased amplitudes for negative stimuli in an early component of attentional capture (the P2) in comparison to healthy controls (Shestyuk & Deldin, 2010). They also found that individuals who were currently depressed showed more positive amplitudes in the late positive potential (LPP) for negative stimuli than the other groups. This suggests that MDD participants selectively attended to negative information (due to the initial P2 amplitude difference from controls) and were engaging in increased cognitive evaluation of negative information (due to the increases in the LPP compared to healthy controls).

Similarly, in a sample of depressed and healthy female adolescents, depressed girls were shown to exhibit greater early (P1) amplitudes in response to negative words, and greater later (LPP) amplitudes to negative words (Auerbach et al., 2015). These findings are consistent with the results of prior work, with the MDD group showing early attention to negative words that continues over the time-course of the ERP.

Further work with a large sample of younger female participants (N = 121) found indications that risk for depression (i.e., maternal history of MDD) was also associated with greater LPP amplitudes to negative words when compared to those at low risk for MDD (Speed et al., 2016). This study used a principle components analysis (PCA), which builds components from the EEG electrode channels that most strongly contribute to an outcome. With the LPP described by the PCA, there

was no difference between positive and negative valence within groups; however, in response to negative words only, the at-risk participants showed increased LPP amplitudes and increased subsequent recall of negative stimuli compared to positive. This study did not find differences in the earlier waveforms (P1 or P2). An additional study investigated depressive response on the SRET from a semantic processing perspective (i.e., the N400 waveform), arguing that a diminished N400 suggests stronger self-reference (Kiang et al., 2017). This study demonstrated that participants with MDD had a diminished N400 in response to negative, but not positive, adjectives.

A recent SRET study using PCA techniques with ERPs in a large community sample of adults (N = 128) found that individuals with elevated depressive symptoms had enhanced negativity to both positive and negative words in early frontal regions (Waters & Tucker, 2016). A waveform that they believed to be an element of the late positive complex or P300, at a similar time frame to the LPP, was attenuated in response to all stimuli in parietal regions. Notably, these findings are in the opposite direction to the results reported above (e.g., the LPP measured at a similar point in time was increased in depressed female adolescents in Auerbach et al., 2015).

In summary, these studies reveal some conflicting results in terms of the waveforms associated with the SRET, raising a question of whether there is attenuation or augmentation of the early selective attention components (P1, P2) and later cognitive evaluation (LPP) in response to negative stimuli in depressed participants relative to healthy controls. Many of the above-reviewed studies were conducted in young, female participants; it is important to determine which of these findings, if any, extend to adult samples. Further, relatively few studies have been completed in a sample with a clinical diagnosis of MDD. Additionally, given the recent emphasis in psychology to replicate novel research (e.g., Munafò et al., 2017), this study's potential to independently replicate prior work in this area is important.

In the current study, we anticipated that behavioral results would follow in the same vein as previous work, with more endorsements of negative words as self-referent in participants diagnosed with MDD compared to healthy controls. Based on prior research and the cognitive model of depression, we predicted that adults with MDD, in comparison to healthy controls, would show early, differential attention between negative and positive words in the P1 and P2. We also predicted that MDD-diagnosed participants, as compared to healthy controls, would show increased cognitive evaluation in later components (similar to the LPP) for negative stimuli. Were this confirmed, it would imply that differential processing of self-referential information results from both the early components involved in perception and selective attention of negative stimuli, and also from the way that these stimuli are elaborated, processed, and encoded.

To better assess the full span of attentional processing in response to word presentations, we conducted analyses using a non-parametric technique often applied to functional neuroimaging analyses (Nichols & Holmes, 2002), which identified spatiotemporal areas that might be strongly differentiating between the MDD and HC groups during self-referential processing. This technique, discussed further below, uses randomized permutations of the data to conduct point-by-point t-tests, correcting for multiple comparisons, which allows spatiotemporal areas with strong differences to rise to the forefront. This data-driven method identifies the onset of differential between-group responses in a manner that is conservative compared to standard parametric approaches because, first, no assumptions of normality are required and, second, no *a priori* (and possibly biased) choices of time window or electrode region are necessary. This is in contrast to the traditional, parametric approach to ERP analysis, which examines activity within a limited number of electrode sites averaged across specified time windows. Nevertheless, we also conducted

a limited number of parametric analyses to allow for easier comparisons of results from past work and the current study.

Method

Participants

Adults were recruited from the Austin, TX community through the use of advertisements posted on websites (Craigslist, a UT-Austin message board, and Indeed) and fliers. Postings described a study on "mood" and "emotional experiences" but highlighted the need for both healthy and depressed subjects. The advertisements directed participants to a website to determine eligibility. At this website, participants provided informed consent for the screening and filled out a brief survey of current mood and demographics. Research assistants and graduate students trained on diagnostic interviewing conducted phone screenings on eligible participants, using the Mini International Neuropsychiatric Interview (version 6.0, Sheehan et al., 1998). The MINI is a standardized instrument used for brief screenings to diagnose a variety of psychiatric disorders. Research assistants took part in a training workshop during which they learned interview skills, role-played interviews, and reviewed diagnostic criteria. After the workshop, they listened to calls conducted by experienced researchers and had their initial screening interviews monitored for fidelity. Phone calls were audio-recorded with consent from participants throughout the study for fidelity analyses. An independent assessor (J.D.B.) randomly selected and rated 20% of MDD and HC interviews. Agreement for MDD diagnosis between study interviewers and the independent assessor was excellent ($k = 1.00, p < .0001$).

Inclusion and Exclusion Criteria

Participants were included in the current study if they were between the ages of 18 and 55 and spoke fluent English. Participants who scored less than 13 or greater than 16 on the Center for Epidemiologic Studies – Depression Scale (CES-D) during online screening were invited to complete the MINI interview over the phone. Participants with CES-D > 16 who met DSM-5 criteria for MDD ($N = 22$; American Psychiatric Association, 2013) were included in the study. Nine participants met criteria for a current single major depressive episode, while 12 met criteria for recurrent MDD. Participants with MDD were included whether or not they met criteria for a current anxiety disorder; for all participants in this group, MDD was the primary diagnosis as assessed during the MINI. Of those diagnosed with MDD, 12 met criteria for one or more DSM-IV anxiety or trauma-based disorders (seven for Generalized Anxiety Disorder, two for Social Anxiety Disorder, five for panic attacks, and four for PTSD).

Healthy control participants had a CES-D < 13 and did not meet diagnostic criteria for past or current MDD or a current anxiety disorder ($N = 24$). Importantly, all participants (HC and MDD) were excluded from the study if they met diagnostic criteria for the following disorders: current alcohol or substance abuse or dependence; mania or hypomania; bipolar disorder; or psychosis. Participants were also excluded based on criteria that could affect EEG collection, including a history of seizures or epilepsy, head trauma, current use of beta-blockers, and current use of anti-psychotic drugs (Keil et al., 2014).

[8]One participant in this group identified as agender.

[9]This participant reported long-term medication usage, but otherwise met inclusion criteria. Analyses were conducted without this participant and showed no significant differences.

	HC ($N = 23$)	MDD ($N = 21$)
Age, mean (SD)	25.3 (7.9)	24.6 (6.7)
Female	15 (65%)	16 (76%)[8]
White	13 (57%)	10 (48%)
Hispanic	6 (26%)	10 (48%)
Psychiatric Medication		
None	22	15
Current medication usage	1[9]	6
Current SSRI, for > 10 weeks	1	4
Other antidepressant	0	2
Anti-anxiety medication	0	1
CES-D, mean (SD)	4.9 (5.2)	34.7 (7.93)
MASQ, Anxious Arousal subscale	11.3 (2.0)	19.6 (6.1)
MASQ, General Distress subscale	13.9 (4.2)	32.3 (8.3)
MASQ, Anhedonic Depression subscale	24.2 (6.4)	40.5 (5.6)

Table 2.1: Characteristics of participants included as healthy controls (HC) and diagnosed as depressed (MDD).

Sample Characteristics

Data from one participant in each condition were excluded because of poor EEG data quality. Demographics are reported in Table 2.2.1.1; participants were in their mid-twenties, mostly female, and approximately half were white (53% white, 28% Hispanic/Latino, 16% Asian, and 2% Black). Groups did not differ significantly on the basis of age ($t[41.82] = -0.33$, $p = .74$), gender ($\chi^2[2] = 2.30, p = .32$), or race ($\chi^2[2] = 0.77, p = .85$). Groups did of course significantly differ on depression severity, as measured by the CES-D, $t(31.6) = -14.0, p < .001$..

Power Analysis

A priori power analyses were conducted to determine sample size necessary to achieve a medium-to-large effect size ($\eta^2 = 0.22$), as per Auerbach et al. (2015) for

multivariate ANOVA; with the current sample size, we would achieve 80% power.

Procedure

The institutional review board at the University of Texas at Austin provided approval for the study (IRB # 2014-08-0078). Participants who were eligible for the study following phone screening were scheduled for a 1.5-hour session in the lab. Informed consent was obtained, and participants completed several self-report questionnaires. Following this, Easycap EEG caps were placed on the participant's head and prepared for EEG collection, and then participants completed experimental tasks on a computer. Participants were paid $20 for completing the study and were provided with a list of local mental health resources.

Measures

Center for Epidemiologic Studies – Depression Scale (CES-D)

The CES-D (Radloff, 1977) is a self-report scale designed to assess depressive symptoms over the past week using 20 items. Scores may range from 0 to 60; a score greater than 16 is often used as a cut-off for elevated depressive symptoms (Radloff, 1977; Santor, Zuroff, Ramsay, Cervantes, & Palacios, 1995). The CES-D was used for screening and was assessed again during the laboratory visit when it was confirmed that the CES-D remained above 16 for participants with MDD and below 13 for healthy controls.

Mood and Anxiety Symptoms Questionnaire (30-item version; MASQ)

The MASQ (Clark & Watson, 1991) is a 90-item self-report scale designed to measure the tripartite model of depression, anxiety, and general distress. A short (30-item) adaptation of this questionnaire has been developed, which maps closely

onto the original questionnaire (Wardenaar et al., 2010), and was used in the present study. The MASQ provides sub-scales of General Distress, Anhedonic Depression, and Anxious Arousal. The Anxious Arousal subscale was used as a covariate in a final set of analyses to determine if anxiety symptoms altered observed findings.

Self-Referent Encoding Task (SRET)

The SRET (Derry & Kuiper, 1981, see Figure 2.1) is an affective decision-making task where participants make binary-choice decisions about whether positive and negative words are self-descriptive. Participants view the words on a computer screen and make rapid judgments following the word's display. In this version, participants viewed 40 negative and 40 positive words (as in Auerbach et al., 2015[10]) selected from the Affective Norms for English Words (M. M. Bradley & Lang, 2010) for a total of 80 trials.

Stimuli were presented in random order for 200 ms, followed by 1,800 ms of a fixation cross. Since subjects were instructed to withhold their motor response during this fixation cross, event-related potentials to word presentation and decision-making remained uncontaminated by motor responses. Only after offset of the fixation cross were participants presented with the question prompt, 'Does this word describe you?" Participants used a Logitech gaming controller's shoulder buttons to respond "yes" or "no". Although the 1800ms period between the stimulus offset

[10]The following 80 words were included in the SRET: Positive words were: admired, adorable, alive, beautiful, bold, bright, capable, carefree, confident, cute, devoted, dignified, elated, engaged, famous, festive, friendly, gentle, grateful, happy, honest, hopeful, inspired, jolly, joyful, lively, loyal, lucky, masterful, outstanding, proud, satisfied, silly, surprised, terrific, thoughtful, untroubled, useful, vigorous, wise. Negative words were: afraid, alone, angry, anguished, bored, brutal, burdened, cruel, crushed, defeated, depressed, disgusted, disloyal, displeased, distressed, dreadful, fearful, frustrated, guilty, helpless, hostile, insane, insecure, lonely, lost, morbid, obnoxious, rejected, rude, scared, shamed, sinful, stupid, terrible, terrified, troubled, unhappy, upset, useless, violent.

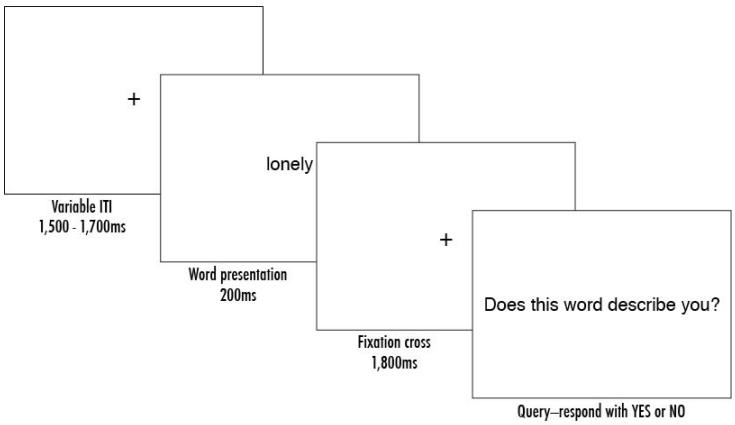

Figure 2.1: The SRET (Self-Referent Encoding Task). Event-related potential epochs are based on the moment of word presentation.

and behavioral response allows for the recording of neural response to stimuli, this extended period renders the reaction time response less meaningful and somewhat difficult to interpret.[11] Thus, we primarily focus on ERP responses to word stimuli rather than reaction time in our analyses, which is consistent with prior work in this area (e.g., Auerbach et al., 2015). Participants completed several neutral practice trials before the task began, to ensure that they knew to wait to respond until the question appeared. A jittered intertrial interval followed each trial, between 1,500 and 1,700 ms in length.

After completing the task, participants completed an image-based task for approximately 12 minutes. Following this distraction, they were asked to recall as many adjectives as possible from those presented during the SRET, within five min-

[11]Indeed, we tested whether group and valence predicted reaction times, and found no significant effect.

utes. Participants were not previously informed that they would be asked to perform this recall task. The primary behavioral outcome from the SRET is the number of positive and negative words endorsed as self-referential. An alternative method (see e.g., Goldstein et al., 2015) involves calculating processing bias scores, with a negative score calculated as $\frac{number\ of\ negative\ words\ endorsed}{number\ of\ any\ words\ endorsed}$ and the reverse for positive valence. Thus, individuals with an increased number of negative words endorsed, or a negative processing bias closer to one, have a stronger negative processing bias.

EEG Recording and Data Analysis

EEG was recorded using a 64-channel active electrode system placed in the Easycap recording cap and recorded with the BrainVision actiCHamp amplifier and PyCorder software. In addition to the 64 cap channels, an additional four channels were collected to track vertical and horizontal eye movements. All head channels were located based on extended 10/20 system locations; cap sizes were chosen based on the circumference of participant's heads. Electrode impedances were reduced using an electrical conducting gel, to below 10 kΩ. Continuous EEG were sampled at 500 Hz, initially referenced to Cz. Offline, data were processed using BrainVision Analyzer 2.0 software where the data were re-referenced to an average reference of all head channels.

Electrooculogram (EOG) channels were created by subtracting active electrodes placed below the eyes from above-eye (Fp1 and Fp2) sites for vertical EOG, and by subtracting sites placed outside of the left and right canthi of the eyes for horizontal EOG. A Butterworth infinite impulse response filter was applied to bandpass filter the data from 0.1 to 30 Hz (slope of 12 dB/oct), and sections with major artifacts identified by visual inspection were marked and excluded from analysis. Several participants ($N = 6$) had one or two faulty electrodes; these participants had the

faulty channel(s) interpolated using a linear triangulation algorithm before further analyses were conducted. An independent component analysis (ICA) transform was then conducted in order to identify and remove the effects of eye blinks and eye movements, using both vertical and horizontal EOG channels. Each participant's whole dataset was used to calculate the ICA matrix, and a restricted Infomax rotation was used to decompose the ICA and remove components relating primarily to eye blinks and eye movements (Jung et al., 2000; Lee, Girolami, & Sejnowski, 1999).

Individual trials were split into 1,500 ms epochs selected from 200 ms before stimulus onset to 1,300 ms following stimulus onset. Epochs were created separately for positively- and negatively-valenced words. Following the creation of these epochs, intervals were further artifact assessed using the following semi-automated criteria based on Auerbach et al. (2015): a maximal voltage step of 50μV/ms; a voltage difference above 300 μV within an epoch; amplitudes above 200 μV or below -200 μV; and periods longer than 100 ms with activity under 0.5 μV. Epochs that did not pass these parameters were rejected from further analysis. A linear de-trend[12] was then applied to the data based on the 100 ms before stimulus onset and the 100 ms at the end of the epoch, and a baseline correction was applied by averaging the period from 200 ms preceding stimulus onset until that onset.

Average responses were created per-participant, for negative and positive words separately. A minimum of 20 valid epochs were required per participant, per each valence; all participants had at least this many. On average, 36 ($SD = 2.9$) epochs per participant were included for negative word stimuli trials, and 36.1

[12]Given the possibility that a linear de-trend would alter differences in longer lasting, late ERP components, all analyses were repeated without the de-trend. Results were not substantially different except in one test indicated in the text, below; nor did the visual inspection of waveforms reveal substantial differences.

($SD = 2.7$) epochs were included for positive word stimuli trials. Difference waves were generated by subtracting positive from negative word stimuli trials.

The data were examined using pointwise non-parametric randomized permutation t-tests, which were corrected for multiple comparisons across time and site (Trujillo, Allen, Schnyer, & Peterson, 2010; Sanguinetti, Trujillo, Schnyer, Allen, & Peterson, 2015; Nichols & Holmes, 2002; Pernet, Latinus, Nichols, & Rousselet, 2015). This cluster-based method allows for the identification of differential responses between MDD and HC groups in a manner that is more conservative than standard t-tests, avoids putative *a priori* choices of regions of analysis by utilizing the full scalp recorded data. This analysis was performed on the difference waves (negative words minus positive words) in a between-subjects analysis.

This non-parametric statistical method consists of a three-step process to create an empirical null distribution for the between-group ERP difference waves to be used for hypothesis testing the between-group differences. First, we computed a statistical significance threshold for the between-group difference waves at electrode and time point. As the ERP responses were recorded from 62 channels over a 1.5 sec (1,500 ms) epoch at a 500 Hz sampling rate, this amounted to a total of $62 \times 1.5 \times 500 = 46{,}500$ independent thresholds. We determined these thresholds by computing a distribution of 20,000 between-group t-statistics for each data point under the null hypothesis. Each t-statistic was computed after exchanging (permuting) the data of a randomized subset of participants in each group (the size of each subset equaled the number of individuals in the smaller of the two subject groups, i.e., $N = 22$). If the null hypothesis is true and there are no between-group differences, then the t-value computed after this exchange is still an element of the null distribution (because exchanging subjects across groups should make no difference if there truly are no between-group differences). This process was repeated 20,000 times to create

a distribution of t-values from which we determined the two-tailed $p = 0.05$ threshold for each of the 46,500 data points.

Because such a large number of independent tests will inflate type-I error, it was necessary to correct for multiple comparisons. We accomplished this in a second step by using these significance thresholds to determine contiguous t-statistic clusters across electrodes and time points. We then computed the distribution of maximal t-statistic clusters under the null hypothesis. This was accomplished by computing a second round of 20,000 between-group data permutations, where during each permutation new t-values were computed for each data point. Those t-values that were above the $p = 0.05$ thresholds determined in the first step of this procedure were then divided into contiguous clusters. Data was arranged into a three-dimensional structure (anterior-posterior electrode dimension, left-right electrode dimension, time dimension) and t-statistic clusters were defined as three-dimensional neighborhoods of contiguous points (26-connected point neighborhoods). Then, for each identified cluster, we computed its exceedance mass, defined as the summed total of t-values within the cluster (i.e. "the integral of the statistic image above the primary threshold within the suprathreshold cluster"; Nichols & Holmes, 2002, p. 8). We then selected the largest exceedance mass for a given permutation, yielding a distribution of 20,000 maximal exceedance mass values under the null hypothesis.

In a final step, we used the previously-obtained null distribution of maximal cluster exceedance masses to hypothesis test the cluster exceedance masses observed in the non-permuted data. Cluster exceedance masses calculated from the non-permuted data with sizes greater than the null distribution's $p = 0.05$ criterion exceedance mass were considered to be significant at the two-tailed level with strong control for type-I error. All p-values were corrected for multiple comparisons using a step-down procedure (A. P. Holmes, Blair, Watson, & Ford, 1996). As each cluster

corresponded to a spatiotemporal extent of between-group ERP differences, this method allowed us to simultaneously identify where on the scalp and when in time those differences were statistically significant.

In an effort to compare our findings with prior work, event-related potential (ERP) components were also calculated as the mean area under the curve for relevant electrode sites and time windows. Consistent with previous studies (e.g., Auerbach et al., 2015), the P1, P2, and the early late positive potential (LPP), components were calculated using averages across Pz, POz, P1, P2, PO3, and PO4. The P1 was quantified as the mean average from 100 to 200 ms following the stimulus; the P2, from 200 to 300 ms; and the early LPP, from 400 to 600 ms. The late LPP was calculated as the average of Fz, FCz, and Cz, from 600 to 1,200 ms following the stimulus.

Behavioral data cleaning, modeling, and visualization was conducted in RStudio (version 1.0.136) running R (version 3.3.2) with the following packages: dplyr (Wickham et al., 2015), tidyr (Wickham, 2015), ggplot2 (Wickham, 2009), lme4 (Bates et al., 2015), lmerTest (Kuznetsova, Brockhoff, & Christensen, 2016), and compute.es (Re, 2013). EEG data were prepared in BrainVision Analyzer 2.0 and analyzed in MATLAB via in-house scripts.

Results

Summary statistics and analyses for behavioral data

Behavioral SRET data are summarized in Table 2.2. Two-way mixed effects ANOVAs were conducted with factors of group and valence. For processing bias, the group $\times$ valence interaction was significant, $F(1, 40) = 138.8, p < .001$, generalized $\eta^2 = .78$, with the MDD group having a greater negative processing bias than the HC group, $t(27.3) = -11.21, p < .001$, Cohen's $d = -3.47$, 95% CI [-4.46, -2.48]. Given

that processing bias is a ratio, the test for positive processing bias return inverse results with opposite signs, and is thus not repeated here.

For each valence, participants could endorsed between 0 and 40 words. In the HC group, participants endorsed from 0-8 negative words, and 12-36 positive words. In the MDD group, participants endorsed 2-30 negative words, and 8-27 positive words. In the two-way mixed effect ANOVA, for word endorsement, the group $\times$ valence interaction was significant, $F(1, 40) = 115.3, p < .001$, generalized $\eta^2 = .62$, with the MDD group endorsing more negative words than the HC group, $t(20.9) = -10.08, p < .001$, Cohen's $d = 3.12$, 95% CI [2.19, 4.06], and the HC group endorsing more positive words than the MDD group, $t(40.0) = 5.90, p < .001$, Cohen's $d = 1.83$, 95% CI [1.08, 2.57]. Within the MDD group, there was a near-significant effect of valence, with more negative words endorsed than positive, $t(33.3) = 2.02, p = .051$, Cohen's $d = 0.66$, 95% CI [-0.02, 1.33]. Within the HC group, there was a significant effect of valence, with significantly more positive words endorsed than negative, $t(27.3) = -16.8, p < .001$, Cohen's $d = 4.94$, 95% CI [3.75, 6.14].

Lastly, for RTs, the group $\times$ valence interaction was not significant, $F(1, 40) = 0.00008, p = .99$, generalized $\eta^2 < .001$. Given the delay between presentation and requested response, this lack of significance is unsurprising.

Non-parametric statistical analysis

Grand averages were generated for HC and MDD participants separately and used to create topographic maps (see Figure 2.2). The topographic maps were used to visualize the scalp distribution of the group by valence differences, and how they change over time. As temporal interval increases, consistent frontal and posterior differences develop between groups, with the MDD group demonstrating greater dif-

	HC		MDD	
	M	*SD*	*M*	*SD*
Positive Processing Bias	0.92	0.091	0.45	0.16
Negative Processing Bias	0.084	0.091	0.55	0.16
Positive Words Endorsed	25.30	6.27	14.84	5.23
Negative Words Endorsed	2.09	2.19	18.89	6.99
Positive Recall	9.78	3.81	7.79	2.86
Negative Recall	6.00	3.19	8.63	3.62
Self-Referential Positive Recall	6.74	3.26	3.32	1.92
Self-Referential Negative Recall	0.78	1.00	4.37	2.61
Positive RT	409.5	183	506.4	274
Negative RT	400.9	169	497.7	272

Table 2.2: Behavioral data for the Self-Referential Encoding Task. Number of positive and negative words endorsed are sums; processing biases are ratios calculated as the number of positive/negative words endorsed over the total number of words endorsed. Recall is the number of words of that valence recalled; self-referential recall only includes words that were endorsed during the task. HC = healthy control; MDD = Major Depressive Disorder; RT = reaction time.

ferences than healthy control participants in responses to negative versus positive words. Central differences were especially apparent from 600 – 1,000 ms, indicating periods where the HC group's responses to negative minus positive words were increased compared to the MDD group. Thus, these topographic maps indicated heightened differential activation at central sites during a time window that is consistent with the LPP. Early time periods show relatively minor differences between groups, primarily in frontocentral regions.

We computed permutation tests using negative word stimuli minus positive word stimuli difference waves, comparing between groups (HC subtracted from MDD). The permutation tests were performed over all electrode sites following data processing, using an interval from 0 – 1,000 ms post-stimulus, thus encompassing the

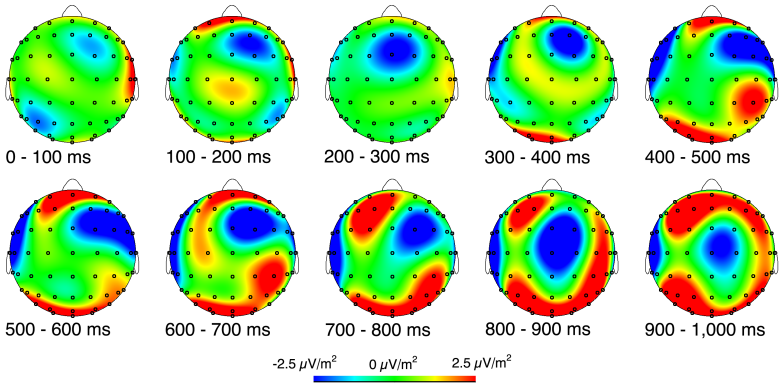

Figure 2.2: Topographies from 0 to 1,000 ms, for negative minus positive, for HC minus MDD participants. Red shading indicates where the HC group showed a greater difference between negative and positive adjectives in a given spatiotemporal area; blue shading indicates where the MDD group showed a greater difference. HC = healthy control; MDD = Major Depressive Disorder.

waveforms that were significantly different between groups in previous work described above, including P1, P2, and the LPP.

The results of the permutation tests are displayed in Figures 2.3 and 2.4. The upper portion of Figure 2.3 depicts the ERPs averaged across all electrodes that showed statistically significant differences between groups. The lower portion of Figure 2.3 indicates the electrodes and time periods where the difference waveform for negative vs. positive word stimuli was significantly different between MDD and HC groups. Darker colors, concentrated at left-frontal and central sites, indicate where the MDD group showed a more negative difference waveform relative to controls; that is, where the negative minus positive difference waveform in the MDD group was more negative compared to controls. Permutation tests indicated significant differences from 380 to 866ms. Lighter colors indicate the converse: that the MDD group showed a more positive difference waveform, primarily in the posterior sites; these differences were most evident in a later time-period, from 380ms through the

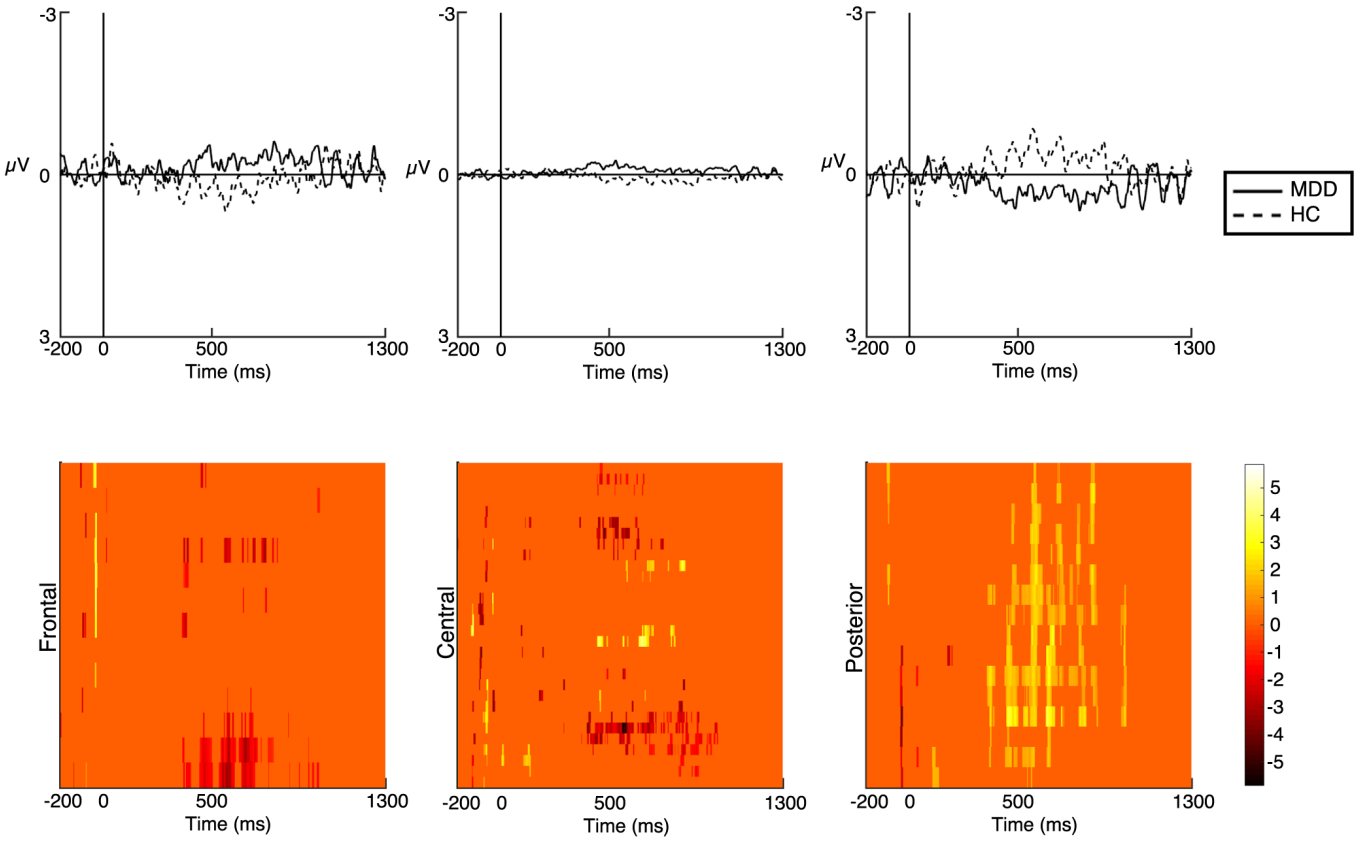

Figure 2.3: Event-related potential (ERP) differences by group. Top: ERP difference waves elicited by negative minus positive words, for MDD group (solid line) and HC group (dashed line), recorded at frontal scalp sites (left), central (middle), and posterior (right). Stimulus presentation is indicated by the solid black line at time 0; negative voltage is plotted up. ERPs are averaged across the electrodes that showed statistically significant differences between groups. Bottom: Color values indicate significant ($p < .05$) t-values for clusters comparing MDD to HC groups across positive and negative images. Clusters are arrayed by time (x-axis) and by laterality (y-axis, with left at the bottom and right on the top). Orange (primary shading) indicates no significant spatiotemporal difference; darker colors indicate that the MDD group showed a more negative waveform, and lighter that the MDD group showed a more positive waveform.

end of the analysis window at 1,000ms.

To facilitate interpretation of the difference waveforms observed in Figure 2.3, we also plotted the average waveforms across frontal, central and posterior scalp locations collapsed across all electrodes that showed significant differences in the statistical maps (i.e., those indicated in the lower section of Figure 2.3), in response to negative and positive stimuli separately for the MDD and HC groups. These plots are shown in Figure 2.4. Beginning around 450ms in frontal and central sites, more positive ERPs to positive words than to negative are evident in data from the MDD participants, whereas HC participants show little difference between stimuli. The posterior differences were inverted relative to the frontal effects. Beginning around 500ms in posterior sites, the MDD participants showed more positive ERPs to negative words than to positive; HC participants showed the opposite effect. This pattern over posterior sites begins late and grows more positive over time. The waveform is less positive in response to positive words for the MDD group, whereas it is less positive in response to negative words for the HC group. There were not apparent differences for either group in early responses to positive and negative words, either between or within groups.

Relationship between behavioral and neural outcomes

Based on the results described above, we exported timepoint-by-timepoint voltage per participant, averaged across all spatiotemporal clusters marked as significantly different between MDD and HC groups in the difference waves (i.e., the mean per-group of the ERPs depicted in Figure 2.3). We then examined whether the ERP response to negative words minus positive words predicted behavioral outcomes (i.e., processing bias and endorsements).

For ERP response predicting endorsements of word stimuli on the SRET, a lin-

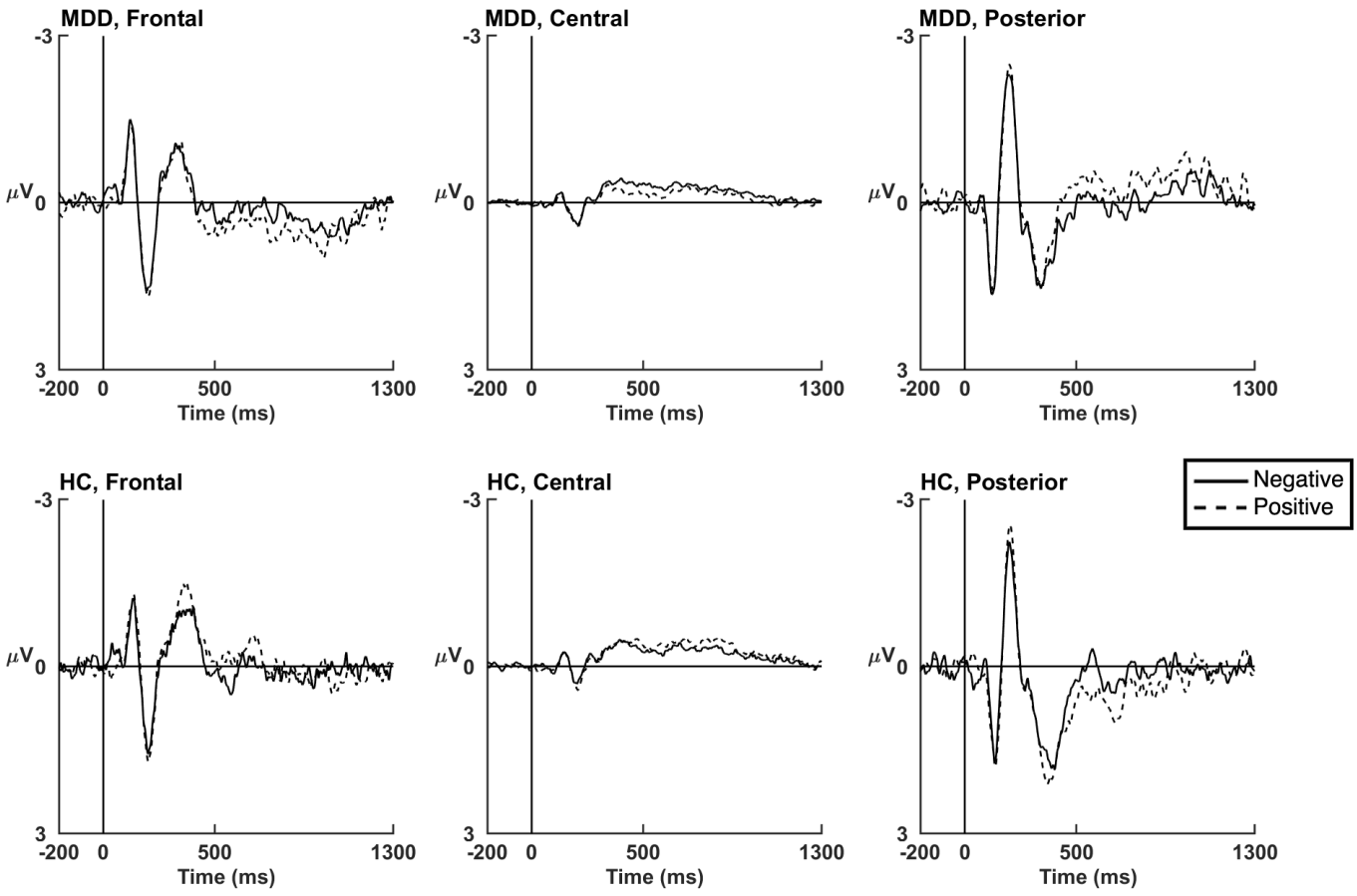

Figure 2.4: Average Event-Related Potentials for each group, collapsed across regional electrode sites that were significantly different in the permutations tests; negative is plotted upwards. Any electrode that showed significant differences in the permutations tests is included in its regional average. The solid line shows negative words and the dashed line positive words; difference waves for permutations tests were calculated as negative minus positive. The MDD group is in the top row, and HC in the bottom row; plots show grand averages across subjects in that group, for significant electrode sites in that region. In frontal and central sites, the MDD group on average shows a more negative amplitude towards negative words, whereas the HC group shows a more negative amplitude towards positive words. In posterior sites, the trend is reversed; MDD shows a more positive amplitude trend towards negative words, whereas the HC group shows a more positive amplitude trend towards positive words.

ear mixed model regression with a factor of valence (positive or negative words) and a continuous predictor of the non-parametric test voltage values found an interaction between valence and the voltage values, $t(78) = -2.80, p = .006$, Cohen's $d = -0.86$, 95% CI [-1.51, -0.21]. A more positive ERP difference was related to an increased number of negative words endorsed, $t(39) = 2.28, p = .029$, Cohen's $d = 0.50$, 95% CI [0.06, 0.95]. Conversely, a less positive ERP difference was non-significant in predicting the number of positive words endorsed, $t(39) = -1.64, p = .11$, Cohen's $d = 0.41$, 95% CI [-0.83, 0.08]. These results are consistent with the above-discussed group effects, as MDD participants are more likely to show more positive ERP difference waves (to negative minus positive words), especially in posterior sites.

A final model tested whether the behavioral and ERP results were independent predictors of depression group status, using logistic generalized linear models with group status modeled as 1 for MDD and 0 for HC. An additive model with three predictors: positive words endorsed, negative words endorsed, and permutations test voltage values (rescaled with a standard deviation of 1, from 0 – 4, to result in odds ratios [OR] that were interpretable), revealed a significant effect for number of negative words endorsed, $OR = 1.66$, 95% CI [1.21, 3.49], $z = 2.04, p = .04$; MDD diagnosis was 1.66 times more likely for every additional negative word endorsed. This model had a non-significant effect of the number of positive words endorsed, $OR = 0.81$, 95% CI [0.51, 1.05], $z = -1.31, p = .19$ and the permutations test voltage values, $OR = 0.82$, 95% CI [0.10, 5.00], $z = -0.23, p = .82$.

These predictors were correlated with one another but measured separate constructs. The number of negative words endorsed was strongly negatively correlated with the number of positive words endorsed, $r = -.63$, 95% CI [-.42, -.79], $p < .001$ and was positively correlated with the permutations test voltage values, $r = .34$, 95% CI [.04, .59], $p = .03$; the number of positive words endorsed was not

significantly correlated with the permutations test voltage values, $r = -.25$, 95% CI [-.52, .06], $p = .11$. There was a marginally-significant positive correlation between permutations test voltage values and score on the CES-D, $r = .26$, 95% CI [-.05, .52], $p = .09$.

ANOVA: Early attentional components

We performed parametric analyses of variance (ANOVA) to compare ERP responses between groups at specific time-windows; these waveforms are shown in Figure 2.5. Area under the curve was calculated for each component and entered into a mixed-level ANOVA, with fixed factors of group and valence, and a random factor of participant. There was no significant interaction of group $\times$ valence for the area under the P1 waveform, 100 to 200 ms following stimulus presentation, $F(1, 40) = 0.03, p = .87$, generalized $\eta^2 < .001$. There were no significant main effects of group or valence: for group, $F(1, 40) = 0.02, p = .90$, generalized $\eta^2 < .001$, or for valence, $F(1, 41) = 0.21, p = .65$, generalized $\eta^2 < .001$.

Likewise, there was no interaction of group $\times$ valence for the area under the P2 waveform, 200 to 300 ms following stimulus presentation, $F(1, 40) = 0.06, p = .80$, generalized $\eta^2 < .001$. There were also no significant main effects of group or valence: for group, $F(1, 40) = 0.12, p = .73$, generalized $\eta^2 = .003$, for valence, $F(1, 41) = 0.11, p = .74$, generalized $\eta^2 < .001$. These results demonstrate that for these early components, neither group showed significant effects of valence nor significantly differed from the other.

ANOVA: Late positive potential (LPP)

For the early LPP, calculated at the same sites as the P1 and P2 components, from 400 to 600 ms, there was a significant group $\times$ valence interaction,

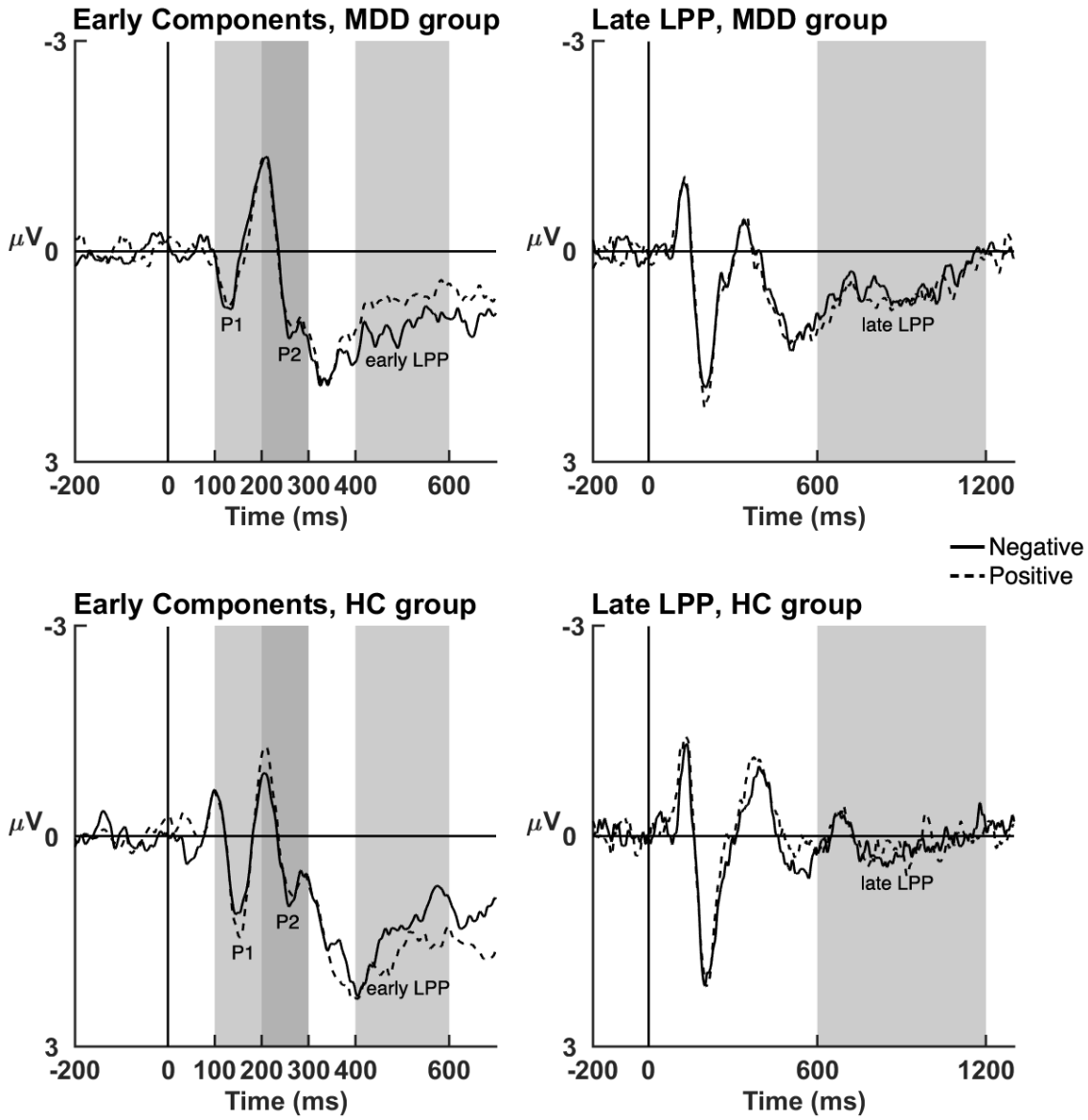

Figure 2.5: Parametric event-related potential components compared between groups. Negative is plotted upwards. The solid line shows negative words and the dashed line positive words. Shaded regions indicate the time-frame for the labeled waveforms. Early attentional components (at left; as indicated, P1, P2, and early LPP) are shown from 200 ms before stimulus presentation to 700 ms following, and are averages across Pz, POz, P1, P2, PO3, and PO4. The late LPP can be seen at right; the waveform is plotted from 200 ms before stimulus presentation to 1,300 ms following, and is an average across Fz, FCz, and Cz. The MDD group is shown at the top, and the HC group at the bottom.

$F(1, 40) = 7.51, p = .009$, generalized $\eta^2 = .01$.[13] This interaction, visualized in Figure 2.5, indicates that the depressed group exhibited greater activity following negative versus positive words and healthy controls demonstrating the opposite pattern. This difference (i.e., amplitudes to positive – negative stimuli between groups) was a medium to large effect (MDD: $M = -0.16\ \mu V, SD = 0.34$; HC: $M = 0.15\ \mu V, SD = 0.38$. For the comparison of difference waves, Cohen's $d = 0.85$, 95% CI [0.20, 1.50]).

The late LPP was calculated across Fz, FCz, and Cz from 600 to 1200 ms. A mixed-effects ANOVA with factors of group and valence found no group × valence interaction, $F(1, 40) = 0.01, p = .92$, generalized $\eta^2 < .001$. After dropping the interaction, there was a significant effect of group, $F(1, 40) = 6.09, p = .018$, generalized $\eta^2 = .13$, with the depressed participants showing greater amplitudes across both valences compared to healthy controls. These group differences are apparent in the right portion of Figure 2.5. There was also a main effect of valence, with a greater amplitude to negative words compared to positive words, $F(1, 41) = 4.89, p = .03$, generalized $\eta^2 = .04$.

Anxiety as a covariate

For three of the primary analyses reported above, we repeated the ANOVA analyses with the Anxious Arousal subscale of the MASQ as a covariate. For the prediction of negative processing bias, the group × valence interaction remained significant after including anxious arousal as a covariate, $F(1, 79) = 274.2, p < .001$, generalized $\eta^2 = .78$. For ERP response predicting endorsements of word stimuli,

[13] For this test only, analyses with data that had not had a linear de-trend applied during the data processing pipeline found a different result—the interaction was no longer significant, $F(1, 39) = 2.65, p = .11$, generalized $\eta^2 = .003$. A main effect of valence was also not significant, $F(1, 40) = 0.872, p = .36$, generalized $\eta^2 = .021$.

with anxious arousal as a covariate, the interaction between word valence and the voltage values remained significant, $t(77) = -2.85, p = .006$, Cohen's $d = -0.86$, 95% CI [-1.50, -0.22]. Finally, for the early LPP (from 400 to 600 ms) as the outcome, a significant group valence interaction remained when including anxious arousal as a covariate, $F(1, 39) = 7.51, p = .009$, generalized $\eta^2 = .01$.

Discussion

This study examined the electrocortical corollaries of positive and negative self-referent processing in depression using the SRET. Using a novel analytic technique, pointwise non-parametric randomized permutations, we compared significant spatiotemporal differences between MDD and healthy control groups. Our primary findings indicated that across posterior sites, beginning immediately but becoming strongest by 380 ms and increasing until the end of the analysis window (1,000 ms), MDD participants demonstrated more positive ERP amplitudes in response to negative words than positive words, whereas HC participants showed an inverse configuration. This pattern of findings may be driven by the late positive potential (LPP), which is associated with sustained attentional engagement and increased cognitive evaluation of negative material in MDD (Shestyuk & Deldin, 2010; Auerbach et al., 2015). Further, these results could not be accounted for by the presence of anxious arousal symptoms.

Findings in the current study indicate differential cortical responses between MDD and healthy controls over centroparietal sites that are generally consistent with previously observed LPP responses in MDD (Shestyuk & Deldin, 2010; Auerbach et al., 2015). As indicated in Figure 2.3, these effects are most evident across central and posterior sites in a time period suggestive of cognitive evaluation (i.e., from approximately 350 to 900 ms). Further, using parametric analyses, we observed a

group × valence interaction for the early portion of the LPP (400 to 600ms). The depressed group exhibited greater activity following negative versus positive words and healthy controls demonstrated the opposite pattern. This effect replicates past work (Auerbach et al., 2015), which found depression group differences for amplitudes to positive versus negative words during this same time window at the same electrode locations.

Recent research in participants with MDD has shown less positive LPP amplitudes in response to rewarding images compared to HC participants (Weinberg et al., 2016; MacNamara, Kotov, & Hajcak, 2016). Our findings of reduced LPPs in response to positive stimuli in the MDD group are consistent with these findings. Indeed, recent work indicates that MDD affects reward-processing as well as negative information-processing (Proudfit, 2015). This reduction in evaluation of reward thus fits with the RDoC concept of depression as a failure of a positive valence system (Proudfit, 2015; Cuthbert & Insel, 2013).

For the later portion of the LPP, the parametric analysis of variance revealed a main effect for group (not an interaction between stimuli condition and group), where MDD participants had more positive amplitudes to both word valences compared to healthy controls. However, valence differences during this time period between MDD and healthy controls were captured in the pointwise non-parametric randomized permutation analyses, suggesting that approach may have been more sensitive than parametric analyses. Thus, results of the non-parametric analyses suggest that depressed individuals show greater late-stage cognitive evaluation of negative stimuli, which is consistent with prior work examining the LPP in depression.

In contrast to the findings at posterior locations, non-parametric analyses also revealed negative amplitudes in frontal and some central regions for negative versus positive words in MDD but not HC participants, with consistent differences

appearing from approximately 380 ms to 866 ms. However, the functional significance of these left frontal differences is unclear. Additionally, interpretation of these frontal effects must be tempered by the fact that we did not employ recording sites over frontopolar regions.

It is possible that these negative amplitudes reflect an N2 wave (although the N2 often appears earlier in the time course). The N2 or N200 waveform is a scalp potential with negative polarity that is located in frontal and central regions and appears to be linked to error monitoring, cognitive monitoring, and response inhibition (Schmajuk, Liotti, Busse, & Woldorff, 2006; Ramautar, Kok, & Ridderinkhof, 2004). The N2 is more negative when inhibition of response is successful in a Stop-Signal task (Schmajuk et al., 2006). Given that in our data, MDD participants showed a more negative amplitude during this time-frame in response to negative words, it is possible that this reflects increased cognitive evaluation and monitoring of the correct response to these negative stimuli. It is also possible, however, that this waveform results from participants inhibiting their behavioral response on the task until they were permitted to respond. Future research will need to tease apart these competing explanations should this pattern be replicated.

Past ERP studies have also found that early attentional responses, including the P1 and P2, differentiated MDD from HC participants (Shestyuk & Deldin, 2010; Auerbach et al., 2015; Waters & Tucker, 2016). However, the current study did not find these distinctions in the nonparametric analysis, nor in parametric analysis of variance performed on time-locked waveforms, although such differences were hinted at in the waveforms seen in Figure 5. These findings indicate that there were not large differences in these stages of processing of emotional stimuli. Instead, the results of the current study seem to indicate that the primary differences between groups were evident in later, more elaborative stages of processing and cognitive evaluation.

Given that we based the design of our task on past work (Auerbach et al., 2015), there were no major methodological differences that should have resulted in this lack of early attentional difference (in the P1 or P2) between groups. Thus, differences across studies could be due in part to developmental stage (adolescents in prior work versus adults in current work), symptom severity, or other factors. Future work that examines the ERP responses to the SRET in depressed samples across the lifespan could address this question directly.

In this study, behavioral data from the SRET consistently showed that participants in the MDD group but not the HC group endorsed more negative words and fewer positive words as self-descriptive compared to healthy controls. We also found a significant relationship between the ERP responses and the behavioral results—most prior studies have not linked these levels of analysis. This relationship indicated that electrical activity within frontal, central, and posterior scalp sites across all electrodes that showed significant differences between the MDD and healthy control groups were predictive of the number of negative words endorsed—i.e., that a more positive amplitude to negative words (relative to positive words) was related to an increased endorsement of negative words. This supports the idea that the brain responses to negative stimuli found herein are associated with self-referential processing, rather than simply representing an ERP response to word valence. Moreover, this relationship falls within the negative valence system (Woody & Gibb, 2015; Cuthbert & Insel, 2013), demonstrating how electrocortical activity in MDD may be strongly connected to negative self-reference. Indeed, the increased cognitive evaluation of negative stimuli may be linked to rumination, which is common in MDD (Nolen-Hoeksema, 2000). How rumination may differ from increased cognitive evaluation bears further scrutiny in psychophysiological work with the SRET.

Although the non-parametric analytic techniques used in this study were con-

servative, and there are several strengths to our methods, it is important to acknowledge that our sample was relatively small, which limits our ability to detect large effects. Larger samples could also further examine the role of anxiety symptoms or other mental illness in self-referent processing, or allow us to fully explore potential gender differences in the results (although results were consistent when a covariate of gender was added). Consistent with past work, the SRET did not include an other-reference condition, focusing solely on self-reference. Given the nature of the stimuli, many of the healthy participants did not endorse many negative words and many of the participants in the MDD group endorsed few positive words. This makes subdividing ERPs into cells by valence and self-reference difficult, due to empty cells for many participants. Including neutral words in future studies using the SRET could provide an alternative ERP difference wave model that would provide further evidence of the presence or absence of LPP differences between groups. Additionally, although this study did perform a structured diagnostic interview with participants, diagnoses were performed over the phone and not by clinicians. It is possible that inclusion in the study would have been modified slightly had participants instead been diagnosed by clinicians.

The current study is also consistent with recent efforts in other areas of research to develop literatures that are robust and replicable, an effort that proves increasingly important in the current neuropsychological landscape (Munafò et al., 2017). One important aspect of this effort is for independent laboratories to conduct replications of prior work to determine whether previously observed results are consistent across settings and are robust to changes in methods or samples. The current study attempted to replicate prior findings and indeed found support for later stage cognitive evaluation of negative information in MDD. We believe that conducting additional replication studies for important clinical phenomenon is a critical direc-

tion for psychopathology research in general. Engaging in large scale, multi-site, and pre-registered collaborative studies should be central to this endeavor (Tackett et al., 2016).

In summary, the current study provides evidence of increased cognitive evaluation of negative compared to positive self-referent stimuli in major depression, without evidence of differential early attentional engagement. These results are evident both behaviorally and in later posterior ERP components thought to reflect cognitive evaluation. Negative stimuli appear to capture and sustain attention among participants with MDD to a greater degree than positive stimuli during the later stages of information processing. Importantly, brain responses during the cognitive evaluation stage were also predictive of the number of negative words endorsed, even after statistically controlling for anxiety symptoms, indicating that this component of the ERP is related to self-referential processing. Given these results, late-stage event-related potentials that support biased processing of self-referential stimuli appear to be a stable feature of depression. As such, future work should investigate whether this processing can be ameliorated through treatment. Results from the present study indicate that interventions should target later, more elaborative stages of information processing and provide important direction for identifying the brain responses that should be targeted by such treatments in major depressive disorder.[14]

[14]This work was supported by awards from the National Institute on Drug Abuse (4R01-DA03245705) and the National Institute of Mental Health (1R56-MH10865001A1) to CGB.

Study 3

Positive Imagery Training Increases Positive Self-Referent Cognition in Depression[15]

Introduction

Over the past half-century of depression-related research, the field has accumulated a great deal of evidence pointing to the role of cognitive biases in maintaining depressive symptoms (Beck, 1967; Gotlib & Joormann, 2010; Gotlib & Krasnoperova, 1998). The relationship of such biases and depressive disorders is well established; individuals diagnosed with depressive disorders are more likely to have negative self-schemas and to view themselves negatively (Beck, 1967). Beck theorized that negative views about the self, the world, and the future (the cognitive triad) contribute to the maintenance of depression.

Such negativity is mirrored by a difficulty relating to positive information; Beck referred to this as a "positive blockade" (1967). Whereas negative self-referent information is preferentially processed with relative ease, positive information tends to be ignored, discounted, or processed with difficulty. Together, these biases help to define the existing memory representations of the world, including self-representations. Individuals with negative self-views (or schemas) are more likely to experience increased depressive symptomatology (Connolly et al., 2015; LeMoult, Kircanski, et al., 2017).

[15]Dainer-Best, Shumake, & Beevers, 2018

The relationship between negative self-schema and the maintenance of depression is sometimes viewed in causal terms, with stronger negative self-schemas thought to produce a more protracted episode of depression (Beck, 1967). It is clear that both positive and negative schema are strongly correlated with depressive symptoms (Phillips et al., 2010; Dainer-Best et al., in press), and can predict worsening of symptoms (Disner et al., 2016); however, such studies are correlational and thus vulnerable to third variable explanations that account for the association between self-referent processing and depression.

As such, attempts to modify self-schema can be viewed as a direct test of whether maladaptive cognitions serve to maintain depression, as reductions in self-schema should precede reductions in depression symptoms, if there is a causal link. Although cognitive-behavioral therapy (CBT; Beck, 1979) targets maladaptive cognitions, few studies to date have directly measured self-referent information processing, with assessments such as the Self-Referent Encoding Task (SRET; Derry & Kuiper, 1981), before and after CBT. While schema change is a hypothesized mechanism for depression improvement, randomized controlled trials rarely track biased self-referent processing over the course of treatment.

Two such studies of adults with social anxiety disorder (Goldin et al., 2013; Thurston, Goldin, Heimberg, & Gross, 2017) measured self-referent processing with the SRET before and after CBT. Both studies found reduced negative self-reference and increased positive self-reference following treatment. With a depressed sample, Quilty, Dozois, Lobo, Ravindran, and Bagby (2014) measured change in self-reference in pharmacotherapy and CBT with the SRET. They reported that self-referent processing changed similarly across both treatment conditions. Further, Dozois et al. (2009) reported that depressed adults who received cognitive therapy in conjunction with pharmacotherapy saw improvement in positive and negative self-schemas that

was not observed in people who received pharmacotherapy alone. Both treatment groups reported similar reductions in depression during treatment.

CBT interventions are multifaceted, and so even if change in self-referent processing is observed, it is unclear what aspect of treatment may be producing those changes. In contrast, there is a nascent literature using cognitive bias modification (CBM) techniques to target and change specific cognitive biases associated with depressive disorders (Hallion & Ruscio, 2011; MacLeod, Koster, & Fox, 2009). Studies that target specific cognitive mechanisms with CBM have generally been focused on attention (Wells & Beevers, 2010; Ferrari, Möbius, van Opdorp, Becker, & Rinck, 2016), overgeneral memory (E. R. Watkins, Baeyens, & Read, 2009; Raes, Williams, & Hermans, 2009; Neshat-Doost et al., 2013), and interpretation (E. A. Holmes, Lang, & Shah, 2009; Joormann, Waugh, & Gotlib, 2015; LeMoult, Colich, et al., 2017). Assessing whether such focused interventions leads to symptom reduction is ideal for theory testing, and continues to be an important focus of clinical research (MacLeod et al., 2009).

Interpretation-focused CBM studies in particular have used imagery-based techniques to remediate maladaptive cognitive processes (E. A. Holmes et al., 2009; Hitchcock et al., 2016). Such techniques rely on participants' mental imagery rather than providing cues to specific stimuli. They may also have a more idiographic target, focusing on individualized information processing rather than directing all participants to specifically attend to the same stimuli. Such cue-based mental imagery interventions may directly target maladaptive schema in depression (E. A. Holmes, Blackwell, Burnett Heyes, Renner, & Raes, 2016) and elicit emotional reactions as though imagined events were happening (E. A. Holmes & Mathews, 2010).

Mental imagery interventions that focus on depression-specific maladaptive cognitions have the potential to change self-referent cognitions. Given a biased self-

schema, any such interventions must change the way that judgments of information about the self are made. Previous research has shown that making self-referential judgments about emotional material enhances specific verbal and visual information and can boost memory for that information (Hamami, Serbun, & Gutchess, 2011). Thus, by repeatedly encoding positive self-referent information through mental imagery, it becomes more salient when self-schema is accessed. This prioritization makes a positive self-schema more accessible, thus reducing negative self-referent processing biases. Further, S. B. Klein (2013) has additionally theorized that such future-oriented imagery is straightforward even when individuals have no episodic memory of a situation. Other work has also shown that simulating positive future events in a single session can increase depressed participants' ratings of the likelihood and importance of other future events (Boland, Riggs, & Anderson, 2017).

Given the robust association between the self-referential bias and dysphoria, we developed a CBM intervention focused on enhancing positive self-referent processing by encouraging participants to repeatedly practice viewing themselves positively and imagining a pleasant and positive future. To provide an adequate control, we developed a neutral intervention that focused instead on detailing objects in a neutral, present situation. This training directly mirrored the positive intervention, but without what we believed to be the active components of that intervention. That is, participants did not focus on positive, self-referent, or future-focused imagery.

We thus compared a positive self-reference training (PSRT), focused on using positive cues to encourage positive, future, self-referent processing, to a neutral training condition (NTC) that used neutral cues to encourage neutral, current non-self-referent processing. We hypothesized that participants in the PSRT—but not the NTC—group would show improvement on measures of self-referential bias (i.e., the SRET), especially in terms of their response to positive items. We also hypothe-

sized that the reduction in depressive symptoms in the PSRT group would be linked to their improvement in self-reference.

Methods

Participants

Participants in the study were eligible if they were adults between the ages of 18-45, had elevated total score > 13 on the Center for Epidemiologic Studies–Depression (CES-D) scale, were fluent in English, and provided informed consent. Participants who completed the study (completers; $N = 87$) were on average 26.4 years old ($SD = 7.0$), female (84%), and white (66%). They were paid between \$25 and \$35 for completing the study, with higher payments being received for completing additional training sessions as discussed below. Participants in the intention-to-treat (ITT) sample, who completed the full baseline assessment and thus were randomized into one of the two training conditions ($N = 264$), were on average 26.8 years old ($SD = 7.0$), female (75%), and white (67%).

Participants were recruited through online postings advertising an online mood study using three online forums: (1) Craigslist, an online bulletin board, in several major cities; (2) ResearchMatch, a national health volunteer registry supported by the U.S. National Institutes of Health as part of the Clinical Translational Science Award (CTSA) program; and (3) online postings through a community events board at the University of Texas at Austin. The advertisements directed participants to a website to determine study eligibility. Here, participants provided informed consent, provided their age, and filled out the CES-D. If they were eligible for the study, further measures (discussed below) were collected, and participants were randomized into one of two training conditions.

Measures

Center for Epidemiologic Studies–Depression Scale (CESD)

The CESD (Radloff, 1977) is used to assess depressive symptoms over the past week using a 20-item self-report questionnaire. Scores may range from 0 to 60, with higher scores indicating elevated depressive symptoms. Scores greater than 16 have been regularly used as an indication of possible diagnosis of major depressive disorder (Radloff, 1977; Santor et al., 1995); scores above 13 may thus indicate mild depressive symptoms. The CES-D was used to determine eligibility at baseline, and assessed again following one week (T1) and two weeks (T2).

Self-Referent Encoding Task (SRET)

The SRET (Derry & Kuiper, 1981) is an affective decision-making task designed to assess schema-related processing. Participants make decisions about whether positive and negative adjectives are self-descriptive. Words appear on their computer screens, and participants make rapid judgments following each word's display. The SRET was presented using Inquisit software (Millisecond Software LLC, Seattle, WA). Inquisit is a commercial platform that runs "full screen" on participant's computers as a Java applet. The SRET was administered at baseline (T0), one week (T1), and two weeks (T2) after baseline.

The SRET consisted of a short practice block followed by the actual task. The practice block allowed participants to view and respond to 5 practice words. After completing the practice block, 90 words were displayed once each in random order. Words were displayed in white text on a black screen and remained on-screen until participants responded. Participants were told to use the Q or P keys on their personal keyboard to answer whether the word described them or not. Each trial was followed by a 1,500 ms intertrial interval. Words were chosen from among the

words used in previous SRET studies (Disner et al., 2016; Auerbach et al., 2015; Dainer-Best et al., 2017) and had been selected from the Affective Norms for English Words (M. M. Bradley & Lang, 2010) and from an additional validated word-list (Doost et al., 1999).

As the SRET was presented three times (at T0, T1, and T2), a subset of words were presented only once. Thus, 80 words were presented every time,[16] and an additional 30 words, chosen randomly from all possible words, were presented once each: 10 at baseline, 10 at T1, and 10 at T2.[17] Although free recall data has often been collected, we did not collect such data in this study due to concerns about learning effects due to repeated administration.

The primary behavioral outcome from the SRET is the probability of endorsing positive vs. negative words, which we have demonstrated to be strongly linked to depression (Dainer-Best et al., in press). Additionally, including an extra subset of words at each time-point allowed us to distinguish between responses that varied due to repetition and those that varied due to changes in self-referential processing.

[16]The following positive words were shown each time: joyful, gentle, cool, funny, surprised, glad, loyal, bold, proud, good, outstanding, awesome, elated, adorable, admired, devoted, untroubled, festive, content, beautiful, satisfied, thoughtful, free, excellent, dignified, confident, excited, kind, capable, brilliant, best, helpful, nice, alive, pleased, terrific, bright, vigorous, carefree, playful. The following negative words were shown each time: disloyal, alone, dumb, hostile, helpless, worried, defeated, annoyed, obnoxious, mad, shamed, morbid, anguished, unhappy, guilty, distressed, frustrated, ashamed, weak, disgusted, wicked, brutal, unloved, displeased, depressed, sad, terrified, fearful, violent, burdened, lonely, crushed, angry, scared, bad, hateful, upset, troubled, afraid, sorry.

[17]These words included, at baseline: wonderful, lucky, useful, fantastic, grateful (positive) and useless, cruel, nasty, terrible, dreadful (negative); at T1: great, fun, jolly, loved, engaged (positive) and rejected, lost, stupid, insane, unwanted (negative); and at T2: cute, hopeful, friendly, famous, masterful (positive) and insecure, sinful, horrible, rude, bored (negative).

Training

There were two training conditions, positive self-reference training (PSRT) and a neutral training condition (NTC). The PSRT prompted participants to focus on (1) future events that were (2) specific, (3) positive or fun, and (4) involved the participants' involvement. Positive cues focused on places or concepts that could engage self-referent positive imagery, like going to a caf or receiving a gift; instructions explicitly suggested scenarios that would be "positive, pleasant, or fun".[18] The full script that participants viewed is included in the supplementary materials. The NTC, conversely, was designed to lack the primary components that were active in the PSRT condition. As such, the NTC training instructed participants to focus on (1) present locations that were (2) generic, (3) neutral, and (4) involved primarily objects and spaces. Neutral cues were all focused on places. The practice cues included describing a caf or describing a gift shop.[19]

Both training conditions were couched in the idea of "target[ting] unhelpful ways of thinking". Participants viewed videos elaborating the details of their training; the full text of each training video is included in the supplemental materials. They were reminded of the important details of their training condition before completing the training each day, and were also emailed a link to view the video again. Both trainings provided participants with cues and asked them to create audio recordings of themselves responding to a prompt. The prompt and cues both differed by condition, although there was some overlap in cues. Two cues were given

[18]The remaining PSRT cues were: going shopping, an achievement, feeling relaxed and comfortable, a celebration, being in a peaceful place, playing a game, feeling capable, listening to music, being somewhere beautiful, being extremely interested in something, being supportive, a garden or museum, a great restaurant, your favorite movie.

[19]The remaining NTC cues were: a clothing store; a theatre; a bedroom; a banquet hall; an office; a game room; a classroom; a music hall; a palace; a science museum; a restaurant; an art museum; a garden; a meeting room.

each day of the training; the first day, these were described as "practice cues".

Following the video training, participants were instructed to complete the practice cue recordings for their condition, talking into a microphone about the cue for 3-5 minutes. They were instructed that they should complete the two practice recordings, and then would receive emailed reminders to complete a minimum of four and a maximum of seven additional pairs of recordings, every other day. Thus, participants who completed the study submitted between 10 and 16 cued recordings. (Participants who submitted fewer than ten recordings were not considered adherent.) Completing more recordings was encouraged with a hierarchical payment structure. Participants were paid $25 if they completed all assessments and a minimum of eight recordings. For each additional set of recordings that participants completed (i.e., 10, 12, or 14), participants were paid an additional amount ($2, $3, and $5), enabling participants to earn a maximum of $35. This information was included in the consent form and reiterated after randomization.

All recordings were voice-only recordings spoken directly into a microphone, on participants' personal phones or computers. Participants were able to complete the recordings either through a "chat bot" in Telegram, a mobile application available for major smartphone platforms (Telegram Messenger LLP, Berlin, Germany), or through a website using a jQuery (jQuery Team) plugin to record from a computer microphone. Both methods uploaded participants' voice recordings to a secure website to ensure training adherence. Of those who completed the study, in the NTC group, 31 participants used Telegram and 12 participants used the web interface; in the PSRT condition, 25 participants used Telegram and 19 participants used the web interface. Following every day's cue recordings, participants were asked to rate how well they believed they had been able to imagine the scenario they were recounting.

Assessment of subjective benefit and acceptance

At the end of the study, before completing other measures, participants completed a ten-item questionnaire assessing their subjective satisfaction with the training program. These questions were drawn from those reported in studies of online interventions (Meyer et al., 2009), modified to best fit the current work.

Procedure

The Institutional Review Board at the University of Texas at Austin approved all procedures. The trial was registered at ClinicalTrials.gov (Identifier NCT03056963; Dainer-Best, 2017).

Participants provided signed informed consent online using REDCap electronic data capture tools hosted at UT-Austin (Harris et al., 2009). REDCap (Research Electronic Data Capture) also presented all questionnaires. Following the completion of CES-D screening, demographic information were collected. Participants were asked to self-define a personal identification phrase that was specific but did not contain protected health information (PHI). This phrase was used to link data between training recordings, self-report questionnaires, and the behavioral task. They were then automatically directed to a separate website to complete the SRET. The SRET took between 8 and 10 minutes.

After completing the SRET, participants were randomized into either a positive (PSRT) or neutral (NTC) condition. Randomization was not stratified in any way and was entirely random for the majority of the data collection period (i.e., three months). Near the end of data collection (i.e., for one week), in order to conclude the study with an approximately equal number of completers between intervention groups, participants were randomized at a rate of 4:1 to the neutral condition. Seven

NTC participants and one PSRT participant completed the study during this 4:1 allocation period.

Once randomized, participants were again provided with information about payment for the study as described above, and asked to confirm that they understood this information. They were then provided with a short video explaining the training for their condition, as described above. They completed a practice set of cues. Every other day henceforth and for the following two weeks, at the same time as they finished watching the training video, each participant received an email reminding them to log in to REDCap to complete the day's training. Thus, participants received a reminder email for seven trainings. On the fourth day of training (i.e., eight days following baseline) participants completed the CES-D and SRET, before being invited to complete that day's training. This schedule can be seen in Figure 3.1.

After all trainings had been completed, on the 15th day following baseline, participants received an email asking them to once more complete the CES-D and SRET; they also completed the subjective benefit questionnaire at this time. Participants were paid $25 to $35, as described above. All study procedures were automated, although research personnel were available to troubleshoot any technical issues that occurred through data collection.

Participant Attrition and Data Filtering

The screening was begun by 1,243 participants, of which 678 (54%) were eligible and completed the remaining questionnaires. (See Figure]3.2 for a CONSORT diagram (Moher, Schulz, & Altman, n.d.) describing participant flow.) Only 264 participants were randomized—the remaining participants were not eligible, did not download the software to complete the SRET, or did not provide their identification phrase. Of participants who were randomized to the PSRT condition ($N = 124$),

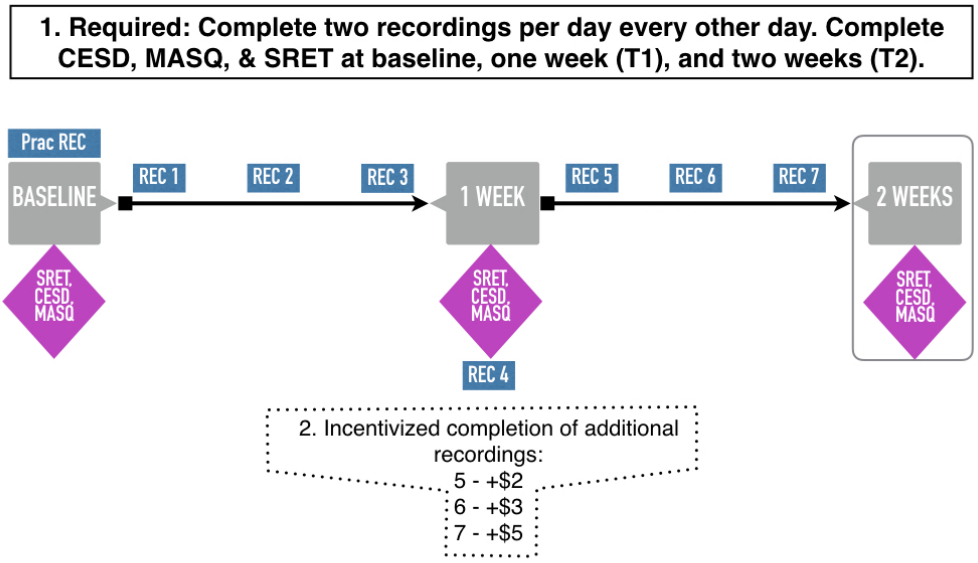

Figure 3.1: Assessment schedule showing how SRET, CES-D, and recordings were administered.

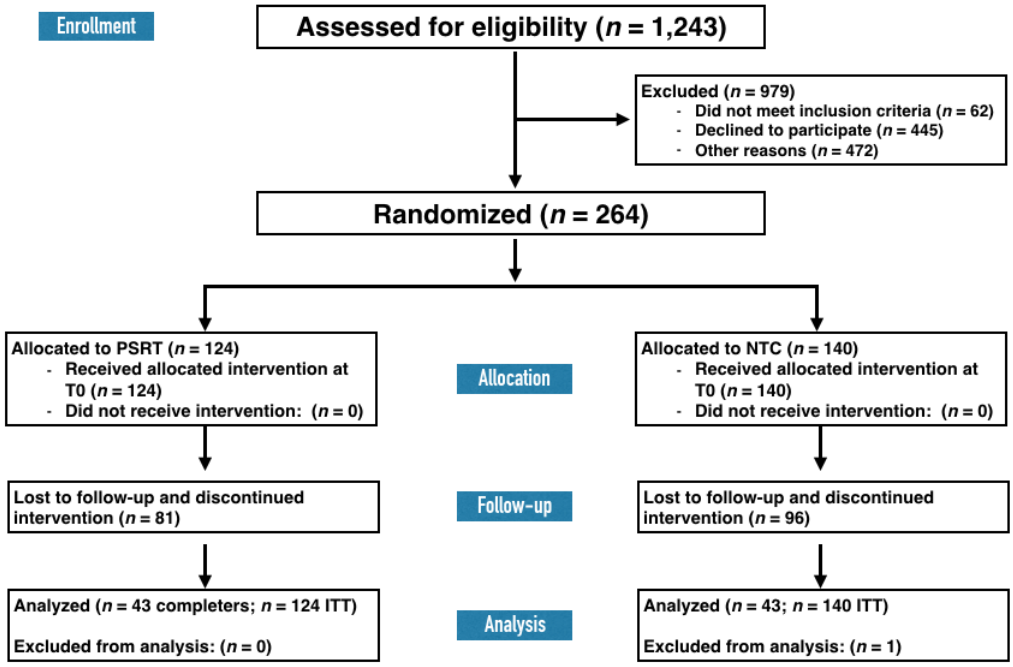

Figure 3.2: CONSORT flow chart. There were 1,243 visitors who began the eligibility screen; their uniqueness cannot be assessed. Following randomization, the interventions were accessed 264 unique times. All such participants were included in either ITT (intention-to-treat) or completer samples.

31% completed the study after randomization (i.e., four or more training sessions as well as baseline, one-week [T1], and two-week [T2] assessments). Of participants who were randomized to the NTC condition ($N = 140$), 35% completed the study after randomization. We refer to these participants as completers. Participants' questionnaire data was dropped for a specific time-point if all questions on the questionnaire were answered the same despite some items being reverse scored or if "catch" items (e.g., directing participants to select a certain answer to a question) were not answered correctly ($N = 10$ [3.8%] at baseline, 1 [1.0%] at T1, and 2 [2.3%] at T2). These metrics were taken as an indication that participants were likely not attending to the self-report questionnaires, and thus individual responses were dropped.

On the SRET, trials were marked as outliers if reaction times (RTs) were less than 200 ms (0.9% of all trials). These trials were dropped from analyses. After dropping trials based on RT, those participants with more than 25% of trials deemed outliers for any reason were excluded for that time-point ($N = 3$ [3.4% of completers]).

Analysis of Missing Data

As described above, a large percentage of data were missing owing mainly to participant attrition and, to a lesser extent, data cleaning. To help diagnose the mechanism of missing response data, a logistic regression predicting dropout as a function of treatment condition and baseline CES-D found no interaction between the two, $\beta = 0.01$, 95% CI [-0.04, 0.06], $p = 0.59$. There was a near-significant effect of CES-D that was independent of treatment assignment, $OR = 1.02$, 95% CI [1.00, 1.05], $p = .07$, with greater baseline depression increasing the likelihood of dropping out. (Median CES-D for completers = 28, interquartile range 19–34; median CES-D for dropouts = 31, interquartile range 21–39.) There was no effect of condition,

OR = 0.93, 95% CI [0.55, 1.59], p = .79. This strongly suggests that data were missing due to similar factors in both groups (i.e., somewhat higher dropout among more severely depressed participants), not the type of training offered.

Lastly, method of recording (Telegram vs. web) did not predict dropout. There was no interaction between method and condition, β = -0.77, 95% CI [-1.87, 0.32], $p = 0.17$, nor was there a significant main effect of method, $OR = 0.83$, 95% CI [0.49, 1.39], $p = .48$.

Data Analytic Plan

Data cleaning, modeling, and visualization were conducted in RStudio (version 1.0.136) running R (version 3.3.2) with the following packages: *dplyr* (Wickham et al., 2015), *tidyr* (Wickham, 2017), *ggplot2* (Wickham, 2009), and *lme4* (Bates et al., 2015).

Because we were most interested in the effect of the treatment protocol on negative self-referent processing, primary analyses focus on the completer sample. However, to assess the impact of non-compliance and missing outcomes on the intervention's efficacy, intention-to-treat analyses are also performed. Notably, non-compliance is particularly an issue with online intervention studies, as it requires relatively little effort on the part of the participant to complete baseline measures and be randomized into a treatment arm. Indeed, many participants may start an internet intervention but then drop out quickly due to time constraints, lack of motivation, technical problems, lack of interpersonal contact, perceptions of intervention ineffectiveness, among other reasons (e.g., Christensen, Griffiths, & Farrer, 2009).

Primary analyses were conducted using generalized linear mixed models that predicted trial-level binomial response (the log-odds of endorsing an SRET word as self-descriptive) using predictors of condition (PSRT, NCT), time-point (baseline,

T1, T2), valence (positive, negative), and all possible interaction terms. Random effects were modeled to allow random intercepts and slopes over time-point by participant, word, and word novelty. These primary analyses focused on the completer sample (i.e., participants who completed the baseline, T1, and T2 assessments). We also repeated analyses with the intention-to-treat sample (ITT; i.e., any participants who were randomized to PSRT or NTC and provided their ID after randomization, as per Hollis & Campbell, 1999) using the same generalized mixed-effects logistic regression models as in the completer analysis.

Secondary analyses were designed to test whether the change in SRET for the PSRT group was more strongly linked to change in CES-D as compared to the NTC group. These exploratory analyses used a simultaneous latent difference score (LDS) model, also referred to as a bivariate LDS model, a structural equation modeling (SEM) approach to the analysis of multivariate change. LDS models emphasize how individual differences in change in one variable predict individual differences in change in another variable (Newsom, 2015). Models were created using *OpenMx* in R (Neale et al., 2016; Pritikin, Hunter, & Boker, 2015; Boker et al., 2017).

Because this model is based on the covariance both within and between the two instruments (the SRET and CES-D), it is important that the same items be used for both instruments at all measurement occasions. Therefore, we did not include the novel SRET words in this analysis. To create a composite difference score for the SRET, we subtracted the number of positive words endorsed from the number of negative words endorsed for each time-point, and then subtracted the minimum of all scores across all time-points so that the measurement scale would be comparable to that of the CES-D in terms of range and variance. In order to diminish the impact of extreme values and achieve multivariate normality, scores greater than 3 median absolute deviations from the median were proportionally reduced to eliminate skew

while preserving the original rank order. This was applied to six CES-D scores at T0, two at T1, and three at T2. For example, the two highest scores, 51 and 56, were reduced to 47 and 48, while remaining the highest scores in the distribution.

In LDS models, a latent variable representing the difference score is specified for each measurement at T1 and T2, and the autoregressive path between each measurement is set equal to 1. We initially fit three major models:

1. A baseline no-change and no-group-differences independence model, that estimated only two means and two variances (one for SRET and one for CES-D), which were independent of each other, constant over time, and equivalent between groups.

2. A no-group-differences model, which, in addition to the T0 mean and variance of SRET and CES-D, estimates all of the following parameters as being equal between groups: the covariance between SRET and CES-D at baseline, the mean and variance of latent change in SRET and CES-D per week, the covariance among T0 and latent change measurements within each measure, and the simultaneous (lag-0) effect of SRET change on CES-D.

3. A model which includes all of the above parameters but estimated separately for the PSRT and NTC groups. A final model was then derived by setting specific parameters to be equal between groups if their difference did not lead to a significant improvement in fit. Models were fit to the raw data using full information maximum likelihood (FIML) estimation.

4. We also evaluated the possibility of a lagged (non-simultaneous) effect of SRET change on CES-D change, in which T1 SRET change predicts T2 CES-D change, but the addition of this path was not significant, $\chi^2(1) = 0.1$, $p = .79$.

We did not evaluate a possible mediation model, in which treatment condition's effect on CES-D was mediated by SRET change, because there was not evidence that these latent variables changed differently between conditions in previous models.

	PSRT ($N = 44$)	NTC ($N = 43$)
Age, mean (SD)	25.5 (6.2)	27.4 (7.6)
Gender		
Female	35 (80%)	37 (86%)
Male	8 (18%)	6 (14%)
Genderqueer	1 (2%)	0
Race		
White	29 (65.9%)	28 (65.1%)
Black	5 (11.3%)	3 (7.0%)
American Indian	0	1 (2.3%)
Asian	5 (11.3%)	7 (16.3%)
No response	5 (11.3%)	4 (9.3%)
Hispanic/Latino	4 (9%)	5 (12%)
CESD, mean (SD)	27.7 (11.3)	29.8 (10.2)
Current self-reported Major Depressive Episode	14 (31.8%)	12 (28%)
SRET, mean (SD) per parameter		
# Positive Words Endorsed	20.5 (11.7)	19.9 (10.3)
# Negative Words Endorsed	19.3 (10.6)	21.5 (9.7)
RT to Positive Words (ms)	1403 (1064)	1260 (559)
RT to Negative Words (ms)	1479 (1954)	1225 (603)

Table 3.1: Characteristics and symptom profiles of completer participants at baseline. CESD = Center for Epidemiologic Studies – Depression Scale, which ranges from 0 to 60. SRET = Self-Referential Encoding Task.

Results

Baseline Summary Statistics

Participants in the PSRT and NTC completer groups did not differ on age ($t(80.7) = 1.32, p = .19$), sex ($\chi^2(1) = 0.06, p = .81$), gender ($\chi^2(2) = 1.61, p = .45$), race ($\chi^2(4) = 2.80, p = .59$), ethnicity ($\chi^2(1) = 0.006, p = .94$), self-reported current major depressive episode ($\chi^2(1) = 0.007, p = .93$), or depression severity as measured by the CESD ($t(84.4) = .88, p = .38$). Both samples had elevated depression symptoms. Demographic information is presented in Table 3.1, as are endorsements and reaction times (RTs) from baseline on the SRET. A logistic mixed-effects model with factors of group and word valence predicting response at baseline found no interaction, $\beta = -0.06$, 95% CI [-0.28, 0.14], $SE = 0.11, p = .52$. Without the interaction, there was neither a significant main effect of valence ($OR = 1.0$, 95% CI [0.61, 1.62], $p = .98$) nor of condition ($OR = 0.90$, 95% CI [0.67, 1.23], $p = .52$), indicating that the treatment groups did not differ at baseline in terms of their SRET endorsements.

Effect of Training on Self-Referent Processing

To test whether the PSRT or NTC group membership changed self-schema as measured by the SRET, we defined generalized linear mixed models as described in the Data Analytic Plan above, modeling endorsement of positive or negative words as a logistic model. Results in completers indicated a significant interaction between group, valence, and a linear trend over time, $\beta = 0.33$, 95% CI [0.12, 0.55], $SE = 0.11, p = .003$.

To facilitate interpretation of the interaction, we refit the simple model of endorsement as a function of time (using the same random effects structure as the full model) within each group $\times$ valence cell, and computed the ratios of post-treatment

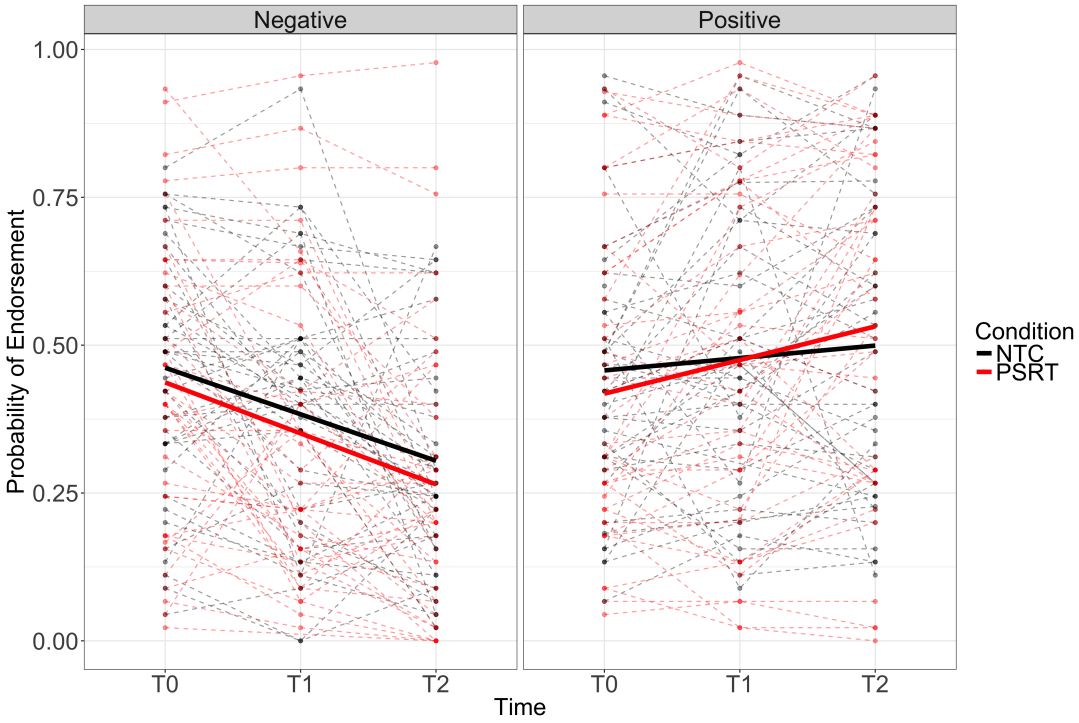

Figure 3.3: Change in probability of endorsing words over time, presented separately for negative (left) and positive (right) words. The PSRT training condition is presented in red; the NTC training condition in black. Dashed lines show individual participants' trajectories. Solid, thick lines show the linear trend of time. The probability of endorsing negative words fell for each group, while the probability of endorsing positive words increased more for the PSRT group than for the NTC group.

endorsement odds over pre-treatment endorsement odds. The PSRT group experienced a significant reduction in the odds of endorsing a negative word, $OR = 0.46$, 95% CI [0.40, 0.53], $p < .001$, as did the NTC group, $OR = 0.51$, 95% CI [0.44, 0.58], $p < .001$. Given that the OR for each group lies within the 95% CI of the other, the magnitude of improvement was not meaningfully different between groups ($p = .22$). In contrast, the PSRT group showed a significant increase in the odds of endorsing a positive word, $OR = 1.59$, 95% CI [1.40, 1.80], $p < .001$, as did the NTC group, $OR = 1.18$, 95% CI [1.04, 1.35], $p = .01$. However, unlike negative endorsements, the ORs and 95% CIs for positive endorsements indicate a meaningfully greater benefit of PSRT over NTC ($p = .002$). Thus, the 3-way interaction can be interpreted as a proportionally greater increase in positive endorsements for the PSRT group with a more equivalent decrease in negative endorsements for both PSRT and NTC groups. This comparison can be seen in Figure 3.3.

Effect of Training on Depression Symptom Change

To test whether the PSRT or NTC conditions reduced self-reported depression symptoms, we ran a mixed-effects linear regression predicting CES-D score, with predictors of training condition and a linear ordered factor of time. The interaction of condition $\times$ time was not significant, $\beta = 0.85$, 95% CI [-1.91, 3.61], $p = .55$. Without the interaction, there was no main effect of condition, $\beta = -0.96$, 95% CI [-4.90, 2.97], $p = .63$. There was a significant main effect of time ($p < .001$), with an improvement of 7.3 points, 95% CI [5.4, 9.2], in CES-D score over the course of 2 weeks. As seen in Figure 3.4, across both groups, CES-D scores reduced in a linear trend from baseline ($M = 28.7$; $SD = 10.8$) to T2 ($M = 21.4$; $SD = 10.2$), a 25% reduction.

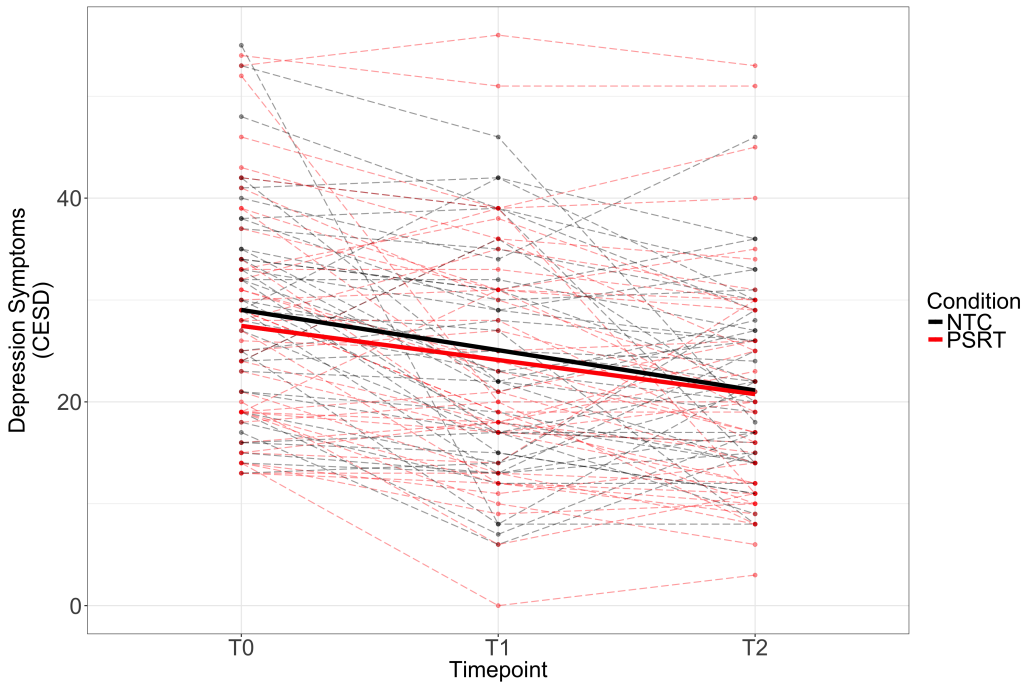

Figure 3.4: Change in depression symptoms over time. The PSRT group is in red; the NTC in black. Dashed lines show individual participants' CESD trajectories. Solid, thick lines show the trend in group means over time.

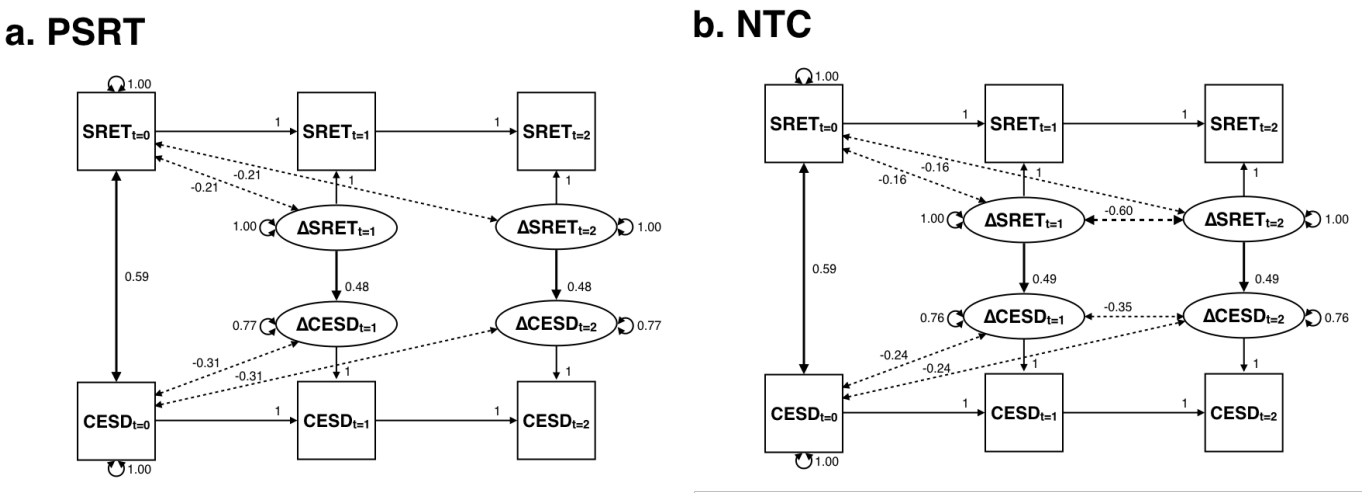

Figure 3.5: Structural equation models (SEM) created separately for PSRT and NTC groups. Lines with double-headed arrows represent covariance between variables, and lines with single-headed arrows represent causal paths. Solid lines indicate positive relationships, and dashed lines indicate negative relationships. Thicker lines represent stronger relationships. Means of SRET and CESD at T0 and the latent change variables (ΔSRET and ΔCESD) were also modeled but not shown in diagram (see Table 3.2). Mean depression, as measured by the CES-D, decreased over time for both groups. Mean SRET (measured as negative – positive) also decreased over time for both groups. Primary differences between groups' LDS models are noted in the relationship between change in SRET from T1 to T2, and change in CES-D from T1 to T2. In the PSRT group only, change in SRET and CESD at T1 is decoupled from change in SRET and CESD at T2; in the NTC group, change in SRET and CESD at T1 is significantly and negatively correlated with change in SRET and CESD at T2. This indicates, for example, that individuals in the NTC group who show more initial improvement are predicted to subsequently worsen (a negative feedback effect), whereas there was no evidence of this effect in the PSRT group. Lines in Figures 3.3 and 3.4 demonstrate varying individual trajectories of participants' change.

Latent Difference Score (LDS) Models Relating SRET Change and CES-D Change

To clarify the relationship between schema and depression change, we built a two- group simultaneous LDS model for the PSRT and NTC groups, as described in the Data Analytic Plan above. We tested hypotheses about the equality of group dynamics by evaluating the likelihood ratios of models fit with and without equality constraints in the group parameters. For this analysis, we needed to construct a composite measure of self-schema negativity that could be treated as a continuous time-varying covariate of CES-D scores. Given that the probability of negative endorsements decreased over time and the probability of positive endorsements increased for both groups, we operationalized aggregate SRET negativity as the difference between total negative endorsements and total positive endorsements, with a unit increase in this score corresponding to either one more negative, or one less positive, endorsement. We hypothesized that, despite similar group mean improvement in CES-D and SRET scores, change in the SRET would predict change in CES-D more strongly in the PSRT group than in the NTC group.

The LDS model that assumed group equivalence for all parameters included 14 parameters and was a large improvement over the baseline no-change model, $\chi^2(10) = 252.6, p < .001$. However, the overall fit of this model was not good: it was significantly different from an exact fit, $\chi^2(40) = 71.8, p = .001$, and had a Root Mean Square Error of Approximation (RMSEA) equal to 0.10, 90% CI [0.06, 0.13]. Thus, we then conducted an omnibus test of group differences by fitting the same model while allowing the groups to differ on every model parameter, thus doubling the number of estimated parameters to 28.

This model led to a further significant improvement in fit, $\chi^2(14) = 31.9, p = .004$. The overall fit of this model was better, $\chi^2(26) = 40.0, p = .04$, RMSEA = 0.08,

	PSRT		NTC			
	b (SE)	β	b (SE)	β	$\chi^2(1)$	p
Path						
SRET change → CES-D change	0.33 (0.04)	0.48	0.33 (0.04)	0.49	0.3	.57
Covariances						
SRET ↔ CES-D at baseline	76.1 (14)	0.59	76.1 (14)	0.59	2.0	.16
Regress-to-mean SRET	-29.7 (8.2)	-0.21	-29.7 (8.2)	-0.16	0.6	.44
Regress-to-mean CES-D	-20.9 (4)	-0.31	-20.9 (4)	-0.24	0.0	.92
Self-feedback SRET change	0	0	-114 (35)	-0.61	16.7	$< .001$
Self-feedback CES-D change	0	0	-29.4 (11)	-0.35	5.2	.022
Means						
SRET at baseline	39.7 (1.5)		39.7 (1.5)		0.1	.73
CES-D at baseline	28.3 (1)		28.3 (1)		0.3	.56
SRET change/week	-5.71 (0.73)		-5.71 (0.73)		0.2	.64
CES-D change/week	-1.8 (0.52)		-1.8 (0.52)		0.1	.73
Variances						
SRET at baseline	187 (27)	1.0	187 (27)	1.0	0.1	.8
CES-D at baseline	88.3 (12)	1.0	88.3 (12)	1.0	0.6	.45
SRET change/week	108 (16)	1.0	189 (35)	1.0	6.1	.014
CES-D change/week	38.3 (5.2)	0.77	63 (11)	0.76	6.5	.01

Table 3.2: For the latent difference score (LDS) models, this table presents the standardized (β) and unstandardized (b) coefficients, with their standard errors (SE). The third column presents the χ^2 and p values from the likelihood ratio test comparing models in which that parameter was either fixed to equality between groups or free to vary between groups. Coefficients that were constrained to equality by the final model are highlighted in gray. ∗The coefficients are standardized against the variance and are by definition 1 for all exogenous variables.

90% CI [0.02, 0.13], but still not good. We subsequently fit a series of models that imposed group-equality constraints on the model parameters one at a time. Those constraints that led to a significant loss in model fit were rejected. This resulted in a final model in which the groups were allowed to differ on only four parameters: variance in SRET change and disturbance in CES-D change, covariance between T1 and T2 SRET change, and covariance between T1 and T2 CES-D change. Overall, the fit of this model was good, $\chi^2(38) = 51.6, p = 0.07$, RMSEA = 0.07, 95% CI

[0, 0.11]. However, the wide confidence interval surrounding the RMSEA estimate suggests the model might benefit from a larger sample size.

The final latent difference score models can be seen in Figure 3.5. Coefficients reported here and in the figure are in standardized correlational units (i.e., -1 and 1 represent perfect negative and positive linear relationships), and Table 3.2 reports the unstandardized coefficients and their standard errors, along with the standardized coefficients (if applicable), and the χ^2 and p values from the likelihood ratio test comparing models in which that parameter was either fixed to equality between groups or free to vary between groups.

Similar to the findings from the mixed effects generalized linear models, these LDS models showed a mean reduction in depression and negative self-schema that was not significantly different between groups. Moreover, the path influence of SRET change on CES-D change was not significantly different between groups, suggesting that the PSRT intervention did not alter the fundamental relationship between self-schema and depression. However, despite these aggregate similarities at the group level, the dynamics governing within-person change appeared to differ markedly between training conditions. This was evidenced by significantly less variance in SRET change for PSRT vs NTC, $\chi^2(1) = 6.1, p = .014$; and significantly less disturbance variance in CES-D change for PSRT vs. NTC, $\chi^2(1) = 6.5, p = .011$. Particularly intriguing was the presence of a significantly greater negative self-feedback effect in the NTC group for both CES-D change, $\chi^2(1) = 5.2, p = .022$, and especially SRET change, $\chi^2(1) = 16.7, p < .001$. This refers to the covariance between successive intervals of change and reflects a tendency for individuals in the NTC to "reverse course" between weeks, e.g., larger initial improvement predicting subsequent setback. In a model in which both groups were allowed to have different non-zero self-feedback dynamics, the correlation between the two latent different scores was

substantial for the NTC group (β = -.35 for CES-D and β = -.60 for SRET) and small for the PSRT group (β = -.11 for CES-D and β = .14 for SRET). The final model reported in Figure 3.5 and Table 3.2 set both PSRT self-feedback coefficients equal to 0 without a significant loss of fit, $\chi^2(2) = 2, p = .36$.

Analyses in the Intention to Treat Sample

We repeated the primary analyses in the ITT sample, identifying the effect of training on self-schema and on depression symptoms. Where data were missing, only data from the baseline assessment were included. For self-schema, the training group × linear time × valence interaction was not significant, β = 0.11, 95% CI [-0.07, 0.30], p = .23. For depression symptoms, the training group × linear time interaction was not significant, β = 0.42, 95% CI [-2.08, 2.92], p = .74. As in the completers sample, without the interaction the main effect of group remained not significant, β = -1.76, 95% CI [-4.21, 0.69], p = .16, whereas the main effect of time was significant, β = -5.17, 95% CI [-6.43, -3.92], $p < .001$.

Subjective benefit and acceptance of training

Because we were piloting a new intervention, we also asked participants who completed the final timepoint of questionnaires to answer several questions about how they liked the training, how easy it was to complete, and how likely they were to recommend the training. These results are included in Table 3.3. Participants in both conditions generally liked the training (90%), felt that it helped them at least a moderate amount (73%), and felt that it met or exceeded their expectations (91%). Many in the PSRT condition (N = 22, 51%), felt that some of the cued scenarios were difficult to imagine. Participants in the NTC condition had marginally less difficulty, with only 14 (33%) reporting that they found imagining scenarios difficult.

	PSRT	NTC
Overall impression of the program		
Positive (4-6)	37	40
Neutral (3)	6	2
Negative (1-2)	0	1
Feeling that the program helped		
Helped a lot	6	6
Helped a little	28	23
Did not help	9	14
Meeting expectations		
Exceeded	6	8
Met	33	31
Failed to meet	4	4
Likelihood of recommending		
Yes	31	30
No	12	12

	PSRT	NTC
Ease in imagining cued scenarios		
Easy (4-5)	19	24
Neutral (3)	1	5
Difficult (1-2)	22	14
Understanding of training		
Straightforward	32	33
Confusing	15	11
Ease of technology interface		
Easy (4-5)	41	39
Neutral (3)	0	0
Difficult (1-2)	2	4
Technical Problems		
None	33	29
Yes	6	10
Yes, and it got in the way	4	4

Table 3.3: Subjective benefit and acceptance of training. Responses of completers participants to the end-of-task questionnaire on their experience of the treatment.

Discussion

These findings support the hypothesis that positive imagery training would reduce self-referential bias, as measured by the self-referential encoding task (SRET). In two randomly assigned groups of participants, over the course of a two-week period, participants in each group completed the SRET three times and recorded daily cues according to their training condition. Analyses showed that participants in the active PSRT condition showed a greater increase in the number of positive words they endorsed as self-descriptive, in comparison to the neutral NTC condition. Participants in both conditions showed a decrease in the number of negative words they endorsed as self-descriptive.

These results support the idea that repeated exposure to positive self-referent information, even if imaginal, might alter subsequent self-referent information processing. The idea that self-referentiality influences memory is not new; it is one of the key points underlying investigation of the self-reference effect (S. B. Klein et al., 1989; Hamami et al., 2011; Derry & Kuiper, 1981). The self-reference effect argues that information relating to one's own experience can facilitate memory (S. B. Klein et al., 1989)—thus, it follows that the repeated exposure in the PSRT condition may provide access to positive self-referential scenes. Importantly, the current work suggests that a focus on self-referential future events may be beneficial to biased self-referenced processing in depression. That is, this link between self-reference and memory appears to be both correlational and causal. Having salient, positive personal experiences increases the endorsement of positive terms in self-evaluation. It also improves the salience of positive self-referent imagery and increases the self-referentiality of those positive concepts.

The active training focused on positive imagery, and participants in this condition showed greater improvements in the number of positive words endorsed as

self-referent on the SRET. We did not include measures of positive affect, and thus do not know whether these participants would also have increased substantially on such questionnaires. Future work may wish to include such measures. Both trainings reduced endorsement of negative words on the SRET. The NTC training succeeded in training participants away from negative stimuli to a similar degree with the PSRT group. Both groups also showed improvement in positive self-referent processing, although the PSRT group showed a significantly greater increase in endorsement of positive adjectives on the SRET than the NTC group. We are unable to conclude whether the results of the NTC group are due to the general reduction over time of elevated depressive symptoms in a non-clinical sample, to a placebo effect, or to an unintentional efficacy of the neutral condition. While a no-training control group might have permitted a comparison with the ameliorative effects of time alone, we cannot fully disentangle our training groups from one another.

Contrary to our predictions, the PSRT condition was not associated with a greater reduction in depression symptoms, as measured by the CES-D, than the NTC condition. Both groups saw a reduction in depression symptoms over the course of the study, and this reduction was not significantly different between groups. Depression change was, however, strongly correlated with change on the SRET. Bivariate latent difference scores models, integrating both change from T0 to T1 and change from T1 to T2, demonstrated that a reduction in a negatively-valenced summary score of the SRET predicted a reduction in the CES-D. For both groups, both constructs changed together, and there was no evidence of a lagged effect of schema change on depression change, at least not over the measured one-week lag.

Despite the absence of group differences in mean change or the relationship between SRET change and CES-D change, there was evidence that the quality of change may have been different, with the NTC group showing much greater variance

in SRET and CES-D change. The starkest difference between groups was a large negative feedback correlation between initial and subsequent change for the NTC group that was completely absent for the PSRT group. This suggests that the PSRT may have disrupted a negative feedback process governing normal fluctuations in self-schema and depression, without altering the fundamental relationship between self-schema change and depression change. Extrapolating from these dynamics, one speculative hypothesis would be that the observed gains in the NTC group would be lost in the subsequent weeks, whereas the gains in the PSRT group would prove more enduring. However, the clinical significance of this finding of reduced variability and disrupted self-feedback—if any—is unknown, and a longer period of observation with a larger sample will be needed to test this idea.

Work in a larger sample could determine whether improvement was moderated by baseline severity of biased self-reference, as has been shown to be the case in larger studies investigating baseline depression severity (Bower et al., 2013). In the current study, the simultaneous change seen in SRET and CES-D supports the theory that self-referent processing plays a role in maintaining depression, given that change on the SRET strongly predicted change in symptomatology as seen in the LDS models. Indeed, such a finding fits with the causal directions predicted by the cognitive model (Beck, 1967), with negatively biased processing (and negative self-schema) helping to maintain depressive biases. Although these findings are preliminary, future work should continue to investigate the causal nature of this relationship.

One key element to the PSRT employed in this study was the aim of creating an intervention that participants could tailor to themselves (as per Hertel & Mathews, 2011). This intervention could thus be accessible to a broad range of participants, improving ease of use. Although many participants reported thinking the training was useful, some in the active PSRT condition also reported having diffi-

culty imagining the cues. Such difficulty may be an active ingredient in the efficacy of imagery interventions (E. A. Holmes et al., 2016), but is also worrisome given research suggesting that imagination may mediate the success of such interventions (Rohrbacher, Blackwell, Holmes, & Reinecke, 2014). This is a problem that bears further scrutiny in this area of research.

This study was limited in part by its taking place in a wholly online environment. Participants were not required to have any contact with the experimenters. Although they received emails every other day, these emails were automated, and many participants never emailed the experimenters directly. Such rare communication limits the ability to generalize the results to in-person therapy, although it is analogous to internet interventions that often have relatively little levels of in-person support (Meyer et al., 2009). Nonetheless, we suggest that the SRET may be very useful in a therapeutic context, and that positive imagery techniques could be useful as an adjunctive treatment alongside CBT or other psychotherapies. Especially given that many participants reported finding even the neutral intervention useful, its appeal in terms of accessibility and ease of use is important to consider.

We also acknowledge that the study had a high degree of participant attrition. Although such attrition is to be expected in online studies, it is unclear whether participants solely did not find the compensation worth their while, or whether other factors beyond those we controlled reduced study completion. This may simply be a peril of online psychology research with such small degrees of participant investment (Postel et al., 2011).

In short, we believe that the current study provides further support for the role of negative self-referent processing in maintaining depression. It additionally suggests that positive imagery may be a promising approach to increase positive self-referent processing. Continuing to explore how positive imagery may increase positive self-

schema—and reduce negative self-schema—is likely important for reducing depressive symptoms. These results may both help us in understanding how the symptoms of depression may be impacted by the salience of emotional information, and in understanding ways of modifying biased views of the self. Modifying such views may be vital to developing treatments targeting depressed mood.

General Discussion

The three studies that together form this dissertation are linked in several ways. All three explore self-referent processing, an important aspect of the self-schema, and consider this construct from different levels of analysis. Additionally, all three explore ways of using novel methods to address the measurement, etiology, and treatment of negative self-referent processing. In their gestalt, these studies provide insight into a pernicious maladaptive cognition that contributes to the experience of low mood and other depressive symptoms.

Summary of Findings

The first study focuses deeply on the self-referential encoding task (SRET), which is used throughout this dissertation as a way of measuring the self-schema. We found strong support for the linkage between participant responses on the SRET and depressive symptoms. Using a cross-validated best subsets analysis, which highlights the most important predictors in a statistically rigorous manner, we identified three or four predictors derived from the SRET that explained from 29% to 43% of the deviance in left-out test samples' depression symptoms, in three large samples. (On the training samples for each group, these predictors explained at minimum 43% of the deviance.)

The most important of these predictors were the number of endorsements of positive and negative adjectives on the SRET, and components of the diffusion model, especially the drift rate. All best models included at least one term that was valenced positive and one valenced negative, highlighting the importance of decreased

positive self-reference, as well as increased negative self-reference, in depression. This study also provided extensive evidence for the strong linkage between self-reference (as measured by the SRET) and depressive symptoms (as measured by the CES-D), underscoring the connection between self-schema and depression.

The drift diffusion model as applied to the SRET, in particular, bears further scrutiny. This computational model of decision-making is based on the assumption that decisions are made following the accumulation of information leading to a threshold for a decision (Voss & Voss, 2007). Thus, for the SRET, the threshold is the point at which a given adjective is judged to be self-referential or not. The diffusion model thus has the potential to provide additional information beyond simply what words were endorsed: it may additionally provide information about the processes underlying that decision-making process.

The drift rate in particular was of interest in this study, representing the speed and direction of the accumulation of information (Voss & Voss, 2007)—i.e., how quickly a judgment of self-reference could be made. That the diffusion model, and the drift rate in particular, were in fact valuable indicators chosen by the best subsets regression is intriguing. It suggests that this computational model—which incorporates response time, response, and the distributions of both—provides information beyond that contained by the more basic, commonly-used predictors. That is, the underlying process of self-referential decision-making may play an important role in the self-referential bias, and the diffusion model provides a window into that process. Thus, although this model is time-intensive and requires a large number of binary-choice trials, it may continue to be useful for future research on self-schema.

The second study investigated the electro-cortical associations of self-referent processing, connecting the process of making decisions about self-reference (i.e., using the SRET) with event-related potentials (ERPs) across the time-course during which

those decisions are made. We found that participants who had been diagnosed with Major Depressive Disorder (MDD) had different waveforms associated with those decisions than healthy control participants. Specifically, participants with MDD had more late-stage positive posterior ERP waveforms in response to negative words than positive words, whereas the healthy participants showed the opposite trend. These waveforms were identified as potentially driven by the late positive potential (LPP), a waveform associated with attentional engagement and increased cognitive evaluation (Shestyuk & Deldin, 2010; Auerbach et al., 2015). These findings thus suggested that the behavioral patterns identified in Study 1 resulted from increased cognitive elaboration following self-evaluation of negative information in MDD, compared to positive information.

The link between behavioral and ERP results from Study 2 provides further evidence for a biological basis to the theorized link between self-schema and depression (Disner et al., 2011). Our findings are consistent with a model that links the activation of self-schema to biased processing. This model thereby distinguishes between participants who were diagnosed with MDD and those who were not. We conclude that the observed self-referential bias may play a mechanistic role in maintaining MDD, in conjunction with other cognitive processes. We found increased cognitive elaboration and evaluative processing in later-stage ERP components, as opposed to early perceptual waveforms. Such a finding suggests that the mechanism by which self-schema may impact depression is at a stage under some degree of cognitive control—not at an early, attentional stage, as has been found in the past (Shestyuk & Deldin, 2010; Auerbach et al., 2015). We believe that the lack of early differences indicates biased control but not pre-attentional distinctions in depressed adults. Impairments in cognitive control have been consistently associated with depression (Snyder, 2013), and this may play a role in the manner in which depressed

individuals in Study 2 had their attention sustained to negative compared to positive words. Indeed, such difficulty disengaging from negative stimuli, and engaging with positive, has been shown to be endemic in depressed individuals (Koster, De Lissnyder, Derakshan, & De Raedt, 2011; Gotlib, Krasnoperova, Yue, & Joormann, 2004; Duque & Vázquez, 2015).

Given that there were consistent behavioral and neural patterns linking self-reference to depressive symptoms, as evidenced by Studies 1 and 2, the third study examined whether an intervention aimed at reducing biased self-reference would change this process, and whether it could additionally ameliorate depressed mood. We thus developed the positive self-referent imagery training (PSRT), a two-week online training which supported our hypothesis that we could modify the self-schema as measured by the SRET. Participants in the PSRT (active) condition showed substantial change on the SRET, endorsing more positive words and fewer negative words compared to their performance at baseline. Participants in the neutral training condition (NTC) had a significantly smaller increase in their endorsement of positive words compared to the PSRT group, although they still increased in their endorsement of positive words from T0 to T2. The NTC group also endorsed fewer negative words during that period, with confidence intervals that overlapped those of the PSRT group. Latent difference score models indicated that the manner of change also differed between groups, with NTC participants showing greater "oscillations" in their trajectories from T0 to T1 to T2, while PSRT groups consistently increased in positive self-schema and retained that reduction. We did not find results indicative of mediation of change in depression by change in self-reference.

Additionally, Study 3 showed that while we were able to improve self-reference differentially, depressive symptoms reduced 25% from T0 to T2 across both groups. There was no effect of condition on this change; participants across groups improved

in depressive symptomatology. Such results imply that while schema may change due to the PSRT intervention, depressive symptoms are not simultaneously impacted. Symptoms do not always change in cognitive bias modification research (MacLeod et al., 2009), but this does bear further scrutiny. Whether such simultaneous change is due to the relatively low bar to inclusion in the study (i.e., CES-D score above 13) or to the relative efficacy of the neutral training (which did inform participants that it was intended to be helpful) is unclear and cannot be explained from the data. Theory predicts that schema change should happen alongside symptom change (Beck, 1967), and these results do not preclude this possibility.

In this third study, we also tried to explicitly manipulate the self-schema, which provides the opportunity to gain further information about the role the self-schema may play in maintaining depressive symptoms. Our findings were not conclusive about this role. Depressive symptoms (measured by the CES-D) and self-reference bias (measured by the SRET) improved in both active and neutral conditions. These results indicate that while the active condition improved positive self-reference more than the neutral condition, both groups experienced reductions in depressive symptoms. Thus, this study did not demonstrate a clear link between self-reference improvement and reduction in depressive symptoms. We conclude that while self-reference may be involved in the maintenance of depressive symptoms, its improvement does not necessarily work in parallel to improvement of those symptoms. Self-reference may constitute a portion of the information contained within the construct of self-schema (S. B. Klein, 2012), and thus it is possible that our training improved self-reference but did not have substantial impact on the self-schema as a whole. Further disentangling these concepts should be a focus of subsequent studies.

Nonetheless, these studies have implications for our understanding of the mechanisms involved in depression. We found further evidence relating a lack of

positive self-schema to depressed mood. While this linkage is not theoretically novel (Beck, 1967; Pyszczynski, Holt, & Greenberg, 1987), it underscores the importance of positive information in sustaining positive mood. All the studies of this dissertation emphasize the idea that self-referent processing is associated with the maintenance of depression. Both groups in Study 3 showed increases in positive self-referent processing that were associated with improvements in depressive symptoms. Moreover, all three studies in this dissertation demonstrate that it is not solely differences in the processing of negative information but also in positive that drive mood-congruent differences. Indeed, this may be the most important conclusion of this research: it provides evidence that depression is strongly associated with a lack of positive self-schema, and that efforts to improve that lack can be effective in terms of changing valence judgments about the self. To be sure, it remains to be seen whether improvement of positive self-schema results in consistent mood improvement, and whether more emphasis on positive training can result in corresponding increased effect on depressive symptoms.

In its totality, this dissertation is able to first provide strong evidence for the psychometric reliability and validity of measuring self-schema through the SRET, then demonstrate that the cognitive elaboration inherent in self-referent decision-making is different in individuals who are depressed, and finally use these findings to develop an intervention aimed at aiding maladaptive self-schema. The statistical techniques used in these studies—best subsets regression with cross-validation in Study 1, nonparametric permutation testing in Study 2, and both linear mixed-effects logistic regression modeling of binary response data and latent difference structural equation models in Study 3—allow us to fully use individual variation in our collected data, and create models that are likely to generalize to new data. These methodologies are exciting for their potential to help us to ask and answer questions about the

contribution of self-referent processing to depression.

Indeed, one of the strengths of the studies reported here are their statistical techniques. In Study 1, we used a cross-validated best subsets regression to identify the predictors from the SRET which best predicted depressive symptoms. Cross-validation has not been extensively used in psychology—its value lies in its ability to assess how an analysis will generalize beyond the current dataset (Hastie et al., 2008). By using it in conjunction with best subsets procedures, which pick the best possible combination of variables in predicting an outcome, we were able to identify models which would generalize well beyond the current study. One potential criticism of this technique, beyond that it is complicated to explain, is its requirement of sub-setting of data—it works best with larger samples, where there is already considerable variance (Hastie et al., 2008). Generally, however, this technique provides greater confidence in estimates of prediction error of the arrived-upon models (Efron & Gong, 1983).

This confidence in both final model and error around that model is key in addressing one of the major concerns of the so-called "replication crisis" (Ioannidis, 2005; Open Science Collaboration, 2015; Maxwell, Lau, & Howard, 2015; Loken & Gelman, 2017). An additional focus identified by this research has been the need to replicate scientific findings. A large part of the impetus for Study 2 was the need to replicate and expand upon past findings, and Study 1 focuses on replicating a similar experiment in three diverse samples. Replications across samples in and out of the laboratory, and across varying ages and demographic variables, is key in identifying consistent, replicable findings. These efforts address many of the points suggested by Munafò et al. (2017); future work should also begin to pre-register studies (as done in a preliminary method in Study 3) and analyses. Additionally, data for Studies 1 and 3 are available online in data repositories, along with processing scripts. These approaches may dramatically influence the development of future

psychological science (Tackett et al., 2016).

Future Directions

Future work may choose to combine additional levels of analysis beyond those described here. For example, integrating neuroimaging with the SRET (Fossati et al., 2003; Yoshimura et al., 2009; Schmitz & Johnson, 2006) suggests that depressive symptoms may change with amygdalar functioning—and that self-referential bias may be a mechanism of change in depressive cognition. This integration has the potential to provide information about the specificity of self-referential processing to depression and to the recruitment of brain networks. By including this analysis level over multiple time-points, we would be able to gain further information as to how self-reference changes over time, and with depressive illness. Integrating event-related potentials (as in Study 2) in an intervention study would provide such evidence, by providing a biomarker to study simultaneous change during treatment. Further, including such potential mediators in analyses—especially when measured repeatedly over the course of treatment—can provide a richer and more complete picture of how treatments effect change (Miller, Dannals, & Zlatev, 2017). Treatment development could be well-served by having regular data collection in-between time-points.

This type of data collection is one of several methodologies that may be effective in producing a more complete picture of the relationship between self-reference, self-schema, and depressive symptoms. In this dissertation, we linked multiple levels of analysis—behavioral data to self-report in Study 1, behavioral and electrocortical data in Study 2, and whether a behavioral intervention can alter behavioral SRET responses across time in Study 3. Continuing in this vein involves dedicated research. We thus aim to identify behavioral tasks that are strongly and consistently linked with depressive symptom clusters; for example, identifying whether biased

self-reference is uniquely linked to cognitive or physical symptoms of depression. Such an effort provides the opportunity to both understand the neurocognitive roots of a specific bias and to determine whether we can ameliorate that bias. Even as we broaden our research to other cognitive biases, we can also hope to further explore biased self-reference. In this pursuit, future work should attempt to determine whether self-schema—using the SRET—is most negative in individuals with specific symptom clusters, or with specific behavioral patterns. Eventually, such work can help to build a model of specific subtypes of depression, and to develop more targets for tailored cognitive therapies.

We know that depression is heterogeneous (Chen et al., 2000; D. N. Klein, 2008), and that individuals display wholly different presentations of their depression. As such, if positive self-reference training eventually shows therapeutic promise, future work should consider identifying participants who are well-suited to such an intervention based on the subgroup of which they are a member. This work can only follow from large-scale data collection which identifies such subgroups that are likely to respond. This identification would additionally permit more "tailored" treatment (Mathews & Mackintosh, 2000; Hoppitt, Mathews, Yiend, & Mackintosh, 2010). Such identification could identify individuals using behavioral measures (i.e., the SRET), with self-report questionnaires, or with ERPs. Further, identifying whether these procedures are effective for individuals with general negative affect or specifically with depressive symptoms could help for further, tailored approaches.

In the near-term, we suggest that carrying out a larger-scale intervention study, targeting biased self-schema (as in Study 3), may provide information that would be useful for moving forward in understanding the mechanism of effect. Collecting longitudinal data across multiple levels of analysis may be a key method of answering the questions that this dissertation raises. Is there a dosing effect? Do

participants who receive in-person training do better? Or those who receive targeted feedback? Such questions can only be answered by a larger study carried out over a longer period—ideally, with more extensive data collection. We have shown that the SRET is sensitive to change in depression. It thus appears to be an excellent instrument for measuring change in self-schema over the course of psychotherapy or cognitive bias modification trainings. Additionally, we suggest that imagery-based practice, as that described in Study 3, may be useful as an adjunctive treatment alongside cognitive behavioral therapy or other therapies that discuss cognitions and self-schema. It may also be that combining such an intervention procedure with other cognitive bias modification techniques may provide multiple active ingredients to a broader cognitive bias modification intervention.

In brief, future randomized clinical trials targeting self-schema through any cognitive bias modification interventions should collect SRET data. Additionally, such clinical trials should include a no-training control group, as well as explicitly including groups with more elevated depression symptoms. Ideally, participants should also receive diagnoses through interviews conducted by trained experimenters. By incorporating such groups, such work would be better able to comment on the efficacy of these interventions.

Future work identifying the mechanisms behind all of the effects studied in this dissertation is also important. Is self-schema the primary driving force of the effects demonstrated by the SRET? Why does an imagery intervention appear to alter behavior on this task? Answering these questions will help in progressing our understanding of psychotherapy research broadly, and in depression specifically (Kazdin, 2007). The results from these studies suggest that the intervention explored in Study 3 altered behavior for the SRET exactly because self-schema does drive self-referential decision-making, although this may be only a subset of the information

contained in the self-schema (S. B. Klein, 2012).

Potential Limitations

Although the linked studies contained in this dissertation provide a great deal of information, they are somewhat limited in their generalizability. Studies 1 and 3 collected a good deal of data but did not diagnose participants. We recognize that non-clinical samples are excellent for providing information in some ways, as they can be representative of a community; however, given that lifetime prevalence rates even of MDD is 16.6% (Kessler et al., 2005), unselected samples are unlikely to have many clinically significant symptoms. However, it remains true that our instruments were flexible enough to be used on these varied samples. The CES-D, for example, was designed for use in the general population (Radloff, 1977). Moreover, random samples are more likely to include participants who may be unselected in studies with stringent inclusion criteria (e.g., Costa, Hari, & Kumar, 2016).

On the other hand, Study 2 diagnosed participants. However, this study was somewhat limited in size and did not provide information across a spectrum of depressive symptom degree. Ideally, such research would have large sample sizes in order to be appropriately powered, and also diagnose participants—providing information about the manifold disorders relating to negative valence. Although online research is promising and important for moving beyond WEIRD data (Western, Educated, Industrialized, Rich, and Democratic; Henrich, Heine, & Norenzayan, 2010), it results in other sample biases that are difficult to control given that we do not know the circumstances under which participants completed study measures and tasks. We nonetheless tried to reduce such biases using published guidelines that identify participants who weren't paying attention to self-report measures (Shapiro et al., 2013) or had extremely variable reaction times (Leys et al., 2013). Development of other

such methods for increasing data quality is important as online research becomes increasingly common.

We also collected some of our data in the lab, particularly in Study 2, where we collected event-linked electroencephalography data. There are, certainly, some drawbacks to such biological data collection. Neuroimaging and electro-cortical data are both limited in what they can tell us, with ERPs showing substantial temporal but poor spatial resolution (Kappenman & Luck, 2012), while fMRI data are better at localization but limited in terms of temporal resolution (Huettel, Song, & McCarthy, 2004). While these fields have advanced substantially over the years, our understanding of the link between behavior and neurobiology remains more limited than we might like (e.g., Uttal, 1998, 2001). As discussed above, one of the key strategies in confirming connections between behavior and neurobiology is repeated replication of findings (Poldrack et al., 2017). Using advanced statistical techniques, as we did in Study 2, may help to find consistent and reproducible results, but such work should be simultaneous with collecting large samples and collecting data across sites and populations. Identifying non-redundant levels of analysis to bolster findings is important for the future of clinical psychological science.

Conclusions

This research has direct clinical implications, even given the samples and moderate impact of a brief intervention. The first two studies provided evidence for the importance of recognizing the relationship between negative self-schema, lack of positive self-schema, and depressive symptoms. They indicated that self-referential decisions are impacted by specific cognitive processes that are different in individuals who are experiencing depressive symptoms. They also, as a matter of course, provided a direction for treatment that was pursued by the third study. That an in-

tervention targeting positive self-reference was successful in a near-transfer provides evidence for the importance of emphasizing positive information, as is often done, e.g., in Behavioral Activation (Martell, Addis, & Jacobson, 2001). How an online treatment such as the one described herein may aid in such work remains to be seen.

As a field, we continue to explore the relationship between negative cognitions and negative affect. It is unlikely that such research will soon reach a conclusion—how could it, with our increased understanding of the relative complexity of depressive disorders? Nonetheless, this research plays its part in defining our understanding of the cognitive structures underlying depression more clearly, and in using data-driven approaches to highlight the best markers of depressive cognition. In sum, this model of using carefully-validated behavioral tasks to assess components of mental illness across levels of analysis is an important one. It provides substantial information both in its component parts and in its whole. This line of research is promising in its outward application as well as its direct continuation. Such work does its part in untangling the complex knot of mental illness.

Contribution

I assert that I, the author of this dissertation, provided a substantial contribution to each of the three papers included herein. I developed the three study concepts with input from my supervisor. For Study 1, I designed the study with input from my supervisor and coauthors. I created all study materials. I carried out online data collection for two of the three samples. I designed analyses with input from my coauthors, and carried them out. (A statistical package to run the statistical models was written by a coauthor.) I created all figures, tables, and supplemental materials. I wrote the manuscript, which was edited and amended by my coauthors. For Study 2, I designed the study with input from my supervisor and coauthors. I created or adapted all study materials. I collected all data in person with the aid of research assistants. I processed all data according to a processing pipeline developed with input from my coauthors. I created the figures with input from my coauthors. I created all tables and carried out the majority of the statistical analyses. (Some statistical programs and analyses were developed by a coauthor.) I wrote a substantial majority of the manuscript, which was edited and amended by my coauthors. For Study 3, I designed the study with my supervisor. I created all study materials, with programming assistance on the web interface. I collected and processed all data. I designed statistical analyses with input from my coauthors. Path models were created by a coauthor. I created all figures and tables. I wrote the manuscript, which was edited and amended by my coauthors.

References

Abramson, L. Y., Seligman, M. E., & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, *87*(1), 49–74. https://doi.org/10.1037/0021-843X.87.1.49

Ahlen, J., & Ghaderi, A. (2017). Evaluation of the Children's Depression Inventory–Short Version (cdi–s). *Psychological Assessment*, *29*(9), 1157–1166. https://doi.org/10.1037/pas0000419

Allgaier, A.-K., Frühe, B., Pietsch, K., Saravo, B., Baethmann, M., & Schulte-Körne, G. (2012). Is the Children's Depression Inventory Short version a valid screening tool in pediatric care? A comparison to its full-length version. *Journal of Psychosomatic Research*, *73*(5), 369–374.

Alloy, L. B., Abramson, L. Y., Murray, L. A., Whitehouse, W. G., & Hogan, M. E. (1997). Self-referent information-processing in individuals at high and low cognitive risk for depression. *Cognition & Emotion*, *11*(5-6), 539–568.

Alloy, L. B., Abramson, L. Y., Whitehouse, W. G., Hogan, M. E., Tashman, N. A., Steinberg, D. L., ... Donovan, P. (1999). Depressogenic cognitive styles: Predictive validity, information processing and personality characteristics, and developmental origins. *Behaviour Research and Therapy*, *37*(6), 503–531. https://doi.org/10.1016/S0005-7967(98)00157-0

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Asarnow, J. R., Carlson, G. A., & Guthrie, D. (1987). Coping strategies, self-perceptions, hopelessness, and perceived family environments in depressed and suicidal children. *Journal of Consulting and Clinical Psychology*, *55*(3), 361–366. https://doi.org/10.1037/0022-006X.55.3.361

Auerbach, R. P., Bondy, E., Stanton, C. H., Webb, C. A., Shankman, S. A., & Pizzagalli, D. A. (2016). Self-referential processing in adolescents: Stability of behavioral and ERP markers. *Psychophysiology*, *53*, 1398–1406.

Auerbach, R. P., Stanton, C. H., Proudfit, G. H., & Pizzagalli, D. A. (2015). Self-referential processing in depressed adolescents: A high-density event-related potential study. *Journal of Abnormal Psychology*, *124*(2), 233-45.

Barlow, D. H., Allen, L. B., & Choate, M. L. (2004). Toward a unified treatment for emotional disorders. *Behavior Therapy*, *35*(2), 205–230. https://doi.org/

10.1016/S0005-7894(04)80036-4
Barnhoorn, J., Haasnoot, E., Bocanegra, B., & van Steenbergen, H. (2015). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, *47*(4), 918–929.
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.
Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects.* University of Pennsylvania Press.
Beck, A. T. (1974). The development of depression: A cognitive model. In R. Friedman & M. Katz (Eds.), *The psychology of depression: Contemporary theory and research.* New York: Winston-Wiley.
Beck, A. T. (1979). *Cognitive therapy and the emotional disorders.* Penguin Books.
Beck, A. T. (2008). The evolution of the cognitive model of depression and its neurobiological correlates. *American Journal of Psychiatry*, *165*, 969–977.
Beck, A. T., Brown, G., Steer, R. A., & Weissman, A. N. (1991). Factor analysis of the Dysfunctional Attitude Scale in a clinical population. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(3), 478–483.
Beck, A. T., Steer, R. A., Epstein, N., & Brown, G. (1990). Beck self-concept test. *Psychological Assessment*, *2*(2), 191–197. https://doi.org/10.1037/1040-3590.2.2.191
Beckham, E. E., Leber, W. R., Watkins, J. T., Boyer, J. L., & Cook, J. B. (1986). Development of an instrument to measure Beck's cognitive triad: The Cognitive Triad Inventory. *Journal of Consulting and Clinical Psychology*, *54*(4), 566–567. https://doi.org/10.1037/0022-006X.54.4.566
Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science. *Circulation Research*, *116*(1), 116–126.
Berto, P., D'Ilario, D., Ruffo, P., Virgilio, R. D., & Rizzo, F. (2000). Depression: Cost-of-illness studies in the international literature, a review. *The Journal of Mental Health Policy and Economics*, *3*(1), 3–10. https://doi.org/10.1002/1099-176X(200003)3:1⟨3::AID-MHP68⟩3.0.CO;2-H
Blackburn, I., Jones, S., & Lewin, R. (1986). Cognitive style in depression. *British Journal of Clinical Psychology*, *25*(4), 241–251. https://doi.org/10.1111/j.2044-8260.1986.tb00704.x
Blackburn, I., Roxborough, H., Muir, W., Glabus, M., & Blackwood, D. (1990). Perceptual and physiological dysfunction in depression. *Psychological Medicine*, *20*(1), 95–103.
Boker, S. M., Neale, M. C., Maes, H. H., Wilde, M. J., Spiegel, M., Brick, T. R.,

... Manjunath B G (2017). OpenMx 2.7.10 user guide [Computer software manual].

Boland, J., Riggs, K. J., & Anderson, R. J. (2017). A brighter future: The effect of positive episodic simulation on future predictions in non-depressed, moderately dysphoric & highly dysphoric individuals. *Behaviour Research and Therapy*, Advance online publication. https://doi.org/10.1016/j.brat.2017.10.010

Bower, P., Kontopantelis, E., Sutton, A., Kendrick, T., Richards, D. A., Gilbody, S., ... Liu, E. T.-H. (2013). Influence of initial severity of depression on effectiveness of low intensity interventions: Meta-analysis of individual patient data. *BMJ*, *346*, f540.

Bradley, B., & Mathews, A. (1983). Negative self-schemata in clinical depression. *British Journal of Clinical Psychology*, *22*(3), 173–181.

Bradley, B. P., & Mathews, A. (1988). Memory bias in recovered clinical depressives. *Cognition and Emotion*, *2*(3), 235–245.

Bradley, M. M., & Lang, P. J. (2010). *Affective norms for english words (ANEW): Stimuli, instruction manual, and affective ratings (Technical report C-2)* (Tech. Rep.). Gainesville, FL: The Center for Research in Psychophysiology, University of Florida.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5.

Cameron, A. C., & Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, *77*(2), 329–342.

Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., ... others (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science*, *301*(5631), 386–389. https://doi.org/10.1126/science.1083968

Chen, L.-S., Eaton, W. W., Gallo, J. J., & Nestadt, G. (2000). Understanding the heterogeneity of depression through the triad of symptoms, course and risk factors: A longitudinal, population-based study. *Journal of Affective Disorders*, *59*(1), 1–11.

Christensen, H., Griffiths, K. M., & Farrer, L. (2009). Adherence in internet interventions for anxiety and depression: Systematic review. *Journal of Medical Internet Research*, *11*(2), e13. https://doi.org/10.2196/jmir.1194

Clark, L. A., & Watson, D. (1991). Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications. *Journal of Abnormal Psy-*

chology, *100*(3), 316–336.

Clifford, P. I., & Hemsley, D. R. (1987). The influence of depression on the processing of personal attributes. *The British Journal of Psychiatry*, *150*(1), 98–103. https://doi.org/10.1192/bjp.150.1.98

Comblain, C., D'Argembeau, A., & Van der Linden, M. (2005). Phenomenal characteristics of autobiographical memories for emotional and neutral events in older and younger adults. *Experimental Aging Research*, *31*(2), 173–189. https://doi.org/10.1080/03610730590915010

Connolly, S. L., Abramson, L. Y., & Alloy, L. B. (2015). Information processing biases concurrently and prospectively predict depressive symptoms in adolescents: Evidence from a self-referent encoding task. *Cognition and Emotion*, *30*(3), 550–560.

Costa, L. J., Hari, P. N., & Kumar, S. K. (2016). Differences between unselected patients and participants in multiple myeloma clinical trials in US: A threat to external validity. *Leukemia & Lymphoma*, *57*(12), 2827–2832. https://doi.org/10.3109/10428194.2016.1170828

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Medicine*, *11*, 126. https://doi.org/10.1186/1741-7015-11-126

Cuthbert, B. N., Schupp, H. T., Bradley, M. M., Birbaumer, N., & Lang, P. J. (2000). Brain potentials in affective picture processing: Covariation with autonomic arousal and affective report. *Biological Psychology*, *52*(2), 95–111.

Dainer-Best, J. (2017). *Guided imagery training and self-referential processing.* Retrieved from `https://clinicaltrials.gov/ct2/show/NCT03056963` (Identifier NCT03056963)

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

Dainer-Best, J., Shumake, J. D., & Beevers, C. G. (2018). *Positive imagery training increases positive self-referent cognition in depression.* https://doi.org/10.17605/OSF.IO/7XKHG

Dainer-Best, J., Trujillo, L. T., Schnyer, D. M., & Beevers, C. G. (2017). Sustained engagement of attention is associated with increased negative self-referent processing in major depressive disorder. *Biological Psychology*, *129*, 231–241. https://doi.org/10.1016/j.biopsycho.2017.09.005

Davis, H. (1979). Self-reference and the encoding of personal information in depres-

sion. *Cognitive Therapy and Research*, *3*(1), 97–110.

Delplanque, S., Lavoie, M. E., Hot, P., Silvert, L., & Sequeira, H. (2004). Modulation of cognitive processing by emotional valence studied through event-related potentials in humans. *Neuroscience Letters*, *356*(1), 1–4.

Derry, P. A., & Kuiper, N. A. (1981). Schematic processing and self-reference in clinical depression. *Journal of Abnormal Psychology*, *90*(4), 286–297.

Dimidjian, S., Hollon, S. D., Dobson, K. S., Schmaling, K. B., Kohlenberg, R. J., Addis, M. E., ... Jacobson, N. S. (2006). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology*, *74*(4), 658–670. https://doi.org/10.1037/0022-006X.74.4.658

Disner, S. G., Beevers, C. G., Haigh, E. A., & Beck, A. T. (2011). Neural mechanisms of the cognitive model of depression. *Nature Reviews Neuroscience*, *12*(8), 467–477.

Disner, S. G., Shumake, J. D., & Beevers, C. G. (2016). Self-referential schemas and attentional bias predict severity and naturalistic course of depression symptoms. *Cognition and Emotion*, *31*(4), 632–644. https://doi.org/10.1080/02699931.2016.1146123

Dobson, K. S. (1989). A meta-analysis of the efficacy of cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, *57*(3), 414–419. https://doi.org/10.1037/0022-006X.57.3.414

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.

Doost, H. T. N., Moradi, A. R., Taghavi, M. R., Yule, W., & Dalgleish, T. (1999). The development of a corpus of emotional words produced by children and adolescents. *Personality and Individual Differences*, *27*(3), 433–451.

Dozois, D. J., & Beck, A. T. (2008). Cognitive schemas, beliefs and assumptions. In K. S. Dobson & D. J. Dozois (Eds.), *Risk factors in depression* (pp. 121–143). Elsevier.

Dozois, D. J., Bieling, P. J., Patelis-Siotis, I., Hoar, L., Chudzik, S., McCabe, K., & Westra, H. A. (2009). Changes in self-schema structure in cognitive therapy for major depressive disorder: A randomized clinical trial. *Journal of Consulting and Clinical Psychology*, *77*(6), 1078–1088. https://doi.org/10.1037/a0016886

Dozois, D. J., & Dobson, K. S. (2001). A longitudinal investigation of information processing and cognitive organization in clinical depression: Stability of schematic interconnectedness. *Journal of Consulting and Clinical Psychology*,

69(6), 914–925.

Duque, A., & Vázquez, C. (2015). Double attention bias for positive and negative emotional faces in clinical depression: Evidence from an eye-tracking study. *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 107–114. https://doi.org/10.1016/j.jbtep.2014.09.005

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*(1), 36–48. https://doi.org/10.1080/00031305.1983.10483087

Everaert, J., Koster, E. H., & Derakshan, N. (2012). The combined cognitive bias hypothesis in depression. *Clinical Psychology Review*, *32*(5), 413–424.

Ferrari, G. R., Möbius, M., van Opdorp, A., Becker, E. S., & Rinck, M. (2016). Can't look away: An eye-tracking based attentional disengagement training for depression. *Cognitive Therapy and Research*, *40*(5), 672–686.

Fossati, P., Hevenor, S. J., Graham, S. J., Grady, C., Keightley, M. L., Craik, F., & Mayberg, H. (2003). In search of the emotional self: An fMRI study using positive and negative emotional words. *American Journal of Psychiatry*, *160*(11), 1938–1945.

Goldin, P. R., Jazaieri, H., Ziv, M., Kraemer, H., Heimberg, R. G., & Gross, J. J. (2013). Changes in positive self-views mediate the effect of cognitive-behavioral therapy for social anxiety disorder. *Clinical Psychological Science*, *1*(3), 301–310.

Goldstein, B. L., Hayden, E. P., & Klein, D. N. (2015). Stability of self-referent encoding task performance and associations with change in depressive symptoms from early to middle childhood. *Cognition and Emotion*, *29*(8), 1445–1455.

Gotlib, I. H., & Joormann, J. (2010). Cognition and depression: Current status and future directions. *Annual Review of Clinical Psychology*, *6*, 285–312.

Gotlib, I. H., Kasch, K. L., Traill, S., Joormann, J., Arnow, B. A., & Johnson, S. L. (2004). Coherence and specificity of information-processing biases in depression and social phobia. *Journal of Abnormal Psychology*, *113*(3), 386–398.

Gotlib, I. H., & Krasnoperova, E. (1998). Biased information processing as a vulnerability factor for depression. *Behavior Therapy*, *29*(4), 603–617.

Gotlib, I. H., Krasnoperova, E., Yue, D. N., & Joormann, J. (2004). Attentional biases for negative interpersonal stimuli in clinical depression. *Journal of Abnormal Psychology*, *113*(1), 127–135. https://doi.org/10.1037/0021-843X.113.1.121

Greenberg, M. S., & Alloy, L. B. (1989). Depression versus anxiety: Processing of self-and other-referent information. *Cognition and Emotion*, *3*(3), 207–223.

Greenberg, M. S., & Beck, A. T. (1989). Depression versus anxiety: A test of the content-specificity hypothesis. *Journal of Abnormal Psychology*, *98*(1), 9–13.

Greenberg, P. E., Fournier, A.-A., Sisitsky, T., Pike, C. T., & Kessler, R. C. (2015). The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *The Journal of Clinical Psychiatry*, *76*(2), 155–162. https://doi.org/10.4088/JCP.14m09298

Haaga, D. A., Dyck, M. J., & Ernst, D. (1991). Empirical status of cognitive theory of depression. *Psychological Bulletin*, *110*(2), 215–236. https://doi.org/10.1037/0033-2909.110.2.215

Hajcak, G., MacNamara, A., Foti, D., Ferri, J., & Keil, A. (2013). The dynamic allocation of attention to emotion: Simultaneous and independent evidence from the late positive potential and steady state visual evoked potentials. *Biological Psychology*, *92*(3), 447–455.

Hajcak, G., Weinberg, A., MacNamara, A., & Foti, D. (2012). ERPs and the study of emotion. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components* (pp. 441–474). Oxford University Press.

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.

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

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.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Hauser, D. J., & Schwarz, N. (2016). Attentive turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hayden, E. P., Olino, T. M., Mackrell, S. V., Jordan, P. L., Desjardins, J., & Katsiroumbas, P. (2013). Cognitive vulnerability to depression during middle childhood: Stability and associations with maternal affective styles and parental depression. *Personality and Individual Differences*, *55*(8), 892–897.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*, 29.

Hertel, P. T., & Mathews, A. (2011). Cognitive bias modification past perspectives, current findings, and future applications. *Perspectives on Psychological Science*, *6*(6), 521–536.

Hitchcock, C., Mueller, V., Hammond, E., Rees, C., Werner-Seidler, A., & Dalgleish, T. (2016). The effects of autobiographical memory flexibility (MemFlex) training: An uncontrolled trial in individuals in remission from depression. *Journal of Behavior Therapy and Experimental Psychiatry*, *52*, 92–98.

Hofmann, S. G., Asnaani, A., Vonk, I. J., Sawyer, A. T., & Fang, A. (2012). The efficacy of cognitive behavioral therapy: A review of meta-analyses. *Cognitive Therapy and Research*, *36*(5), 427–440. https://doi.org/10.1007/s10608-013-9595-3

Hollis, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ*, *319*(7211), 670–674.

Holmes, A. P., Blair, R., Watson, J., & Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism*, *16*(1), 7–22.

Holmes, E. A., Blackwell, S. E., Burnett Heyes, S., Renner, F., & Raes, F. (2016). Mental imagery in depression: Phenomenology, potential mechanisms, and treatment implications. *Annual Review of Clinical Psychology*, *12*, 249–280.

Holmes, E. A., Lang, T. J., & Shah, D. M. (2009). Developing interpretation bias modification as a "cognitive vaccine" for depressed mood: Imagining positive events makes you feel better than thinking about them verbally. *Journal of Abnormal Psychology*, *118*(1), 76–88.

Holmes, E. A., & Mathews, A. (2010). Mental imagery in emotion and emotional disorders. *Clinical Psychology Review*, *30*(3), 349–362.

Hoppitt, L., Mathews, A., Yiend, J., & Mackintosh, B. (2010). Cognitive bias modification: The critical role of active training in modifying emotional responses. *Behavior Therapy*, *41*(1), 73–81.

Huettel, S. A., Song, A. W., & McCarthy, G. (2004). *Functional magnetic resonance imaging* (2nd ed.). Sunderland, MA: Sinauer Associates.

Iacoviello, B. M., Wu, G., Abend, R., Murrough, J. W., Feder, A., Fruchter, E., ... others (2014). Attention bias variability and symptoms of posttraumatic stress disorder. *Journal of Traumatic Stress*, *27*(2), 232–239.

Ingram, R. E., Kendall, P. C., Smith, T. W., Donnell, C., & Ronan, K. (1987). Cognitive specificity in emotional distress. *Journal of Personality and Social Psychology*, *53*(4), 734–742.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS*

Medicine, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jacobson, N. S., Dobson, K. S., Truax, P. A., Addis, M. E., Koerner, K., Gollan, J. K., ... Prince, S. E. (1996). A component analysis of cognitive-behavioral treatment for depression. *Journal of Consulting and Clinical Psychology*, *64*(2), 295–304. https://doi.org/10.1037/0022-006X.64.2.295

Johnson, S. L., Joormann, J., & Gotlib, I. H. (2007). Does processing of emotional stimuli predict symptomatic improvement and diagnostic recovery from major depression? *Emotion*, *7*(1), 201–206. https://doi.org/10.1037/1528-3542.7.1.201

Joormann, J., Waugh, C. E., & Gotlib, I. H. (2015). Cognitive bias modification for interpretation in major depression: Effects on memory and stress reactivity. *Clinical Psychological Science*, *3*(1), 126–139.

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., Mckeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(02), 163–178.

Kappenman, E. S., & Luck, S. J. (2012). ERP components: The ups and downs of brainwave recordings. In S. J. Luck & E. S. Kappenman (Eds.), *The Oxford handbook of event-related potential components* (pp. 3–30). Oxford University Press.

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology Annual Review of Clinical Psychology*, *3*, 1–27.

Keil, A., Debener, S., Gratton, G., Junghöfer, M., Kappenman, E. S., Luck, S. J., ... Yee, C. M. (2014, Jan). Committee report: Publication guidelines and recommendations for studies using electroencephalography and magnetoencephalography. *Psychophysiology*, *51*(1), 1–21.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*, *14*(5), 785–794.

Kelvin, R. G., Goodyer, I. M., Teasdale, J. D., & Brechin, D. (1999). Latent negative self-schema and high emotionality in well adolescents at risk for psychopathology. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *40*(6), 959–968. https://doi.org/10.1111/1469-7610.00513

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the national comorbidity survey replication. *Archives of general psychiatry*, *62*(6), 593–602.

Kiang, M., Farzan, F., Blumberger, D. M., Kutas, M., McKinnon, M. C., Kansal, V., ... Daskalakis, Z. J. (2017). Abnormal self-schema in semantic memory in major depressive disorder: Evidence from event-related brain potentials. *Biological Psychology*, *126*, 41–47.

Klein, D. N. (2008). Classification of depressive disorders in the DSM-V: Proposal for a two-dimension system. *Journal of Abnormal Psychology*, *117*(3), 552–560.

Klein, S. B. (2012). Self, memory, and the self-reference effect: An examination of conceptual and methodological issues. *Personality and Social Psychology Review*, *16*(3), 283–300. https://doi.org/10.1177/1088868311434214

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

Klein, S. B., & Loftus, J. (1988). The nature of self-referent encoding: The contributions of elaborative and organizational processes. *Journal of Personality and Social Psychology*, *55*(1), 5–11. https://doi.org/10.1037/0022-3514.55.1.5

Klein, S. B., Loftus, J., & Burton, H. A. (1989). Two self-reference effects: The importance of distinguishing between self-descriptiveness judgments and autobiographical retrieval in self-referent encoding. *Journal of Personality and Social Psychology*, *56*(6), 853–865.

Koster, E. H., De Lissnyder, E., Derakshan, N., & De Raedt, R. (2011). Understanding depressive rumination from a cognitive science perspective: The impaired disengagement hypothesis. *Clinical psychology review*, *31*(1), 138–145. https://doi.org/10.1016/j.cpr.2010.08.005

Kovacs, M. (1981). Rating scales to assess depression in school-aged children. *Acta Paedopsychiatrica: International Journal of Child & Adolescent Psychiatry*.

Kovacs, M. (1992). *Children's Depression Inventory*. Multi-Health System North Tonawanda, NY.

Kovacs, M. (2003). *Children's depression inventory (CDI): Technical manual update* (Tech. Rep.). North Tonawanda, NY: Multi-Health Systems.

Kruijt, A.-W., Field, A. P., & Fox, E. (2016). Capturing dynamics of biased attention: Are new attention variability measures the way forward? *PloS One*, *11*(11), e0166600.

Kuiper, N. A., & Derry, P. A. (1982). Depressed and nondepressed content self-reference in mild depressives. *Journal of Personality*, *50*(1), 67–80.

Kuiper, N. A., Olinger, L. J., MacDonald, M. R., & Shaw, B. F. (1985). Self-schema processing of depressed and nondepressed content: The effects of vulnerability to depression. *Social Cognition*, *3*(1), 77–93.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in linear mixed effects models [Computer software manual]. (R package version 2.0-32)

Lee, T.-W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation*, *11*(2), 417–441.

LeMoult, J., Colich, N., Joormann, J., Singh, M. K., Eggleston, C., & Gotlib, I. H. (2017). Interpretation bias training in depressed adolescents: Near-and far-transfer effects. *Journal of Abnormal Child Psychology*, 1–9.

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.

Leys, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766.

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 1–10.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://doi.org/10.1126/science.aal3618

Luck, S. J., & Hillyard, S. A. (1994). Spatial filtering during visual search: Evidence from human electrophysiology. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1000.

MacDonald, M. R., & Kuiper, N. A. (1985). Efficiency and automaticity of self-schema processing in clinical depressives. *Motivation and Emotion*, *9*(2), 171–184.

MacLeod, C., Koster, E. H., & Fox, E. (2009). Whither cognitive bias modification research? Commentary on the special section articles. *Journal of Abnormal Psychology*, *118*(1), 89–99.

MacNamara, A., Kotov, R., & Hajcak, G. (2016). Diagnostic and symptom-based predictors of emotional processing in generalized anxiety disorder and major depressive disorder: An event-related potential study. *Cognitive Therapy and Research*, *40*(3), 275–289. https://doi.org/10.1007/s10608-015-9717-1

Martell, C. R., Addis, M. E., & Jacobson, N. S. (Eds.). (2001). *Depression in context: Strategies for guided action*. New York: W.W. Norton & Co.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23.

Mathews, A., & Mackintosh, B. (2000). Induced emotional interpretation bias and anxiety. *Journal of abnormal psychology*, *109*(4), 602.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. https://doi.org/10.1037/a0039400

McCauley, E., Mitchell, J. R., Burke, P. M., & Moss, S. J. (1988). Cognitive attributes of depression in children and adolescents. *Journal of Consulting and Clinical Psychology*, *56*(6), 903–908. https://doi.org/10.1037/0022-006X.56.6.903

Meyer, B., Berger, T., Caspar, F., Beevers, C. G., Andersson, G., & Weiss, M. (2009). Effectiveness of a novel integrative online treatment for depression (Deprexis): Randomized controlled trial. *Journal of Medical Internet Research*, *11*(2), e15.

Miller, D. T., Dannals, J. E., & Zlatev, J. J. (2017). Behavioral processes in long-lag intervention studies. *Perspectives on Psychological Science*, *12*(3), 454–467. https://doi.org/10.1177/1745691616681645

Mineka, S., & Nugent, K. (1995). Mood-congruent memory biases in anxiety and depression. In D. L. Schacter (Ed.), *Memory distortion: How minds, brains, and societies reconstruct the past* (pp. 173–193). Cambridge, MA: Harvard University Press.

Moher, D., Schulz, K. F., & Altman, D. G. (n.d.). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *The Lancet*, *357*(9263), 1191–1194. https://doi.org/10.1016/S0140-6736(00)04337-3

Monroe, S. M., & Simons, A. D. (1991). Diathesis-stress theories in the context of life stress research: Implications for the depressive disorders. *Psychological Bulletin*, *110*(3), 406–425. https://doi.org/10.1037/0033-2909.110.3.406

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective*. MIT press.

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kickpatrick, R. M., ... Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535–549. https://doi.org/10.1007/s11336-014-9435-8

Neshat-Doost, H. T., Dalgleish, T., Yule, W., Kalantari, M., Ahmadi, S. J., Dyregrov, A., & Jobson, L. (2013). Enhancing autobiographical memory specificity through cognitive training: An intervention for depression translated from basic

science. *Clinical Psychological Science*, *1*(1), 84–92. https://doi.org/10.1177/2167702612454613

Newsom, J. T. (2015). Longitudinal structural equation modeling: A comprehensive introduction. In (p. 258). New York, NY: Routledge.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25.

Nolen-Hoeksema, S. (2000). The role of rumination in depressive disorders and mixed anxiety/depressive symptoms. *Journal of Abnormal Psychology*, *109*(3), 504–511. https://doi.org/10.1037/0021-843X.109.3.504

Oliver, J., & Baumgart, E. P. (1985). The Dysfunctional Attitude Scale: Psychometric properties and relation to depression in an unselected adult population. *Cognitive Therapy and Research*, *9*(2), 161–167.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Overmier, J. B., & Seligman, M. E. (1967). Effects of inescapable shock upon subsequent escape and avoidance responding. *Journal of Comparative and Physiological Psychology*, *63*(1), 28–33. https://doi.org/10.1037/h0024166

Padesky, C. A. (1994). Schema change processes in cognitive therapy. *Clinical Psychology & Psychotherapy*, *1*(5), 267–278. https://doi.org/10.1002/cpp.5640010502

Pernet, C., Latinus, M., Nichols, T. E., & Rousselet, G. A. (2015). Cluster-based computational methods for mass univariate analyses of event-related brain potentials/fields: A simulation study. *Journal of Neuroscience Methods*, *250*, 85–93.

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

Poldrack, R. A., Baker, C. I., Durnez, J., Gorgolewski, K. J., Matthews, P. M., Munafò, M. R., ... Yarkoni, T. (2017). Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nature Reviews Neuroscience*, *18*(2), 115–126. https://doi.org/doi:10.1038/nrn.2016.167

Postel, M. G., de Haan, H. A., Ter Huurne, E. D., van der Palen, J., Becker, E. S., & de Jong, C. A. (2011). Attrition in web-based treatment for problem drinkers. *Journal of Medical Internet Research*, *13*(4), e117. https://doi.org/10.2196/jmir.1811

Price, J., Sloman, L., Gardner, R., Gilbert, P., & Rohde, P. (1994). The social com-

petition hypothesis of depression. *The British Journal of Psychiatry*, *164*(3), 309–315. https://doi.org/10.1192/bjp.164.3.309

Prieto, S. L., Cole, D. A., & Tageson, C. W. (1992). Depressive self-schemas in clinic and nonclinic children. *Cognitive Therapy and Research*, *16*(5), 521–534.

Pritikin, J. N., Hunter, M. D., & Boker, S. M. (2015). Modular open-source software for Item Factor Analysis. *Educational and Psychological Measurement*, *75*(3), 458–474.

Proudfit, G. H. (2015). The reward positivity: From basic research on reward to a biomarker for depression. *Psychophysiology*, *52*(4), 449–459. https://doi.org/10.1111/psyp.12370

Proudfit, G. H., Bress, J. N., Foti, D., Kujawa, A., & Klein, D. N. (2015). Depression and event-related potentials: Emotional disengagement and reward insensitivity. *Current Opinion in Psychology*, *4*, 110–113.

Pyszczynski, T., Holt, K., & Greenberg, J. (1987). Depression, self-focused attention, and expectancies for positive and negative future life events for self and others. *Journal of Personality and Social Psychology*, *52*(5), 994–1001. https://doi.org/10.1037/0022-3514.52.5.994

Quilty, L. C., Dozois, D. J., Lobo, D. 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.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401.

Raes, F., Williams, J. M. G., & Hermans, D. (2009). Reducing cognitive vulnerability to depression: A preliminary investigation of MEmory Specificity Training (MEST) in inpatients with depressive symptomatology. *Journal of Behavior Therapy and Experimental Psychiatry*, *40*(1), 24–38. https://doi.org/10.1016/j.jbtep.2008.03.001

Ramautar, J., Kok, A., & Ridderinkhof, K. (2004). Effects of stop-signal probability in the stop-signal paradigm: The N2/P3 complex further validated. *Brain and Cognition*, *56*(2), 234–252. https://doi.org/10.1016/j.bandc.2004.07.002

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Re, A. C. D. (2013). compute.es: Compute effect sizes [Computer software manual].

Rehm, L. P. (1977). A self-control model of depression. *Behavior Therapy*, *8*(5),

787–804. https://doi.org/10.1016/S0005-7894(77)80150-0

Revelle, W. (2016). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. (R package version 1.6.4)

Rodebaugh, T. L., Scullin, R. B., Langer, J. K., Dixon, D. J., Huppert, J. D., Bernstein, A., ... Lenze, E. J. (2016). Unreliability as a threat to understanding psychopathology: The cautionary tale of attentional bias. *Journal of Abnormal Psychology*, *125*(6), 840–851.

Rohrbacher, H., Blackwell, S. E., Holmes, E. A., & Reinecke, A. (2014). Optimizing the ingredients for imagery-based interpretation bias modification for depressed mood: Is self-generation more effective than imagination alone? *Journal of Affective Disorders*, *152-154*, 212–218.

Sanguinetti, J. L., Trujillo, L. T., Schnyer, D. M., Allen, J. J., & Peterson, M. A. (2015). Increased alpha band activity indexes inhibitory competition across a border during figure assignment. *Vision Research*, *126*, 120–130.

Santor, D. A., Zuroff, D. C., Ramsay, J., Cervantes, P., & Palacios, J. (1995). Examining scale discriminability in the BDI and CES-D as a function of depressive severity. *Psychological Assessment*, *7*(2), 131–139.

Sanz, J. (1996). Memory biases in social anxiety and depression. *Cognition and Emotion*, *10*(1), 87–106.

Schmajuk, M., Liotti, M., Busse, L., & Woldorff, M. G. (2006). Electrophysiological activity underlying inhibitory control processes in normal adults. *Neuropsychologia*, *44*(3), 384–395. https://doi.org/10.1016/j.neuropsychologia.2005.06.005

Schmidt, N. B., Joiner, T. E., Young, J. E., & Telch, M. J. (1995). The schema questionnaire: Investigation of psychometric properties and the hierarchical structure of a measure of maladaptive schemas. *Cognitive Therapy and Research*, *19*(3), 295–321.

Schmitz, T. W., & Johnson, S. C. (2006). Self-appraisal decisions evoke dissociated dorsal—ventral aMPFC networks. *NeuroImage*, *30*(3), 1050–1058. https://doi.org/10.1016/j.neuroimage.2005.10.030

Schupp, H., Cuthbert, B., Bradley, M., Hillman, C., Hamm, A., & Lang, P. (2004). Brain processes in emotional perception: Motivated attention. *Cognition and Emotion*, *18*(5), 593–611.

Segal, Z. V. (1988). Appraisal of the self-schema construct in cognitive models of depression. *Psychological Bulletin*, *103*(2), 147–162.

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to

study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220.
Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., ... Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *Journal of Clinical Psychiatry*, *59*(Suppl 20), 22–33.
Shestyuk, A. Y., & Deldin, P. J. (2010). Automatic and strategic representation of the self in major depression: Trait and state abnormalities. *American Journal of Psychiatry*, *167*, 536–544.
Shumake, J. (2016). beset: Tools for data exploration and mining [Computer software manual]. (R package version 0.0.0.9000)
Snyder, H. R. (2013). Major depressive disorder is associated with broad impairments on neuropsychological measures of executive function: A meta-analysis and review. *Psychological Bulletin*, *139*(1), 81–132. https://doi.org/10.1037/a0028727
Speed, B. C., Nelson, B. D., Auerbach, R. P., Klein, D. N., & Hajcak, G. (2016). Depression risk and electrocortical reactivity during self-referential emotional processing in 8 to 14 year-old girls. *Journal of Abnormal Psychology*, *125*(5), 607–619.
Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*(3), 371–394.
Tackett, J. L., Lilienfeld, S., Patrick, C., Johnson, S., Krueger, R., Miller, J., ... Shrout, P. E. (2016, Dec). *Clinical science and the replicability crisis.* Open Science Framework.
Tarlow, E. M., & Haaga, D. A. (1996). Negative self-concept: Specificity to depressive symptoms and relation to positive and negative affectivity. *Journal of Research in Personality*, *30*(1), 120–127. https://doi.org/10.1006/jrpe.1996.0008
Taylor, L., & Ingram, R. E. (1999). Cognitive reactivity and depressotypic information processing in children of depressed mothers. *Journal of Abnormal Psychology*, *108*(2), 202–210. https://doi.org/10.1037/0021-843X.108.2.202
Teasdale, J. D. (1988). Cognitive vulnerability to persistent depression. *Cognition and Emotion*, *2*(3), 247–274.
Teasdale, J. D., & Russell, M. L. (1983). Differential effects of induced mood on the recall of positive, negative and neutral words. *British Journal of Clinical Psychology*, *22*(3), 163–171. https://doi.org/10.1111/j.2044-8260.1983.tb00597.x
Thurston, M. D., Goldin, P., Heimberg, R., & Gross, J. J. (2017). Self-views in social anxiety disorder: The impact of CBT versus MBSR. *Journal of Anxiety*

Disorders, *47*, 83–90.

Timbremont, B., & Braet, C. (2004). Cognitive vulnerability in remitted depressed children and adolescents. *Behaviour Research and Therapy*, *42*(4), 423–437. https://doi.org/10.1016/S0005-7967(03)00151-7

Timbremont, B., Braet, C., Bosmans, G., & Van Vlierberghe, L. (2008). Cognitive biases in depressed and non-depressed referred youth. *Clinical Psychology and Psychotherapy*, *15*(5), 329–339. https://doi.org/10.1002/cpp.579

Trujillo, L. T., Allen, J. J. B., Schnyer, D. M., & Peterson, M. A. (2010). Neurophysiological evidence for the influence of past experience on figure-ground perception. *Journal of Vision*, *10*(2), 5. 1–21.

Uttal, W. R. (1998). *Toward a new behaviorism: The case against perceptual reductionism*. Mahwah, NJ: Lawrence Erlbaum Associates.

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. Cambridge, MA: The MIT press.

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology. *Experimental Psychology*, *60*, 385–402.

Voss, A., & Voss, J. (2007). Fast-dm: A free program for efficient diffusion model analysis. *Behavior Research Methods*, *39*(4), 767–775.

Walker, W. R., Skowronski, J. J., & Thompson, C. P. (2003). Life is pleasant–and memory helps to keep it that way! *Review of General Psychology*, *7*(2), 203–210.

Wardenaar, K. J., van Veen, T., Giltay, E. J., de Beurs, E., Penninx, B. W., & 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.

Waters, A. C., & Tucker, D. M. (2016). Principal components of electrocortical activity during self-evaluation indicate depressive symptom severity. *Social Cognitive and Affective Neuroscience*, *11*(8), 1335-1343.

Watkins, E. R., Baeyens, C. B., & Read, R. (2009). Concreteness training reduces dysphoria: proof-of-principle for repeated cognitive bias modification in depression. *Journal of abnormal psychology*, *118*(1), 55.

Watkins, P. C., Mathews, A., Williamson, D. A., & Fuller, R. D. (1992). Mood-congruent memory in depression: Emotional priming or elaboration? *Journal of Abnormal Psychology*, *101*(3), 581–586. https://doi.org/10.1037/0021-843X.101.3.581

Watkins, P. C., Vache, K., Verney, S. P., & Mathews, A. (1996). Unconscious mood-congruent memory bias in depression. *Journal of Abnormal Psychology*,

105(1), 34–41.

Weinberg, A., Perlman, G., Kotov, R., & Hajcak, G. (2016). Depression and reduced neural response to emotional images: Distinction from anxiety, and importance of symptom dimensions and age of onset. *Journal of Abnormal Psychology*, *125*(1), 26–39.

Weissman, A. N., & Beck, A. T. (1978, November). Development and validation of the Dysfunctional Attitude Scale: A preliminary investigation. In *Paper presented at the meeting of the association for the advancement of behavior therapy, chicago.*

Wells, T. T., & Beevers, C. G. (2010). Biased attention and dysphoria: Manipulating selective attention reduces subsequent depressive symptoms. *Cognition and Emotion*, *24*(4), 719–728.

White, C. N., Ratcliff, R., Vasey, M. W., & McKoon, G. (2010). Using diffusion models to understand clinical disorders. *Journal of Mathematical Psychology*, *54*, 39–52.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, *7:1832*.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York.

Wickham, H. (2015). tidyr: Easily tidy data with 'spread()' and 'gather()' functions [Computer software manual]. (R package version 0.3.1)

Wickham, H. (2017). tidyr: Easily tidy data with 'spread()' and 'gather()' functions [Computer software manual]. (R package version 0.6.1)

Wickham, H., Francois, R., Henry, L., & Müller, K. (2015). dplyr: A grammar of data manipulation [Computer software manual]. (R package version 0.7.1)

Wisco, B. E. (2009). Depressive cognition: Self-reference and depth of processing. *Clinical Psychology Review*, *29*(4), 382–392.

Woody, M. L., & Gibb, B. E. (2015). Integrating NIMH research domain criteria (RDoC) into depression research. *Current Opinion in Psychology*, *4*, 6–12. https://doi.org/10.1016%2Fj.copsyc.2015.01.004

Yoshimura, S., Ueda, K., Suzuki, S.-i., Onoda, K., Okamoto, Y., & Yamawaki, S. (2009). Self-referential processing of negative stimuli within the ventral anterior cingulate gyrus and right amygdala. *Brain and Cognition*, *69*(1), 218–225.

Young, J. E. (1994). *Cognitive therapy for personality disorders: A schema-focused approach* (2nd ed.). Sarasota, FL: Professional Resource Press.

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *11*(4), 493–504. https://doi.org/10.1037/pspa0000056