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

Justin Dainer-Best, Hae Yeon Lee, Jason D. Shumake, David S. Yeager, & Christopher G. Beevers
The University of Texas at Austin

This is a post-print of a manuscript which has been published in *Psychological Assessment*.

Although the Self-Referent Encoding Task (SRET) is commonly used to measure self-referent cognition in depression, many different SRET metrics can be obtained. The current study used best subsets regression with cross-validation and independent test samples to identify the SRET metrics most reliably associated with depression symptoms in three large samples: a college student sample (n = 572), a sample of adults from Amazon Mechanical Turk (n = 293), and an adolescent sample from a school field study (n = 270). Across all three samples, SRET metrics associated most strongly with depression severity included number of words endorsed as self-descriptive and rate of accumulation of information required to decide whether adjectives were self-descriptive (i.e., drift rate). These metrics had strong intra-task and split-half reliability and high test-retest reliability across a 1-week period. Recall of SRET stimuli and traditional reaction time metrics were not robustly associated with depression severity.

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, Hayden, & Klein, 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 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 et al., 2004; Dozois & Dobson, 2001; 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., Bradley & Mathews, 1983; Gotlib 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,

Data are located on the Texas Data Repository, https://doi.org/10.18738/T8/XK5PXX (Dainer-Best, Lee, Shumake, Yeager, & Beevers, 2017) and additional information on Github, https://jdbest.github.io/sretmodels/. An interactive Shiny interface to explore the relationships between these variables can be viewed online at https://jdbest.shinyapps.io/shiny-comparisons/

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. An early study found that depressed participants endorsed more negative words than healthy participants at baseline and after two weeks (Dobson & Shaw, 1987). Researchers also showed consistent endorsements of negative information when tested six months apart, which were only consistent for participants who remained depressed (Dozois & Dobson, 2001). Additionally, researchers who have examined the psychometrics of the SRET within study samples have shown high internal consistency in the responses (Auerbach et al., 2016). However, these studies all had relatively small samples (ns < 100) and internal reliability has not been 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.

One focus was thus to investigate the relative utility of reaction time measures on the SRET, which may have led to ambiguous past results. In fact, 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 = 270). 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 = 13.49, SD = 0.93); the majority was female and white (see Table 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 = 270)
Age, mean (SD)	19.12 (1.14)	37.51 (10.7)	13.49 (0.93)
	n = 336	n = 285	n = 255
Female	212 (62.9%)	165 (57.9%)	151 (57.8%)
	n = 337	n = 285	n = 261
Race	n = 144	n = 285	n = 255
White	83 (58%)	225 (79%)	138 (54%)
Black	8 (6%)	23 (8%)	2 (1%)
Indigenous/ Pacific Islander	3 (2%)	0 (0%)	20 (8%)
Asian	24 (17%)	24 (8%)	52 (20%)
Multiracial/Other	26 (18%)	5 (1%)	—
Hispanic/Latino	—	8 (3%)	43 (17%)
Depression symptoms, mean (SD)	14.19 (9.71)	13.39 (11.85)	2.41 (3.03)
Depression symptoms score above cut-off	188 (33%)	98 (33%)	71 (26%)

Table 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 (with a common cut-off of 16); for adolescents, from the Children's Depression Inventory: Short (CDI:S), which ranges from 0 to 20 (with a common cut-off of 3).

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 3 for the short-form (Allgaier et al., 2012). The CDI:S was used in the adolescent sample, chosen to be time-efficient and to replicate past work which had used the CDI:S (e.g., Miu & Yeager, 2015).

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.[1] 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. Many adolescents did not provide recall data.

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 en-

[1]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".

dorsed as self-descriptive.[2] 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 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 using the Kolmogorov-Smirnov estimation method (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. No structured interview or questionnaire was used to assess psychopathology or past depression; thus, no participants were excluded on this basis.

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 Encoding 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).[3] 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

[2] 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.

[3] 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.

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' 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.[4]

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 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 301 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 timepoints, 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 1/3 of the time were considered not to have been endorsed, whereas those that were endorsed 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 supplemental Figure S1). 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

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

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 $(\hat{y} - y)^2$), 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, and follows guidelines (Hastie, Tibshirani, & Friedman, 2008). 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 et al., 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, 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. 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 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.[5]

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. 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 (β = 0.05, SE = 0.005), drift rate (v) to positive words (β = −0.13, SE = 0.02), and inter-trial variability of starting point (szr; β = −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 (β = 0.03, SE = 0.01), drift rate (v) to negative words (β = 0.13, SE = 0.07), drift rate (v) to positive words (β = −0.11, SE = 0.04), and inter-trial variability of non-decisional components ($st0$; β = 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 (β = −0.24, SE = 0.06), drift rate (v) to negative words (β = 0.43, SE = 0.07), relative starting point (zr) for negative words (β = 1.97, SE = 0.47), and threshold separation (a; β = −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

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

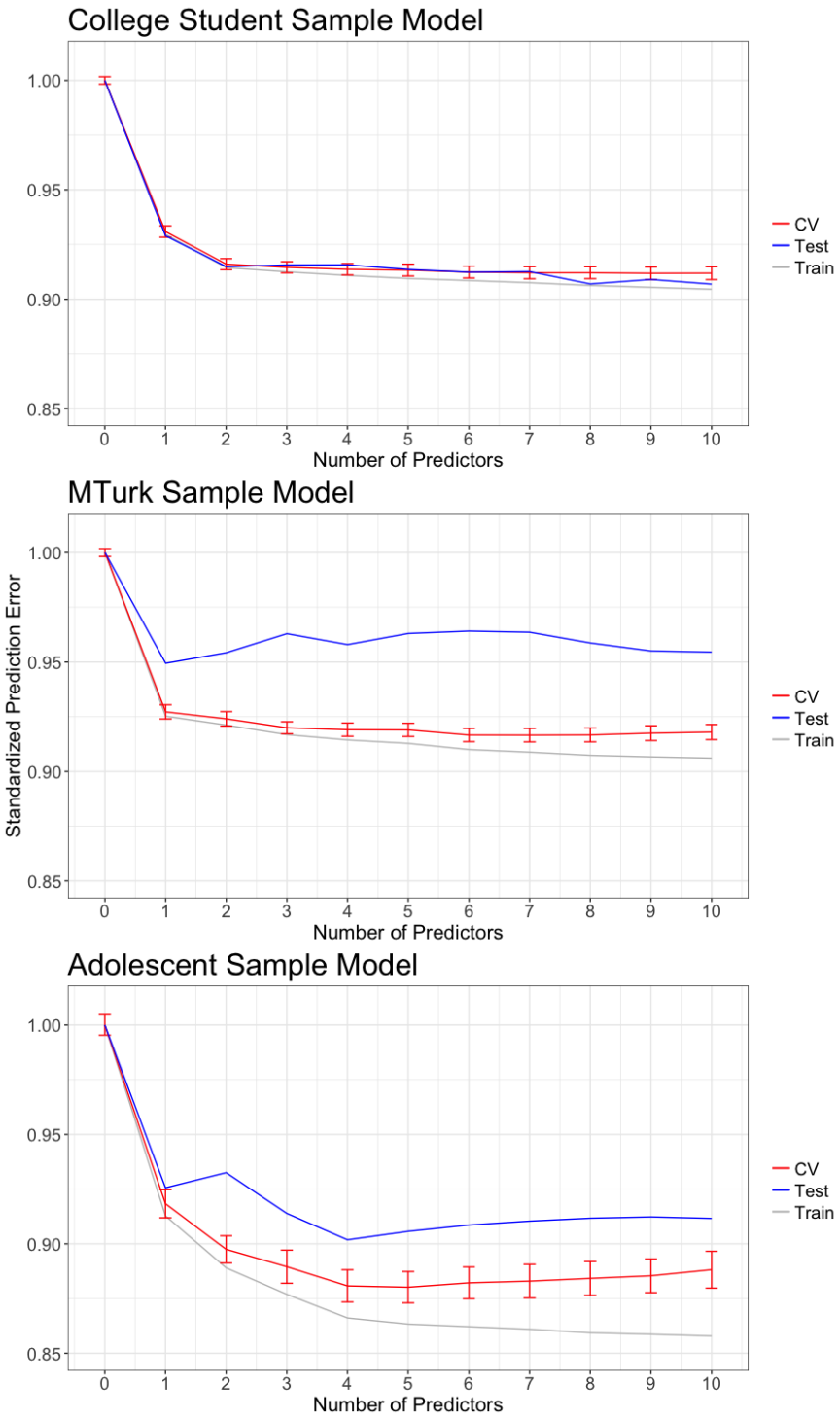

Figure 1. Standardized cross-entropy errors, as a function of increasing the number of predictors for each model. Cross entropy is an index of the discrepancy between model predictions and observed data. Errors are standardized to the null model (i.e., 0 predictors) for each data type, so that the error 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 = 270), 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.

	College Students (n = 572)	MTurk (n = 293)	Adolescent (n = 270)
Positive Endorsements	20.38 (7.09)	20.00 (8.12)	20.02 (5.15)
Negative Endorsements	5.41 (5.36)	5.24 (6.41)	2.62 (4.03)
Positive, Self-Referential Words Recalled	7.39 (3.42)	5.23 (3.15)	5.14 (2.73)
Negative, Self-Referential Words Recalled	2.81 (2.91)	2.06 (2.55)	0.74 (1.44)
Positive Words Recalled	9.35 (3.51)	6.78 (3.47)	6.19 (2.88)
Negative Words Recalled	9.59 (3.60)	7.10 (3.50)	4.75 (3.43)
RT to Positive Words (ms)	829.14 (229.99)	708.33 (172.74)	906 (252.8)
RT to Negative Words (ms)	852.60 (220.56)	719.42 (171.97)	920.5 (231.2)
Drift Rate (v) to Positive Words	1.56 (1.71)	1.79 (2.11)	1.58 (1.36)
Drift Rate (v) to Negative Words	-1.56 (1.39)	-2.06 (1.82)	-2.02 (1.28)
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)	0.43 (0.13)
Threshold Separation (a)	1.33 (0.33)	1.22 (0.34)	1.57 (0.41)
Response Time Constant ($t0$)	0.521 (0.083)	0.508 (0.091)	0.594 (0.112)
Differences in Speed of Response Execution (d)	0.011 (0.034)	0.015 (0.029)	0.014 (0.046)
Inter-Trial Variability of Starting Point (szr)	0.287 (0.118)	0.297 (0.101)	0.270 (0.107)
Inter-Trial Variability of Drift (sv)	0.572 (0.208)	0.595 (0.245)	0.544 (0.214)
Inter-Trial Variability of Non-Decisional Components ($st0$)	0.228 (0.107)	0.183 (0.089)	0.236 (0.115)
Percentage of Contaminants (p)	0.198 (0.214)	0.163 (0.170)	0.179 (0.185)

Table 2

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

three samples; the best model with four predictors for the college students 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 2 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.

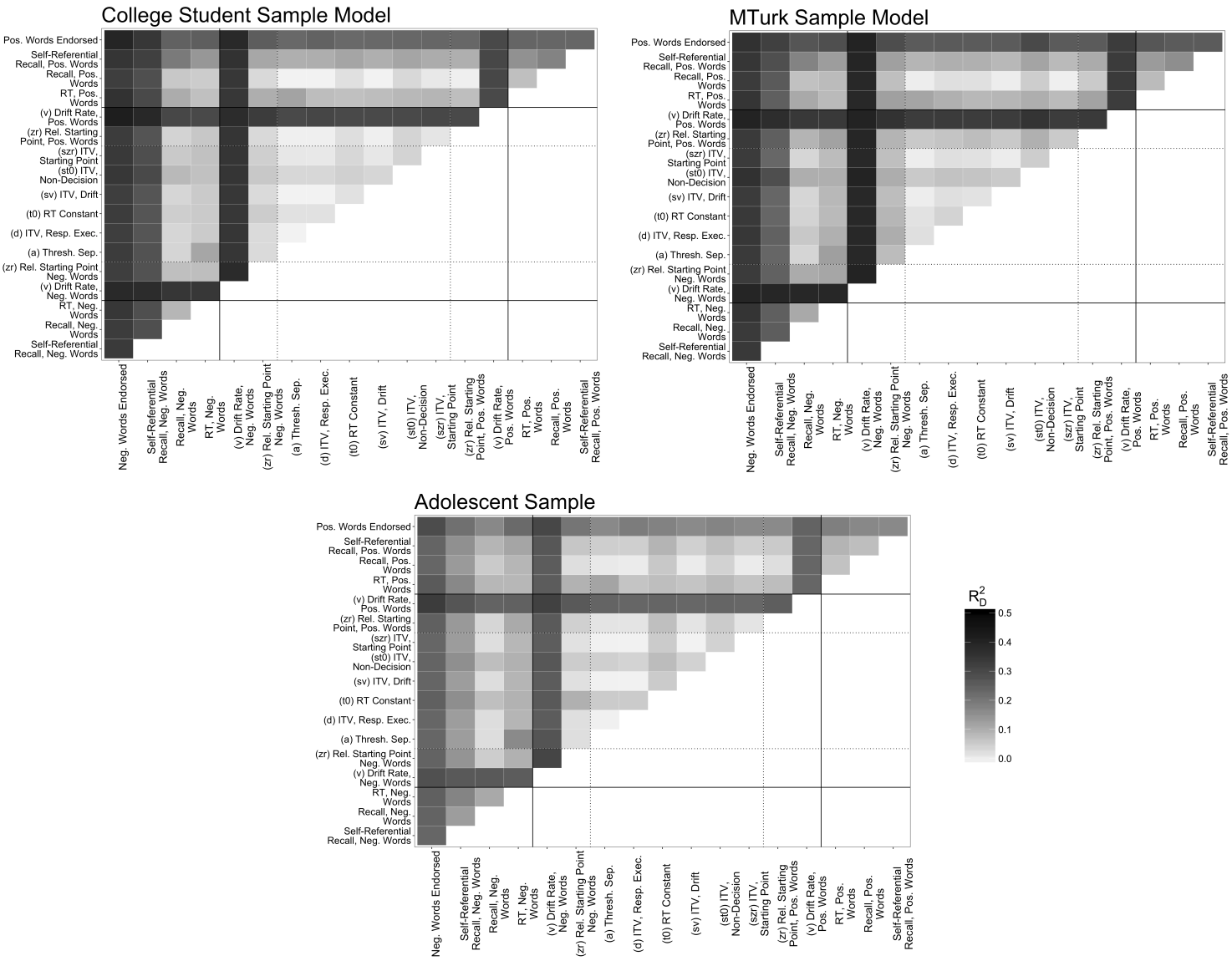

Figure 2. Deviance explained (R^2_D) for each possible two-predictor model predicting depression for each sample. Darker shading indicates a model that explains more of the deviance in depression symptoms, and therefore better fits the data—thus the best (highest R^2_D) two-predictor model for each sample 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.

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 experimen-

tal 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 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 1.) 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 al-

though 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 ≥ .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 2 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 2), 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 [*st*0]; 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 2, 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, *st*0 is correlated with depression in the MTurk sample, and both *st*0 and *a* are strongly correlated with reaction times to both positive and negative words in the MTurk and adolescent samples (see Table S1). (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 as noted in Table 2. 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 self-referent processing in depression, and our samples are similar in makeup to these studies. 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 & Russell, 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 (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. However, we continue to recommend additional numbers of trials—repeating blocks for the college student and MTurk samples provided consistency in endorsement data, strong reliability, and the ability to more effectively make use of the diffusion model.

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 (as recommended by Voss et al., 2013, a minimum of 40 trials for maximum likelihood estimation and of 100 trials for Kolmogorov-Smirnov estimation, although the 52 words presented in the adolescent sample proved sufficient for Kolmogorov-Smirnov estimation in this study), (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. 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. Research was supported in part by the National Institute of Mental Health (grant number R56MH108650) awarded to CGB. Research was also supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (grants R01HD084772 & P2CHD042849) awarded to DSY and the Population Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors declared no conflicts of interest with respect to the authorship or the publication of this article.

References

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. https://doi.org/10.1016/j.jpsychores.2012.08.016

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. https://doi.org/10.1080/026999397379854a

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. https://doi.org/10.1111/psyp.12686

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. https://doi.org/10.3758/a13428-014-0530-7

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beck, A. T. (1967). *Depression: Clinical, experimental, and theoretical aspects*. University of Pennsylvania Press.

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science. *Circulation Research*, *116*(1), 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Bradley, B., & Mathews, A. (1983). Negative self-schemata in clinical depression. *British Journal of Clinical Psychology*, *22*(3), 173–181. https://doi.org/10.1111/j.2044-8260.1983.tb00598.x

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. https://doi.org/10.1177/1745691610393980

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. https://doi.org/10.1016/S0304-4076(96)01818-0

Dainer-Best, J., Lee, H. Y., Shumake, J. D., Yeager, D. S., & Beevers, C. G. (2017). *Replication data and code for: Determining optimal parameters of the Self Referent Encoding Task: A large-scale examination of self-referent cognition and depression, Texas Data Repository Dataverse.* Data set. https://doi.org/10.18738/T8/XK5PXX

Davis, H. (1979). Self-reference and the encoding of personal information in depression. *Cognitive Therapy and Research*, *3*(1), 97–110. https://doi.org/10.1007/BF01172724

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

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*, 1–13. https://doi.org/10.1080/02699931.2016.1146123

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

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. https://doi.org/10.1016/S0191-8869(98)00253-0

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. https://doi.org/10.1037/0022-006X.69.6.914

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. https://doi.org/10.1080/02699931.2014.990358

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. https://doi.org/10.1037/0021-843X.113.3.386

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. https://doi.org/10.1037/0021-843X.98.1.9

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. https://doi.org/10.1016/j.paid.2013.07.016

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. https://doi.org/10.1002/jts.21899

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

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

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. https://doi.org/10.1371/journal.pone.0166600

Kuiper, N. A., & Derry, P. A. (1982). Depressed and nondepressed content self-reference in mild depressives. *Journal of Personality*, *50*(1), 67–80. https://doi.org/10.1111/j.1467-6494.1982.tb00746.x

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. https://doi.org/10.1016/j.jesp.2013.03.013

Litman, L., Robinson, J., & Abberbock, T. (2016). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, 1–10. https://doi.org/10.3758/s13428-016-0727-z

MacDonald, M. R., & Kuiper, N. A. (1985). Efficiency and automaticity of self-schema processing in clinical depressives. *Motivation and Emotion*, *9*(2), 171–184. https://doi.org/10.1007/BF00991574

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6

Miu, A. S., & Yeager, D. S. (2015). Preventing symptoms of depression by teaching adolescents that people can change: Effects of a brief incremental theory of personality intervention at 9-month follow-up. *Clinical Psychological Science*, *3*(5), 726–743. https://doi.org/10.1177/2167702614548317

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. https://doi.org/10.1038/s41562-016-0021

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

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. https://doi.org/10.1007/BF01175139

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. https://doi.org/10.1177/014662167700100306

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Revelle, W. (2016). psych: Procedures for psychological, psychometric, and personality research [Computer software manual]. Evanston, Illinois. Retrieved from `http://CRAN.R-project.org/package=psych` (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. https://doi.org/10.1037/abn0000184

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220. https://doi.org/10.1177/2167702612469015

Shumake, J. (2016). beset: Tools for data exploration and mining [Computer software manual]. (R package version 0.0.0.9000)

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. https://doi.org/10.1080/02699938808410927

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

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

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology. *Experimental Psychology*, *60*, 385–402. https://doi.org/10.1027/1618-3169/a000218

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. https://doi.org/10.1037/1089-2680.7.2.203

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. https://doi.org/10.1016/j.jmp.2010.01.004

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*. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from `http://ggplot2.org`

Wickham, H. (2015). tidyr: Easily tidy data with 'spread()' and 'gather()' functions [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=tidyr` (R package version 0.3.1)

Wickham, H., & Francois, R. (2015). dplyr: A grammar of data manipulation [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=dplyr` (R package version 0.4.3)

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