The world dangerous it is—The scrambled sentences task in the context of posttraumatic stress symptoms

Felix Würtz
Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, Germany

Simon E Blackwell
Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, Germany

Jürgen Margraf
Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, Germany

Marcella L Woud
Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, Germany

Abstract
The Scrambled Sentences Task (SST) is a robust measure of interpretational processes in psychopathology. However, there is little evidence of its utility in measuring dysfunctional appraisals (DAs) of potentially traumatic events. We developed a novel SST for trauma-related DAs and examined its psychometric properties including convergent validity (correlations with PTSD-related symptoms and self-reported DAs), divergent validity (e.g., symptoms of depression and eating disorders), and retest reliability via an online study. Our sample (T1: N = 214, T2: N = 145) included participants who reported a potentially traumatic life event still eliciting distress. We found high correlations between the SST, PTSD-related symptoms (r = .37-.51), and self-report measures of DAs (r = .41-.58), indicating good convergent validity. Internal consistency (split-half = .78-.90) and retest reliability (ICC(3,1) = .73-.81) were also good. However, moderate to large correlations with symptoms of other disorders (r = .17-.58) indicated limited divergent validity. Finally, the SST explained unique variance in PTSD-related symptoms above self-report measures of DAs. The results demonstrate the promise of the SST as a valid and reliable tool to assess DAs in the context of potentially traumatic life events. Further research should investigate the transdiagnostic role of trauma-related DAs in psychopathology and the relationship between the SST and self-report measures of DAs.

Keywords
dysfunctional appraisals, interpretation bias, trauma, posttraumatic stress disorder, Scrambled Sentences Task

Date received: 1 February 2022; accepted: 22 August 2022

Dysfunctional appraisals (DAs) are a central element in cognitive models of the development and maintenance of posttraumatic stress disorder (PTSD) (Dalgleish, 2004; Ehlers & Clark, 2000; Foa et al., 1989). For example, the model of Ehlers and Clark (2000) suggests a causal pathway

Corresponding author:
Felix Würtz, Mental Health Research and Treatment Center, Faculty of Psychology, Ruhr-Universität Bochum, Massenbergrasse 9-13, 44787 Bochum, Germany.
Email: felix.wuertz@rub.de
from DAs of the trauma and its sequelae to other aspects of PTSD, such as a current feeling of threat and reoccurring, intrusive memories of the traumatic event. The key role of DAs in PTSD has been further underscored by the inclusion of negative cognitive changes following a trauma as a criterion in the diagnosis of PTSD in the Diagnostic and Statistical Manual of Mental Disorders – Fifth edition (DSM-5) (American Psychiatric Association, 2013).

Dysfunctional appraisals can be understood as negative evaluations of an event or stimulus. In the context of PTSD, dysfunctional appraisals reflect the patients’ idiosyncratic appraisals/beliefs about their reactions to the event and their trauma symptoms. To illustrate, patients may appraise their reactions during the traumatic event as a sign of personal weakness or consider their intrusive memories as a proof that they are losing control (e.g., Ehlers & Clark, 2000). Accordingly, DAs in the context of PTSD reflect the evaluation of the meaning of a trauma-relevant stimulus and may involve extrapolations to broader, more wide-ranging themes, with implications for oneself, the world, or one’s future. As such, DAs are distinct from other cognitive phenomena in the context of emotional psychopathology, such as interpretation biases. Although interpretation biases are also the product of an evaluative process, they focus more specifically on the evaluation of ambiguous stimuli or situations (cf. Mathews & MacLeod, 2005) and are restricted to narrower themes.

Evidence for a central role of DAs in PTSD comes from several strands of research. To illustrate, studies have found that DAs are associated with symptoms of PTSD, both cross-sectionally (Gómez de La Cuesta et al., 2019) and longitudinally (Bryant & Guthrie, 2005; Ehring et al., 2008; Kleim et al., 2007). Further, the reduction of DAs during treatment has been shown to predict subsequent symptom reduction, but not vice versa (Brown et al., 2019; Kleim et al., 2013). Additional evidence comes from studies with an experimental focus, for example, manipulating DAs in order to test their potential causal effect on analog or actual PTSD-related symptoms. For example, Cheung and Bryant (2017) induced DAs about intrusions via informing participants that intrusive memories of distressing films were indicative of poor psychological functioning. This induction in turn led to more frequent intrusive memories of the films than a non-dysfunctional induction (Schartau et al., 2009). Studies using computerized paradigms developed in the context of Cognitive Bias Modification-Appraisal (CBM-App; MacLeod et al., 2009; McNally & Woud, 2019; Woud & Becker, 2014), however, have revealed more mixed findings. One study (Woud et al., 2012) found fewer intrusions of an analog trauma following training aiming to induce a functional (positive) appraisal style compared to training aiming to induce a dysfunctional (negative) appraisal style (positive vs. negative CBM-App). Other studies found an effect only on intrusion distress, with less distress in the positively trained group compared to the negatively trained group (Woud, Cwik et al., 2018; Woud et al., 2013) or no effects at all on analog trauma symptoms (Vermeulen et al., 2019; Würzt et al., 2021). In a recent randomized controlled trial (RCT) amongst inpatients diagnosed with PTSD, the effects of positive CBM-App as an adjunct to treatment as usual were compared to a sham-control training. Positive CBM-App, compared to the control condition, successfully reduced DAs and led to a reduction in PTSD symptoms (Woud et al., 2021). However, in another recent RCT, CBM-App did not reduce DAs in comparison to a neutral control condition, and there was no between-group difference in PTSD symptoms after CBM-App (de Kleine et al., 2019). An important note here is, however, that the latter RCT delivered the CBM-App training online, to outpatients with PTSD while they waited for therapy, and this could partly explain the deviating results.

In summary, there is substantial evidence that DAs are a core feature of PTSD, and this evidence is consistent with the proposal that DAs can causally affect trauma symptoms, supporting key assumptions of cognitive models of PTSD. However, past research has mainly used self-report questionnaires such as the Posttraumatic Cognitions Inventory (PTCI) (Foa et al., 1999) to assess DAs (for exceptions, see Blackwell et al., 2021; Boffia et al., 2018; Lindgren et al., 2013), and it is unclear to what extent this may or may not be a problem. For example, self-report measures are associated with several disadvantages such as response or demand effects (MacLeod et al., 2009). Further, actively evaluating the questionnaires’ statements might conceal aspects of DAs that are difficult to verbalize (Brewin et al., 1996), or that are prone to be actively suppressed (Ehlers & Clark, 2000; Rosebrock et al., 2019). Finally, DAs may involve automatically activated counterparts, as assumed by cognitive models (e.g., Ehlers & Clark, 2000), which self-report questionnaires may not be able to capture. The findings of Lindgren et al. (2013), for example, provide some evidence for the idea that self-report measures might not fully capture DAs relevant to PTSD. In their study, an indirect measure of DAs based on the reaction time of participants to categorize self-related words as either traumatized or non-traumatized (an implicit association test) explained additional variance in PTSD-related symptoms beyond that explained by a self-report measure of DAs (the PTCI).

Importantly, other research areas in psychopathology are also characterized by such limitations, and this has led to the development of various novel assessment tasks. Such tasks can include, for example, measuring responses to ambiguous stimuli (Hirsch et al., 2016; Schoth & Liossi, 2017). As different kinds of tasks may tap into slightly different aspects of a process under investigation, examining new ways of measuring a phenomenon such as DAs can help in delineating the conditions under which the DA is or is not expressed. Further, having access to a broad range of measures can also provide a form of triangulation, increasing our confidence that results found are robust rather than an artifact.
of one measurement method (Munafo & Davey Smith, 2018). One example of a measure with potential utility in the context of DAs and PTSD is the Scrambled Sentences Task (SST) (Wenzlaff & Bates, 1998), which was initially developed to assess interpretational processes in the context of depression. During the SST, participants are instructed to sort words into grammatically correct sentences (e.g., bright the future very dismal) and each set of words is designed such that it can be formed into either a disorder-congruent (The future looks very dismal) or -incongruent (The future looks very bright) sentence. To reduce the impact of active evaluation of the chosen words, the SST is applied with a short time limit and participants are simultaneously engaged in a cognitively demanding task, usually retaining a 6-digit number in mind. The SST is most commonly scored by calculating the percentage or proportion of correctly formed sentences that were unscrambled in a disorder-congruent manner. Performance on the SST has been shown to be associated with symptoms of depression (Lee et al., 2016; O’Connor et al., 2021) and to predict a subsequent diagnosis of major depression (Rude et al., 2010). Recent research has successfully adapted the stimuli used in the SST to other psychopathology contexts such as anorexia nervosa or body dissatisfaction (Bradtach et al., 2020; Brockmeyer et al., 2018), paranoia (Savulich et al., 2020), anxiety sensitivity (Zahler et al., 2020), symptoms of social anxiety (de Voogd et al., 2017), female sexual dysfunction (Zahler et al., 2021), or generalized anxiety disorder (Krahé et al., 2022). Across all these studies, the SST score was associated with symptoms of the target disorder, supporting the validity of the SST as a measure of dysfunctional interpretational processes in various disorders. Further, a recent meta-analysis has shown that the average correlation between the SST score and symptoms of the targeted disorder is moderate ($r = .46 [ .41, .51]$; Würtz et al., 2022). In addition, there is evidence that the SST is a sensitive measure for the assessment of change in interpretational processes in experimental research using CBM-paradigms in both clinical (Yiend et al., 2014) and analog samples (Bradtach et al., 2020; Würtz et al., 2022). Finally, the recent study by O’Connor et al., (2021) showed that the SST was strongly associated with symptoms of depression, even after taking an explicit measure of dysfunctional interpretational processing into account. However, reaction times derived from the Word-Sentence-Association Paradigm, which is another indirect measure and uses reaction times to infer interpretational processing, failed to predict depressive symptoms.

Moving in a more trauma-related direction, the SST has also been successfully adapted to assess DAs in the context of negative life events (Viviani, Mahler et al., 2018). In the study by Viviani et al., the SST score was associated with questionnaire measures of DAs related to the event and symptoms of depression. However, that SST was specifically designed to be feasible in an fMRI setting and thus differed from the usual version (Viviani, Dommes et al., 2018). Specifically, participants unscrambled the sentences in their mind to avoid artifacts of hand movement, and no cognitive load was applied since the alternative task instruction already presented a cognitive challenge. Further, the SST stimuli neither referred to a potentially traumatic life event nor targeted directly related cognitions, enabling its application in a sample that was not necessarily exposed to such an event. As such, findings from this study might not be fully generalizable to the SST in its common computerized form or the context of an actual potentially traumatic life event.

To summarize, the SST is a potentially valuable tool to assess DAs in different areas of psychopathology. However, there has only been limited research into its applicability to the kind of DAs relevant to PTSD, that is, DAs in relation to potentially traumatic life events. Given its robust findings in a broad range of disorders, the SST might offer a promising approach also for the assessment of DAs in this context. The current study therefore aimed to develop a novel, computerized version of the SST to assess DAs in relation to a potentially traumatic life event, examining the SST’s psychometric properties including its 2-week test–retest reliability. A second aim was to test whether the SST would be a useful predictor of PTSD-related symptoms over and above self-report. Finally, considering the necessity for measures in experimental research to assess changes in cognitive processing pre-post a manipulation, we aimed to develop two parallel SST versions. These were presented in an online study to a sample of participants who reported a potentially traumatic life event that was still causing distress and therefore potentially the object of negative appraisals. Further, we included different questionnaires that were either directly related to PTSD, as measures of convergent validity, or (relatively) unrelated, as measures of divergent validity. To allow an estimation of test–retest reliability, participants were invited to a second assessment session in which we applied the SST again, together with measures of convergent validity.

Our hypotheses were as follows: We expected the SST score to be strongly correlated with both PTSD-related symptoms and DAs measured via self-report (convergent validity). Correlations with measures not directly related to PTSD symptoms, however, were expected to be relatively weak (divergent validity). Further, we expected that the SST would explain unique variance in symptoms, that is, would be a significant predictor of PTSD-related symptoms after controlling for DAs as measured via self-report and other relevant variables (unique variance).1

**Methods**

**Recruitment**

Prior to the start of participant recruitment, the study was pre-registered by uploading the study protocol to the Open Science Framework (OSF: https://osf.io/bnzp3/). Participants were recruited by distributing the study’s link via social media and the participant noticeboard of Ruhr-University Bochum. Inclusion criteria were age ≥ 18 and
prior exposure to a potentially traumatic life event still causing distress (and therefore being potentially subject to negative appraisals). Participants were included when they rated their current distress caused by the event as ≥ 3 on a 10-point scale ranging from 1 (not distressing at all) to 10 (highly distressing) (for a similar approach, see Schartau et al. 2009; Woud et al. 2019). Further, participants could only take part using either a laptop or desktop computer to avoid potential artifacts from the touchscreens of mobile devices. In the study advertisements, participants were informed that high fluency in German (e.g., being a native speaker) was necessary to take part in the study; however, no participants were excluded based on this criterion. Participants received course credit for participation and could take part in a prize draw for a 10€ voucher.

We aimed to recruit a total sample of \( N = 170 \) for T1, expecting a dropout of ~ 15% which would have provided a sample of \( N = 145 \) for T2. This sample size was informed by an a priori power analysis (Erdfelder et al., 2009) aiming to achieve 80% power to find a small to medium effect size \( (f^2 = .055) \) of a single predictor in a multiple regression model with ten predictors in total at an alpha level of \( \alpha = .05 \). However, during recruitment, it became apparent that the dropout was larger than expected and that some participants would have to be excluded due to technical problems. As a consequence, we recruited a larger sample for T1 than initially planned. Specifically, we did not use the sample size at T1 as a stopping criterion but stopped the recruitment at a time-point when \( N = 145 \) participants had provided complete data of both T1 and T2. However, after the recruitment was stopped, participants who had already taken part at T1 could still complete T2.

**Material**

**Scrambled Sentences Task.** The SST (Wenzlaff & Bates, 1998) was implemented as a web application programmed using JavaScript on a designated university-based website (Blackwell et al., 2021), and participants completed the SST in their preferred web browser. During the SST, participants were presented with sets of six words and were instructed to sort these into grammatically correct sentences by selecting five of them. Each trial started with a fixation-cross that was presented for 1s in the middle of the screen. After the fixation-cross disappeared, the word set was presented, and participants could sort the words by clicking on them. After clicking on a word, a number appeared above it, indicating the position of that word in the sentence formed. Participants could not correct their responses. For each trial, there was a time limit of 12s and a trial ended automatically either after the time limit or after five words had been clicked on. As in previous studies, the SST was presented with a cognitive load to reduce participants’ deliberate evaluation of the presented stimuli: At the start of the SST, a 6-digit number was presented for 7s, and participants were instructed to keep the number in mind during the task since they would be asked to recall it afterward. Stimuli for the two SST versions were created based on the German version of the PTCI (Ehlers, 1999; Foa et al., 1999). Each SST version consisted of 20 stimuli, of which 12 were based on the “self,” six on the “world,” and two on the “self-blame” subscale of the PTCI. After the initial stimuli development, the sets were independently discussed with two clinicians outside of the study team who were experienced in the treatment of PTSD. Stimuli were then further adapted to resemble typical trauma-related appraisals of PTSD patients. Word sets were designed such that two words were target words of which one had to be omitted to form a grammatically correct sentence. Depending on which word was omitted, the sentence formed represented either a trauma-related functional or dysfunctional appraisal (example for world subscale: “dangerous the safe is very world”—functional appraisal: the world is very safe, or a dysfunctional appraisal: the world is very dangerous; target words in italics). The order of the two SST versions was counterbalanced (A first or B first) based on the participant number assigned after providing consent, and the order was the same for T1 and T2 (participants receiving version A first at T1 also received version A first at T2 and vice versa). After each SST version, participants were asked to type in the memorized number, received feedback on whether it was correct, and were invited to take a break. Prior to the first version, five neutral practice trials were presented to make participants familiar with the procedure. During the practice trials, participants were also instructed to keep a 6-digit number in mind which had to be recalled after the practice sentences.

The SST score was calculated by dividing the number of correctly generated, dysfunctional sentences by the total number of correctly formed sentences, resulting in a negativity score between 0 and 1, with higher values indicating a more dysfunctional appraisal style. Grammatically incorrect sentences, sentences formed as questions, and sentences with fewer than five words were counted as errors. A list of correct solutions was uploaded together with the pre-registration. However, some additional correct solutions were added after manual inspection of collected data (all materials including the stimuli, the correct solutions, and the source code of the web application can be found on the OSF (https://osf.io/bnzp3/)).

**Convergent validity**

**Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5).** PTSD-related symptoms were assessed using the German translation of the PCL-5 (Krüger-Gottschalk et al., 2017; Weathers, Litz et al., 2013). The PCL-5 consists of 20 items with each item representing a PTSD symptom as specified in the DSM-5. Participants were instructed to rate how much distress was caused by each symptom in the past 2 weeks on a 5-point scale ranging from 0 (not at all) to 4...
A total score of ≥ 33 has been suggested as a cut-off for a potential PTSD diagnosis (Krüger-Gottschalk et al., 2017). Internal consistency (Cronbach’s α) was .94 [.93, .95] at T1 and .95 [.94, .97] at T2, and the test–retest reliability was ICC (3, 1) = .79 [.73, .84].

Posttraumatic Cognitions Inventory (PTCI). DAs related to PTSD were assessed via the German version of the PTCI (Ehlers, 1999; Foa et al., 1999). The PTCI consists of 33 statements representing typical PTSD-related appraisals. Participants were instructed to rate their agreement with each statement on a 7-point scale ranging from 1 (strongly disagree) to 7 (strongly agree) with higher values indicating more dysfunctional DAs. Internal consistency at T1 was .94 [.93, .96] and at T2 .96 [.95, .97]. Test–retest reliability was ICC (3, 1) = .83 [.78, .87].

Divergent validity

Quick Inventory of Depressive Symptomatology (QIDS). Symptoms of depression were assessed via the QIDS (Roniger et al., 2015; Rush et al., 2003). This questionnaire consists of 16 items grouped in nine depression-related domains. Each domain consists of one item except for the domains sleep, appetite, and psychomotor agitation/restlessness, consisting of two to four items. Participants rated how strongly they were affected by each symptom in the past week using a 4-point scale ranging from 0 to 3 with higher scores indicating more severe symptoms. The questionnaire is scored by calculating the sum score of each disorder domain, using the highest score for domains with more than one item (Roniger et al., 2015). Total scores of ≤ 5 are indicative of no depression, 6–10 of mild, 11–15 of moderate, 16–20 of severe, and ≥ 21 of very severe depression, respectively (Rush et al., 2003). Internal consistency for the full scale was .83 [.80, .86].

Ambiguous Scenarios Test for Depression (AST). Depression-related interpretation biases were assessed using version A of the German AST (Berna et al., 2011; Rohrbacher & Reinecke, 2014). The AST consists of 15 ambiguous scenarios describing everyday life situations, for example, having an oral examination. Participants were instructed to briefly imagine each scenario and to rate its pleasantness on an 11-point Likert scale ranging from −5 (very unpleasant) to 5 (very pleasant), with higher scores indicating a more positive interpretation bias. The AST’s internal consistency was .81 [.77, .84].

Acrophobia Questionnaire – Anxiety Subscale (AQ). Fear of heights was assessed via the German translation of the Anxiety subscale of the AQ (Cohen, 1977). The Anxiety subscale consists of 20 height-related items. Participants were instructed to rate how anxious they would feel in each situation using a 7-point Likert-scale ranging from 0 (not anxious at all) to 6 (very anxious). The AQ’s internal consistency was .93 [.91, .94].

Fear of Spiders Screening (FSS). Fear of spiders was assessed via the FSS (Rinck et al., 2002), assessing both fear and avoidance of spiders. The questionnaire consists of 4 statements concerning fear of spiders, associated body symptoms, avoidance, and distress, respectively. Each item is rated on a 7-point Likert-scale ranging from 0 (does not apply at all) to 6 (fully applies). Internal consistency was .91 [.89, .93].

Eating Disorder Examination Questionnaire (EDEQ). Symptoms of eating disorders in the past 4 weeks were assessed via the German translation of the EDEQ (Fairburn & Beglin, 1994; Hilbert et al., 2007). The EDEQ is a 28-item scale, consisting of four subscales: Restrain (5 items), Eating concerns (5 items), Weight concerns (5 items), and Shape concerns (8 items), and five additional items that are not used in scoring. Participants were instructed to answer each item on a 7-point scale ranging from 0 to 6 with higher scores indicating more severe eating disorder symptoms. The EDEQ is scored by calculating the mean score of each subscale and then calculating the overall mean of all subscales. Internal consistency for the full scale was .96 [.95, .96].

Additional measures

Demographic data. Participants were first asked to indicate their age and their gender (male, female, or non-binary), and whether they were a native German speaker. Further they were asked to indicate their self-rated language fluency (“How secure are you in reading German texts?”) on a 5-point scale ranging from 0 (very insecure) to 4 (very secure). Finally, participants were asked for their highest level of formal education.

Life Event Checklist for DSM-5 (LEC-5) and distress ratings. Exposure to potentially traumatic negative life events was assessed via the German translation of the LEC-5 (Weathers, Blake et al., 2013). The LEC-5 is a checklist assessing the exposure to 17 traumatic events following the criteria of DSM-5. For each traumatic event, participants were instructed to indicate whether (i) it happened to them, (ii) they witnessed it, (iii) learned about that it happened to someone close, (iv) it happened in an occupational context, (v) they are not sure whether it happened to them, or (vi) they were not exposed to it. For each item, multiple answers were possible, and we scored the LEC by counting the number of events to which participants select another answer than “not exposed.” Scores on the LEC were calculated to help better characterize the sample in terms of
trauma exposure, but not used in analyses. To assess further information about the distressing event, we supplemented the LEC with additional questions. That is, participants were asked to indicate which of the selected events caused the most distress, and when the event happened. Further, they had to rate how much distress the event caused when the event happened, and how distressing the event is now, using a 10-point scale ranging from 1 (not distressing at all) to 10 (highly distressing).

**Procedure**

Prior to participation, all participants were informed that they would be asked about a traumatic event during the study and recommended not to take part if they would anticipate experiencing too much distress from questions related to the event. Participants then gave informed consent by completing an online form. Next, they were asked whether they were using a built-in laptop mouse or external mouse to take part in the study. This was followed by the LEC and the additional questions related to the potentially traumatic event. Participants who rated their current distress < 3 were excluded from further participation (Schartau et al., 2009; Woud et al., 2019). Subsequently, participants completed the SST, after which they were asked for their email address in order to be able to contact them for the second assessment. Next, participants were directed to another survey platform where the questionnaire measures were administered, in the following order: Demographic data, PTCI, PCL-5, QIDS, AST, EDEQ, AQ, and FSS. Two questionnaires assessing resilience and emotion regulation were included at the end of T1 for potential exploratory analyses; however, these are not further analyzed in the present paper (see Supplementary Material for details). Two weeks later, participants received an email containing a link to the second assessment. During this second assessment, participants were asked again to indicate what type of mouse they were using, followed by assessing the current distress caused by the potentially traumatic negative event and the application of the SST. Next, the PTCI and the PCL-5 were completed, again using a separate survey platform. Finally, participants were debriefed, received their course credit, and could enter their information for the prize draw. At the end of each assessment time point, participants were provided with information on an online platform with contact details for (trauma-specialized) psychotherapists. All study procedures were approved by the local ethics committee of the Psychology department at Ruhr-University Bochum (approval number: 596).

**Statistical approach**

For convergent and divergent validity, bivariate correlations were calculated. However, upon visual inspection, both the SST and questionnaire data were found to be positively skewed and not normally distributed. Therefore, we calculated Spearman’s rank correlations with bootstrapped 95% confidence intervals (CIs; 10,000 resamples) instead of the Pearson correlations planned in the pre-registration, in order to obtain a more robust estimation of the examined associations (de Winter et al., 2016). As the investigation of our hypotheses concerning con- and divergent validity depended on multiple significance tests, the significance level of $\alpha = .05$ was adjusted via the Bonferroni correction and thus was $\alpha = .008$ for convergent validity (.05/number of versions (full SST, v. A, v. B) = $3 \times$ number of questionnaires (PTCI, PCL) = 2) and $\alpha = .003$ for divergent validity (.05/number of versions = $3 \times$ number of questionnaires (QIDS, AST, EDEQ, FSS, AQ) = 5). Whether the SST would explain unique variance in current PTSD-related symptoms above self-report was examined via a series of multiple regression models with the PCL total score of the relevant time point as the outcome. In a first regression step, we controlled for basic characteristics of the participants and the event: age, gender (female vs. non-female), distress caused by the negative life event when it occurred, and time passed since the event (in months). In a second step, we then added scores on the SST to examine how much variance in PTSD symptoms it explained above these basic variables. In a third step, we added scores on the PTCI, allowing us to examine whether SST scores explained additional variance in PTSD symptoms above that explained by self-report DAs. Analyses were run separately for both assessment time points. For all regression models, full models were examined for influential outliers (standardized residuals > 3 and Cook’s distance > 1) (Field et al., 2012) and multicollinearity (variance inflation factors; VIFs). Bootstrapped 95% CIs (10,000 resamples) for regression coefficients were calculated as robustness checks.

In exploratory analyses, we investigated Spearman’s rank correlations between the PCL-5 and the PTCI and measures of con- and divergent validity at T1 to investigate the association between the symptoms of different disorder domains and compare the results for the SST to those for the PTCI. Further, we analyzed whether participants who dropped out for T2 differed from participants who did not drop out on their age, self-reported language fluency, number of correct solutions on the SST or their bias score, depressive symptoms, PTSD symptoms, or the PTCI using two-sample t-tests or Welch’s tests where appropriate. Finally, we investigated rehearsal effects and differences between the two SST versions. Specifically, we investigated whether participants were more likely to make errors or to form sentences in a dysfunctional way in the first compared to the second application at each time point, or on one or other of the versions. This was done by fitting mixed model logistic regression with fixed effects of “application” (first and second) and “version” (version A and version B) with
random intercepts for “participant” and “stimulus.” Outcomes were “incorrect = 1 versus correct = 0” and “dysfunctional = 1 versus functional = 0,” respectively. Models were fitted separately for T1 and T2 data. We further investigated whether participants were more likely to make errors on the SST or to form sentences in a dysfunctional way, based on how many times the SST had already been applied. This was investigated by fitting mixed model logistic regression with the fixed effect “time” (first, second, third, or fourth application) and random intercepts for “participant” and “stimulus” using maximum likelihood estimation with the outcomes “incorrect = 1 versus correct = 0” and “negative = 1 versus positive = 0,” respectively. Tukey’s pairwise post-hoc comparisons were applied. For additional analyses (such as predictive validity analyses and analyses concerning rehearsal effects), the associated statistical approach, and the results, see Supplementary Material.

All analyses were run in R version 4.0.0 (R Core Team, 2021) using RStudio version 1.4.1106 (RStudio Team, 2022). Spearman’s rank correlations and CIs were calculated using the R-package “psych” (Revelle, 2019). All reliability indices were calculated using the package “psych” (Revelle, 2019).

Results

Sample

In total, N = 318 participants took part in T1. Out of these, 83.33% (N = 265) rated the current distress caused by the negative life event to be ≥ 3 and were therefore eligible to proceed with the study. Of those, 31 participants then were excluded at T1 as they did not complete any of the questionnaires. Further, two participants were excluded as they dropped out after the demographic questionnaire. Additionally, at T1, a further 14 participants were excluded due to technical difficulties with the saving of the SST data. Finally, one participant (0.7%) was excluded as they did not form a grammatically correct sentence on either one of the SST versions, resulting in a final sample of N = 145 for T2. In total, 3 participants (partially) took part twice in the study. Their data of the first participation was included in all analyses, yet data for the second participation was excluded.

On average, participants at T1 reported mild symptoms of depression on the QIDS, M = 6.59, SD = 4.62, and a subthreshold level of PTSD-related symptoms on the PCL-5, M = 18.66, SD = 15.49. The most common index event was “life-threatening illness” (19.16%, n = 41), followed by “other” (16.36%, n = 35), “severe human suffering” (11.21%, n = 24), “unwanted sexual experiences” (9.81%, n = 21), “physical or sexual assault” (7.94%, n = 17, respectively), and “sudden accidental death” or “transportation accidents” (7.01%, n = 15, respectively) with each other category being endorsed by fewer than 10 participants. Further, the index event was most commonly personally experienced (39.25%, n = 84), followed by witnessed (20.56%, n = 44), learned that it happened (12.15%, n = 26), happened in an occupational context (2.34%, n = 5), or participants were not sure whether the event happened to them (1.87%, n = 4). The remaining participants (23.36%, n = 50) endorsed multiple exposure forms for the index event, such as that they personally experienced the event but also witnessed it. For sample characteristics including questionnaire data, see Table 1.

Psychometric properties of the Scrambled Sentences Task

Internal consistency at T1 was split-half [95% CI] = .88 [.84, .91] for the full SST, .78 [.71, .84] for version A, and .80 [.73, .85] for version B. Internal consistency at T2 was .90 [.86, .93] for the full SST, .80 [.73, .85] for version A, and .82 [.76, .87] for version B. Test–retest reliability (mean re-test interval: 16.39 days12) was ICC (3, 1) [95% CI] = .81 [.75, .85] for the full SST, .75 [.68, .81] for version A, and .73 [.65, .79] for version B. Parallel versions reliability was .71 [.65, .76] for T1 and .84 [.79, .87] for T2. Reliability between versions across time points was the following: For version A at T1 and B at T2, ICC (3, 1) [95% CI] = .75 [.68, .80] and for version B at T1 and A at T2, .67 [.59, .74]. The cognitive load retention rates at T1 were 77.57% after the practice trials, 50.47% after the first application, and 61.68% after the second application, and for T2, 78.62% after the practice trials, 56.55% after the first application, and 57.24% after the second application. Finally, the error rates and thus the rate of excluded trials were the following at T1: 19.15% for the full SST, 16.75% for version A, and 21.54% for version B. At T2, the error rates were 14.40% for the full SST, 13.28% for version A, and
15.52% for version B. For further details on error rates and analyses on stimulus level, see the section Exploratory Analyses and the Supplementary Material.

Convergent validity (T1 and T2) and divergent validity (T1)

As expected, for convergent validity, we found that the SST total score was strongly and positively correlated with both the PCL-5, \( r_s = .37 - .51 \) and the PTCI, \( r_s = .41 - .58 \), indicating that generating a higher proportion of dysfunctional sentences was associated with both more severe PTSD-related symptoms and a more maladaptive, trauma-related appraisal style. This pattern was found for both assessment timepoints, that is, T1 and T2. Unexpectedly, for divergent validity, we found the SST score to be strongly and positively correlated with the QIDS, \( r_s = .40 - .48 \), and strongly and negatively correlated with the AST, \( r_s = -.58 \) to \(-.53 \), showing that higher, dysfunctional scores on the SST were associated with more severe depressive symptoms and a more maladaptive, depression-related interpretation bias. For other measures of divergent validity, namely, the EDEQ, \( r_s = .26 - .30 \), the FSS, \( r_s = .17 - .24 \), and the AQ,
The PTCI was found to be strongly associated with the PCL. For detailed correlations, see Table 2.

Unique variance explained at T1 and T2
For T1, we found that in step 1, there were no significant predictors of concurrent PTSD-related symptoms. As expected, adding the SST total score in step 2 significantly increased the explained variance. Next, in line with our expectations, the SST remained a significant predictor after adding the PTCI total score. Results for T2 mirrored those for T1 as such that the SST was a significant predictor of PTSD-related symptoms and remained significant after including the PTCI total. There were no influential outliers and an investigation of VIFs indicated that the influence of multicollinearity on the model interpretability was negligible (VIFs < 1.85). For details, see Table 3.

Exploratory analyses—Posttraumatic Cognitions Inventory correlations
The PCL was found to be strongly associated with the PTCI and the QIDS and moderately correlated with the AST, the EDEQ, and the AQ. A weak correlation was found for the FSS. The same pattern of correlations was found for the PTCI. For details, see Table 4. In summary, the pattern of correlations between the PCL, the PTCI, and other questionnaire measures followed the same pattern as for the SST.

Exploratory analyses—Analyses of dropout
Further exploratory analyses comparing participants who dropped out for T2 with participants who did not revealed that they did not differ in age, self-reported language fluency, their SST bias score, or their severity of depressive symptoms. However, participants who dropped out of the study had overall lower rates of correct solutions at T1, slightly higher PTSD symptom levels, and showed higher PTCI scores, compared to those who completed both T1 and T2. For descriptive statistics of analyses as well as results of the significance tests concerning dropout, see Table 5.

Exploratory analyses—Rehearsal effects
In further exploratory analyses, we examined rehearsal effects. Investigating error rates at T1, we found that the chance for participants making an error was lower in the second than in the first application of the SST, as indicated by a significant main effect of application, OR [95% CI] = 0.62 [0.46, 0.81], SE = 0.11, p = 0.001 (error rate first application: 21.92%, second application: 16.38%). This effect did not differ between the two SST versions as indicated by a non-significant main effect of version, OR [95% CI] = 1.38 [0.95, 2.01], SE = 0.27, p = 0.08. Further exploratory analyses comparing participants who dropped out with participants who did not did not reveal that they did not differ in age, self-reported language fluency, their SST bias score, or their severity of depressive symptoms. However, participants who dropped out showed higher PTCI and AQ scores, compared to those who did not. For descriptive statistics of analyses as well as results of the significance tests concerning dropout, see Table 5.

Table 2. Bivariate correlations between the SST and questionnaire measures.

|          | M (Range) | SD  | PT CI T1 | PCL T1 | QIDS T1 | SST T1 | EDEQ T1 | FSS T1 | AQ T1 | PTCI T2 | PCL T2 |
|----------|-----------|-----|----------|--------|---------|--------|---------|--------|-------|---------|--------|
| T1 SST   | -0.26 –0.78 | 0.18 | -0.58 [0.18, 0.67] | 0.51 [0.39, 0.61] | 0.48 [0.36, 0.59] | -0.58 [0.68, 0.42] | -0.30 [0.16, 0.24] | 0.16 | 0.24 [0.11, | 0.32 | 0.18, 0.45 | 0.50 | 0.35, 0.45 |
| T2 SST   | -0.17 –0.73 | 0.19 | -0.56 [0.44, 0.65] | 0.50 [0.39, 0.61] | 0.49 [0.37, 0.60] | -0.53 [0.63, 0.42] | -0.30 [0.17, 0.23] | 0.10 | 0.30 [0.16, | 0.50 | 0.36, 0.47 | 0.33 | 0.63 |
| v. A     | -0.26 –0.72 | 0.20 | -0.51 [0.40, 0.61] | 0.43 [0.31, 0.54] | 0.40 [0.28, 0.52] | -0.55 [0.64, 0.45] | -0.26 [0.12, 0.31] | 0.12 | 0.41 [0.25, | 0.37 | 0.21, 0.51 |
| v. B     | -0.27 –0.73 | 0.21 | -0.50 [0.41, 0.54] | 0.45 [0.33, 0.52] | 0.40 [0.28, 0.52] | -0.55 [0.64, 0.45] | -0.26 [0.12, 0.31] | 0.12 | 0.41 [0.25, | 0.37 | 0.21, 0.51 |

Spearman’s correlations with bootstrapped 95% CIs are presented. For T1, the sample size is N = 214.
SST: Scrambled Sentences Task; v.: version, PTCI: Posttraumatic Cognitions Inventory. PCL: PTSD Checklist for DSM-5; QIDS: Quick Inventory of Depressive Symptomatology – Self-report; AST: Ambiguous Scenarios Task; EDEQ: Eating Disorder Examination Questionnaire; FSS: Fear of Spiders Screening; AQ: Acrophobia Questionnaire.

N = 213, for T2, the sample size is N = 145. All correlations > .17 are significant after correcting for multiple testing while the correlation = .17 did not remain significant.

r_s = .29...32, correlations with the SST total score were positive, indicating that higher scores on the SST were associated with more severe symptoms of eating disorders, spider phobia, and fear of heights. However, as expected, these correlations were relatively weak. The pattern of results concerning divergent and convergent validity was the same for the SST total score, versions A and B. For detailed correlations, see Table 2.
the first or second application, or between versions (overall error rate: 14.40%).

Investigating the rates of dysfunctional sentences at T1, we found no significant main effects or interactions ($p$s > .220), indicating that the chance that participants generated sentences in a dysfunctional way did not differ between the first and second application (first: 25.94%, second: 25.23%) or between versions (dysfunctional sentences: version A: 25.85%, version B: 25.28%) on T1. The same pattern of results was found for T2 ($p$s > .341, overall percentage of dysfunctional sentences: 27.21%). In summary, we found that on T1, there was an initial decrease in the error rate after

| Table 3. Results of stepwise multiple regression testing the unique variance on the concurrent PCL explained by the SST. |
|---------------------------------------------|
| **T1** | **B [95% CI]** | **SE B** | **β [95% CI]** | **SEβ** | **t** | **p** |
| **Step 1** | $R^2_{adj} = .01$ | .19 | -.06 [-.20, .08] | .07 | .85 | .398 |
| Age | -.16 [-.43, .16] | .29 | -.20 [-.57, .18] | .19 | 1.03 | .302 |
| Gender (non-female) | -3.05 [-7.61, 1.52] | .56 | .13 [-.01, .27] | .07 | 1.86 | .064 |
| Event distress then | 1.04 [-.05, 1.99] | .07 | -.13 [-.15, .07] | .09 | 9.28 |
| Months since event | .00 [-.02, .03] | .07 | [.01, .13] | .07 | 1.03 | .302 |
| **Step 2** | $ΔR^2_{adj} = .30$ | .16 | .02 [-.10, .13] | .06 | .27 | .785 |
| Age | .04 [-.23, .33] | .29 | -.34 [-.27, .24] | .16 | 1.16 | .876 |
| Gender (non-female) | -3.9 [-4.90, 5.55] | .47 | -.04 [-.19] | .06 | 1.22 | .223 |
| Event distress then | .57 [-.34, 1.42] | .11 | .02 [-.20] | .06 | 2.33 | .01 |
| Months since event | .00 [-.02, .03] | .06 | -.10 [-.13] | .06 | .23 | .817 |
| SST total score (T1) | 48.52 [36.69, 60.16] | .25 | 45.68 | .06 | 9.67 | .001 |
| **Step 3** | $ΔR^2_{adj} = .24$ | .13 | -.07 [-.16, .02] | .05 | 1.46 | .148 |
| Age | .19 [-.43, .07] | .25 | -.12 [-.27, .24] | .13 | 1.3 | .896 |
| Gender (non-female) | -.26 [-.29, 3.01] | .38 | .11 [.02, .20] | .05 | 2.33 | .021 |
| Event distress then | .88 [.27, 1.56] | .07 | -.01 [-.08] | .06 | 1.03 | .318 |
| Months since event | .01 [-.01, .03] | .05 | .03 [-.07, .12] | .05 | .6 | .550 |
| SST total score (T1) | 13.70 [7.96, 26.30] | .00 | 16.04 | .06 | 2.65 | .009 |
| PTCI total score (T1) | .31 [.24, .37] | .03 | .63 [.52, .75] | .06 | 10.72 | .001 |

| **T2** | **B [95% CI]** | **SE B** | **β [95% CI]** | **SEβ** | **t** | **p** |
|---------------------------------------------|
| **Step 1** | $R^2_{adj} = .00$ | .25 | -.09 [-.26, .08] | .09 | 1.02 | .308 |
| Age | -.25 [-.60, .23] | .25 | -.04 [-.49, .38] | .22 | 2.5 | .803 |
| Gender (non-female) | .93 [-7.06, 8.71] | .73 | .15 [-.02, .32] | .09 | 1.75 | .083 |
| Event distress then | 1.28 [-1.12, 2.49] | .02 | .09 [-.08, .26] | .09 | 1.00 | .318 |
| Months since event | .02 [-.01, .07] | .02 | -.01 [-.08, .26] | .09 | 1.03 | .318 |
| **Step 2** | $ΔR^2_{adj} = .36$ | .13 | -.07 [-.16, .02] | .05 | 1.46 | .148 |
| Age | .19 [-.43, .07] | .25 | -.12 [-.27, .24] | .13 | 1.3 | .896 |
| Gender (non-female) | -.26 [-.29, 3.01] | .38 | .11 [.02, .20] | .05 | 2.33 | .021 |
| Event distress then | .88 [.27, 1.56] | .07 | -.01 [.02, .20] | .06 | 2.33 | .021 |
| Months since event | .01 [-.01, .03] | .05 | .03 [-.07, .12] | .05 | .6 | .550 |
| SST total score (T2) | 52.89 [40.65, 66.47] | .00 | 48.76 | .06 | 8.88 | <.001 |
| **Step 3** | $ΔR^2_{adj} = .28$ | .15 | -.05 [-.16, .05] | .05 | .97 | .332 |
| Age | -.15 [-.42, .13] | .15 | -.05 [-.16, .05] | .05 | .97 | .332 |
| Gender (non-female) | .39 [-3.74, 7.58] | .22 | .02 [-.25, .29] | .14 | 1.7 | .863 |
| Event distress then | .65 [-.03, 1.45] | .08 [-.03, .18] | .05 | 1.47 | .143 |
| Months since event | .02 [.00, .04] | .01 | .09 [-.02, .19] | .05 | 1.64 | .103 |
| SST total score (T2) | 14.51 [3.24, 29.79] | .03 | .17 [.03, .31] | .07 | 2.49 | .014 |
| PTCI total score (T2) | .34 [.26, .40] | .03 | .68 [.55, .81] | .07 | 10.31 | <.001 |

CIs for B are bootstrapped. T1: N = 214, T2: N = 144 (for T2, one participant was excluded due to missing values on “Age” and “Gender.” QIDS: Quick Inventory of Depressive Symptomatology – Self-report; PTCI: Posttraumatic Cognitions Inventory; SST: Scrambled Sentences Task score across both versions.)
the first application. This effect was not found at T2. Further, rates of dysfunctional sentences did not differ between versions or applications within each assessment time point.

Investigating error rates across time, we found that participants were more likely to make errors on the SST in the first application (error rate first application: 21.92%), compared to the later applications: Application 2 (error rate: 16.38%), OR [95% CI] = 1.60 [1.37, 1.87], SE = .10, p < .001, application 3 (error rate: 14.90%), OR = 1.63 [1.35, 1.96], SE = .12, p < .001, application 4 (error rate: 13.90%), OR = 1.82 [1.50, 2.20], SE = .13, p < .001. However, after the initial decrease, the error rate remained relatively stable as indicated by non-significant pairwise comparisons between the other applications (ps > .310). We further found that participants were less likely to form negative sentences in application 2 (negativity rate: 25.23%) compared to application 4 (negativity rate: 27.51%), OR = .83 [.69, .99], SE = .06, p = .035. Finally, differences between the other applications were non-significant (negativity rate application 1: 25.94%, application 3: 26.90%, ps > .078) indicating that there was a slight increase in negativity rates between the second application on T1 compared to the second application on T2, but not between the others.

### Discussion

The aim of the present study was to develop a novel SST version to assess DAs in the context of potentially traumatic life events and to examine its psychometric properties in a sample exposed to such an event. Further, we aimed to investigate whether the SST would explain unique variance in concurrent PTSD-related symptoms above self-report. Finally, we aimed to develop two parallel versions of the SST to be used in experimental settings.

As expected, the SST was strongly correlated with both concurrent PTSD-related symptoms and PTSD-related DAs. This pattern was found within both time points, indicating good convergent validity. Concerning divergent validity, we found the expected small correlations with symptoms of eating disorders and specific phobias. However, we also found large correlations with both symptoms of depression as well as with depression-related interpretation biases, thus questioning the SST’s divergent validity. Further, the SST was a significant predictor of concurrent PTSD symptoms when taking the PTCI into account, which indicates that the SST could explain variance in PTSD symptoms beyond self-reported DAs. Concerning reliability, we found good internal consistency and test-retest reliability and consistency between both parallel versions of the SST. In exploratory analyses, we further found that participants made fewer errors on the SST after the repeated application, yet their bias score remained stable.

To summarize, results on the SST’s psychometric properties are mostly as expected and mirror those of the earlier study in the context of negative life events. That is, Viviani, Mahler et al. (2018) found that the SST was highly correlated with measures of PTSD-related DAs such as the PTCI; however, the association with depression was also strong. Further, our results are in line with those of a recent

### Table 4. Exploratory analyses of associations between the PCL-5, the PTCI, and divergent validity measures.

| Measure | PCL T1 | PTCI T1 |
|---------|--------|---------|
| PCL T1  | -      | -       |
| PTCI T1 | 75 [67, .81] | -       |
| QIDS    | .66 [57.74] | .64 [55.72] |
| AST     | -.31 [-.44, -.17] | -.41 [-.53, -.28] |
| EDEQ    | .37 [.25, .49] | .34 [.22, .46] |
| FSS     | .18 [.06, .31] | .17 [.04, .30] |
| AQ      | .34 [.21, .46] | .33 [.19, .45] |

Bivariate Spearman’s correlations are presented. PTCI: Posttraumatic Cognitions Inventory; PCL: PTSD Checklist for DSM-5; QIDS: Quick Inventory of Depressive Symptomatology – Self-report; AST: Ambiguous Scenarios Task; EDEQ: Eating Disorder Examination Questionnaire; FSS: Fear of Spiders Screening; AQ: Acrophobia Questionnaire.

* N = 213. All correlations are statistically significant at p < .001, with the exception of those with r = .18 and r = .17, which are statistically significant at p < .01 and p < .05, respectively.

### Table 5. Descriptive statistics for the analyses of participants who dropped out.

| Measure                        | Dropout (N = 68) | No Dropout (N = 146) | t (df) | d [95% CI] | p   |
|-------------------------------|-----------------|----------------------|--------|------------|-----|
| Age                           | 25.46 (6.09)    | 24.72 (5.58)         | t (121.08) = 85 | .13 [-.16, .42] | .400 |
| Fluency German                | 3.60 (.99)      | 3.83 (7.1)           | t (99.80) = 1.68 | .28 [-.01, .57] | .095 |
| SST total score               | .28 (.19)       | .24 (.17)            | t (121.54) = 1.39 | .21 [-.08, .50] | .168 |
| QIDS total score              | 7.34 (5.15)     | 6.24 (4.33)          | t (112.66) = 1.53 | .24 [-.05, .53] | .130 |
| Number correct solutions full SST | 30.87 (6.19)  | 33.03 (3.99)         | t (93.81) = 2.63 | .45 [.15, .75] | .010 |
| PCL-5 total score             | 22.03 (17.09)   | 17.09 (14.47)        | t (113.38) = 2.06 | .32 [.03, .62] | .041 |
| PTCI total score              | 86.25 (38.19)   | 75.17 (27.96)        | t (101.69) = 2.14 | .35 [.06, .65] | .035 |
meta-analysis investigating the psychometric properties of the SST (Wüst et al., 2022). Here, it was found that the SST generally had good convergent validity as indicated by overall high correlations between the SST and symptoms of the target disorders, yet correlations with unspecific symptom measures were also high. Further, our findings on the SST’s reliability are in line with Wüst et al. (2022) as we replicated good internal consistency and retest reliability. The latter is an important addition to previous research as previous findings on the temporal stability of DAs measured via the SST have been limited to only a small number of studies.

Broadening the scope of our interpretations, our results allow the examination of the SST’s divergent validity in relation to the comorbidity of symptoms of certain disorders. While we found the SST to be strongly correlated with both symptoms of PTSD and depression, correlations with specific phobias and eating disorders were overall weaker. Regarding the former, both depression and PTSD are common disorders following a traumatic event and frequently occur comorbidly while sharing common symptom clusters (e.g., concentration problems or negative affect) (Ehring et al., 2006; Lewis et al., 2019; Marmar et al., 2015). Our findings might therefore indicate that this comorbidity could limit the SST’s divergent validity specifically in relation to PTSD and depression. Conversely, divergent validity may be better in relation to disorders that co-occur less frequently with PTSD (Mitchell et al., 2012). However, in exploratory analyses, we found similar patterns of correlations between the PTCI (and also for the PCL) and measures included for assessing divergent validity, suggesting that the limited divergent validity of the SST (in relation to PTSD vs. depression) is not necessarily inherent to the SST itself but may rather reflect a general association between trauma-related DAs and symptoms of this psychopathology following a potentially traumatic life event (Ehring et al., 2006, 2008). Hence, while the aim of our study was to validate an SST assessing DAs specific to PTSD, we interpret our findings as such that our SST may currently be best conceptualized as assessing trauma-related DAs. These DAs in turn are associated with symptoms of PTSD, but also symptoms of depression. This could be a result of comorbidity and shared symptoms between depression and PTSD. Additionally, although results of a recent network analysis amongst children and adolescents suggest that both syndromes have distinct features, trauma-related DAs were found to be equally associated with both depression and PTSD (de Haan et al., 2020). Future work using a prospective design could investigate the relative specificity of the current SST and the original depression-focused version, or specific subsets of their items, in predicting depression versus PTSD outcomes following potentially traumatic life events (for an approach to investigate the SST’s specificity, see Krahé et al., 2022).

Moving on to a further aim of our study, we found preliminary evidence that the SST would predict variance in PTSD-related symptoms beyond self-report as measured via the PTCI. This finding suggests that the SST adds value to the assessment of DAs solely via self-report which is in line with other studies investigating the benefit of non-self-report measures to assess trauma-related DAs (Lindgren et al., 2013) and our results also support the SST’s general value in the investigation of DAs and interpretation biases. To illustrate, in the context of depression, the inclusion of the cognitive load appears to provide the SST with benefits over simple self-report (e.g., the SST itself without a cognitive load), for example, in predicting a future diagnosis of depression (Rude et al., 2003) or showing differential response to cognitive bias modification versus CBT (Bowler et al., 2012). One explanation for this is that the cognitive load prevents suppression of dysfunctional cognitions that participants would otherwise engage in. More generally, these findings highlight the potential utility of developing DA assessments that tap into more automatic processes than self-report, such as the SST with a cognitive load, and including them in further studies. This may be particularly relevant for interventional studies where demand effects can be especially strong, or longitudinal studies where there is an interest in measuring an underlying cognitive vulnerability while someone is not currently experiencing symptoms of psychopathology. In fact, how useful such measures are in different contexts is an empirical question that can only be addressed via including them in a range of different kinds of studies.

Overall, our findings indicate the potential value and utility of the SST as a measure of DAs in the context of potentially traumatic life events or PTSD, particularly when also taking into account the broader literature on assessing dysfunctional interpretational processing in the context of emotional psychopathology. To illustrate, in earlier studies, the SST has successfully been applied together with eye-tracking measures enabling the simultaneous investigation of DAs/interpretation biases and attentional biases (Sanchez et al., 2015; Sñirlea et al., 2020). While both types of biases play an important role in PTSD (Bomyea et al., 2017) and emotional psychopathology in general, there is little evidence for their interplay. Hence, applying the SST together with measures of other cognitive biases has the potential to advance the field of PTSD and anxiety-relevant research, by facilitating the investigation of combined cognitive biases, a strain of research that has made important contributions to the investigation of other disorders (i.e., the combined cognitive bias hypothesis, see Everaert et al., 2017). Further, in a recent study (O’Connor et al., 2021), the SST was compared to purely performance-based (indirect) and self-report measures of interpretation biases. Results showed that the SST’s correlation with disorder-related symptoms was stronger than that for unrelated symptoms.
Interestingly, the SST was uncorrelated with fully indirect and only moderately correlated with self-report measures. These results point out that the SST is a highly useful tool for the assessment of interpretational processing, while the SST seems to be neither a fully direct nor indirect measure (O’Connor et al., 2021). While these results were found in the context of depression, further studies should investigate whether similar results concerning the relationship between the SST and other measures of DAs can be found in the context of negative life events and PTSD.

Previous research has also demonstrated the SST’s potential to validly assess changes in DAs, for example, in experimental settings such as in the context of CBM (Bradatsch et al., 2020) where it has been found that in particular the cognitive load enables the SST to be able to detect effects of CBM training paradigms (Bowler et al., 2012). This potential is further underlined by our promising findings on the psychometric properties of the SST, namely, good retest reliability and the availability of two parallel versions which can be used for the repeated assessment of DAs, and our exploratory findings that although participants make fewer errors after repeated application, their bias score remains stable in absence of a manipulation. This is especially valuable in research on the manipulation of DAs, for example, in analog or actual trauma. While near-transfer effects, that is, effects of CBM on an Encoding Recognition Task (Mathews & Mackintosh, 2000), are commonly found, the investigation of far-transfer effects on, for example, symptom measures provides a less consistent picture (Woud, Zlomuzica et al., 2018; Würtz et al., 2021). This makes it necessary to cross-validate and triangulate current findings across different measures of DAs to shed light on these current issues to potentially open novel routes for the improvement of such paradigms (Munafo & Davey Smith, 2018).

Our results need to be interpreted in the context of a number of limitations. A first limitation to our study is the dropout from T1 to T2 of ~ 30%. While our estimations of test–retest reliability were based on a reasonably sized sub-sample, our findings still need to be interpreted cautiously as a potentially non-random dropout could have influenced the results.

As an initial validation study, this study was carried out in an online setting to recruit a large sample. We therefore could not ensure that all participants were highly concentrated during the study, and we could not standardize the application of the SST concerning technical aspects such as screen size, the used mouse and browser. This limitation becomes more apparent considering previously mentioned differences in the presentation of the SST between the web browsers. While we do not consider this difference to be a threat to the validity of our results, which in fact indicate the robustness of the basic task given a potentially sub-optimal implementation, it is possible that a cleaner pattern of results would be found in a more controlled lab administration of this SST. This limitation also becomes relevant considering the deviation from the pre-registered participant exclusion criterion based on the finding that the error rate was higher than expected. While most results did not differ between our pre-registered (more than 30% errors on the SST) and our post-hoc applied participant exclusion criterion (no valid sentences on the SST), our finding that the SST explains variance beyond the PTQI was only found using the more lenient post-hoc criterion. It is encouraging that a strict accuracy criterion is not required to use the measure, as excluding participants based on a strict error rate criterion poses the risk of reducing the generalizability of results obtained and the utility of the SST as a measure. However, the difference in results that this exclusion criterion made means we have to be more cautious about the robustness of our findings.

Some limitations concern our sample. Using the LEC, we aimed to include only participants who had been exposed to a self-reported, potentially traumatic negative life event, but it is unclear whether these events would always fulfill DSM-5 criteria for a traumatic event. Further, our sample was considerably homogenous, consisting mostly of female participants who were native speakers and finished at least A-level education. Results from this study therefore might not be generalizable to more diverse samples, especially samples less fluent in German. Finally, our sample consisted of participants with on average mild depression and PTSD symptoms on a subclinical level. Our results are therefore potentially not generalizable to participants with higher levels of psychopathology such as patients with PTSD and/or depression. We need to be particularly cautious in relation to the longitudinal data given our exploratory analyses indicating that participants who dropped out made on average more errors and had slightly higher symptoms of PTSD compared to those who completed the study.

In conclusion, we found that the SST for assessing DAs in the context of potentially traumatic life events had good convergent validity, while findings on the divergent validity were more mixed, especially for symptoms of depression. Further results indicated the SST’s capability of explaining variance in PTSD-related symptoms beyond self-report, and its utility was further underlined by our findings of good internal consistency and test–retest reliability. Finally, we found that the two versions were approximately parallel, and applying the SST repeatedly provided a stable assessment of DAs suggesting that the SST is a potentially valuable tool in research on the experimental manipulation of DAs in the context of potentially traumatic life and analog trauma.

Acknowledgments
We would like to thank Dr. Tobias Teismann and Dr. Jan C. Cwik for their feedback on the stimuli for the SST. We would further like
to thank Dr. Christian Leson for his support with the server maintenance. We acknowledge support by the Open Access Publication Funds of the Ruhr-Universität Bochum.

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
Felix Würtz is supported by a doctoral scholarship of the Studienstiftung des deutschen Volkes. Marcella L. Woud is funded by the Deutsche Forschungsgemeinschaft (DFG) via the Emmy Noether Programme (WO 2018/3-1). Jürgen Margraf and Marcella L. Woud are Principal Investigators of a project within the SFB 1280 “Extinction Learning” awarded by the DFG (Projectnr.: 316803389). Project title: “The impact of modification of stimulus-outcome contingencies on extinction and exposure in anxiety disorders.” The funding bodies had no role in the design of the study, the collection, analysis, and interpretation of the data, or the preparation of the manuscript.

Ethics approval statement
All study procedures were approved by the local ethics committee of the Psychology department at Ruhr-University Bochum (approval number: 596).

Data availability
This study’s design was pre-registered and the registration protocol and the anonymized dataset generated and analyzed, as well as the used analysis script and material (excluding standardized questionnaires), are available on the Open Science Framework (OSF) via the following link: https://osf.io/bnzp3/

ORCID IDs
Felix Würtz © https://orcid.org/0000-0003-1627-9432
Simon E. Blackwell © https://orcid.org/0000-0002-3313-7084
Jürgen Margraf © https://orcid.org/0000-0001-5207-7016
Marcella L. Woud © https://orcid.org/0000-0002-4974-505X

Supplemental Material
Supplemental material for this article is available online.

Notes
1. Unique variance was labeled “specificity” in the pre-registration. For analyses concerning predictive validity, specified in the pre-registration, see Supplementary Material.
2. The web application was programmed to be compatible with common web browsers yet was optimized for Mozilla Firefox. In other browsers than Firefox, the numbers remained on the screen for a short interval after the buttons disappeared (~500 ms) yet disappeared before the fixation cross for the next trial was presented. In Firefox, the numbers immediately disappeared with the buttons. However, while we consider this to be a limitation, we do not consider this to be a threat to the validity of the data as all trials were separated by the presentation of a fixation cross (1 s), and thus the attention of participants was directed toward the center of the screen before each trial, irrespective of the used browser. Further, while we found slight differences in the error rates between browsers, SST scores did not differ. For analyses comparing browsers, see Supplementary Material. An improved version of the SST that fixes this issue is available on the OSF alongside the original version used in this study.
3. Due to an administration error, for one item (item 11—self-image), the answer option for “2” was presented for both the answers “2” and “3.” This error was accounted for by recoding answers with the value “3” to “2” (three participants) as the correct option was only presented for the latter.
4. Note that item 8 is included in both the “shape concern” and “weight concern” subscales. Due to an administration error, items 19–21 were not presented to participants. These items are part of the “Eating concerns” subscale and thus the mean for this subscale was calculated using only 2 items (items 7 and 9).
5. When participants indicated that they were not exposed to an event in addition to another option (e.g., by checking both “not sure whether it happened” and “not exposed”), this was still counted as “exposed.” This procedure was decided on after data inspection to take all potential traumatic events into account.
6. Gender was assessed using 3 categories (female, male, and non-binary/diverse) with one participant in the third category. For the inclusion of Gender as a variable in regression models, we therefore collapsed the categories “male” and “non-binary” to the category “non-female” to avoid estimating gender differences based on one data point.
7. We note that in the pre-registered protocol, we planned to add both PTCI and QIDS scores together in the third regression step. However, based on reviewer feedback, we noticed that the inclusion of depressive symptoms as a predictor would obscure our conclusions on whether the SST would explain variance beyond self-report in PTSD-related symptoms. We therefore did not include the QIDS as a predictor. Results of the regression including the QIDS as a predictor can be found in the Supplementary material.
8. First, analyses were carried out including also the effect of “time” (T1, T2) and all two- and three-way interactions. However, due to convergence problems, we decided to analyze error rates/negativity rates between versions separately for T1 and T2.
9. The error was found to be associated with the server settings. As all data was generated on the client side, the technical difficulties led to data loss yet did not impair the validity of the data.
10. The dropout rate was calculated as the percentage of participants who answered at least the PTCI on T1 but did not start the survey on T2.

11. The procedure for excluding participants based on their error rate deviated from the protocol. Based on the laboratory-based study by Brockmeyer et al. (2018), we planned to exclude participants with error rates > 30% which would have resulted in 33 excluded participants for T1 (15.14%) and 9 for T2 (6.16%). However, after data inspection, we decided to apply a more liberal criterion, excluding only participants with no valid trials in one of the SST versions which was similarly applied in other online studies (Hirsch et al., 2020). Results of analyses applying the 30% criterion did not differ from results presented here if not otherwise indicated. For additional results, see Supplementary Material.

12. Five participants completed T2 thirty or more days after T1. Repeating the analyses on the test–retest reliability excluding these participants did not affect the results.

13. Applying the pre-registered 30% error-rate exclusion criterion, the SST did not remain a unique predictor of PTSD-related symptoms in step 3 when taking the PTCI into account. However, at T2, results did not differ applying either exclusion criterion.

References

American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders (5th ed.). American Psychiatric Association.

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

Berna, C., Lang, T. J., Goodwin, G. M., & Holmes, E. A. (2011). Developing a measure of interpretation bias for depressed mood: An ambiguous scenarios test. Personality and Individual Differences, 51(3), 349–354. https://doi.org/10.1016/j.paid.2011.04.005

Blackwell, S. E., Ehring, T., Gladwin, T. E., Margraf, J., & Woud, M. L. (2021). The relationship between self-traumatized and self-vulnerable automatic associations and posttraumatic stress symptoms among adults who have experienced a distressing life event. Current Psychology. Advance online publication. https://doi.org/10.1007/s12144-021-02441-9

Bofòa, J. W., Norr, A. M., Tock, J. L., Amir, N., & Schmidt, N. B. (2018). Development of the Interpretation bias Index for PTSD. Cognitive Therapy and Research, 42(5), 720–734. https://doi.org/10.1007/s10608-018-9915-8

Bomyea, J., Johnson, A., & Lang, A. J. (2017). Information processing in PTSD: Evidence for biased attentional, interpretation, and memory processes. Psychopathology Review, 4(3), 218–243. https://doi.org/10.5127/pr.037214

Bowler, J. O., MacKintosh, B., Dunn, B. D., Mathews, A., Dalgleish, T., & Hoppit, L. (2012). A comparison of cognitive bias modification for interpretation and computerized cognitive behavior therapy: Effects on anxiety, depression, attentional control, and interpretive bias. Journal of Consulting and Clinical Psychology, 80(6), 1021–1033. https://doi.org/10.1037/a0029932

Bradatsch, S., Vahl, M. D., Potterton, R., Gordon, G., Schmidt, U., & Brockmeyer, T. (2020). Interpretation bias modification to reduce body dissatisfaction – A randomized controlled pilot study in women with elevated weight and shape concerns. Journal of Eating Disorders, 8(1), 34. https://doi.org/10.1186/s40337-020-00305-4

Brewin, C. R., Dalgleish, T., & Joseph, S. (1996). A dual representation theory of posttraumatic stress disorder. Psychological Review, 103(4), 670–686. https://doi.org/10.1037/0033-295X.103.4.670

Brockmeyer, T., Anderle, A., Schmidt, H., Febry, S., Wünsch-Leiteritz, W., Leiteritz, A., & Friederich, H. C. (2018). Body image related negative interpretation bias in anorexia nervosa. Behaviour Research and Therapy, 104(May), 69–73. https://doi.org/10.1016/j.brat.2018.03.003

Brown, L. A., Belli, G. M., Asnaani, A., & Foa, E. B. (2019). A review of the role of negative cognitions about oneself, others, and the world in the treatment of PTSD. Cognitive Therapy and Research, 43(1), 143–173. https://doi.org/10.1007/s10608-018-9938-1

Bryant, R. A., & Guthrie, R. M. (2005). Maladaptive appraisals as a risk factor for posttraumatic stress. Psychological Science, 16(10), 749–752. https://doi.org/10.1111/j.1467-9280.2005.01608.x

Cheung, J., & Bryant, R. A. (2017). The impact of appraisals on intrusive memories. Journal of Behavior Therapy and Experimental Psychiatry, 54(March), 108–111. https://doi.org/10.1016/j.jbtep.2016.07.005

Cohen, D. C. (1977). Comparison of self-report and overt-behavioral procedures for assessing acrophobia. Behavior Therapy, 8(1), 17–23. https://doi.org/10.1016/S0005-7894(77)80116-0

Dalgleish, T. (2004). Cognitive approaches to posttraumatic stress disorder: The evolution of multirepresentational theorizing. Psychological Bulletin, 130(2), 228–260. https://doi.org/10.1037/0033-2909.130.2.228

de Haan, A., Landolt, M. A., Fried, E. I., Kleinke, K., Alisc, E., Bryant, R., Salmon, K., Chen, S., Liu, S., Dalgleish, T., McKinnon, A., Alberici, A., Claxton, J., Diehle, J., Lindauer, R., de Roos, C., Halligan, S. L., Hiller, R., Kristensen, C. H., & Meiser-Stedman, R. (2020). Dysfunctional posttraumatic cognitions, posttraumatic stress and depression in children and adolescents exposed to trauma: A network analysis. Journal of Child Psychology and Psychiatry, 61(1), 77–87. https://doi.org/10.1111/jcpp.13101

de Kleine, R. A., Woud, M. L., Ferentzi, H., Hendriks, G.-J., Broekman, T. G., Becker, E. S., & Van Minnen, A. (2019). Appraisal-based cognitive bias modification in patients with posttraumatic stress disorder: A randomised clinical trial.
European Journal of Psychotraumatology, 10(1), 1625690. https://doi.org/10.1080/20008198.2019.1625690

de Voogd, E. L., De Hullu, E., Heyes, S. B., Blackwell, S. E., Wiers, R. W., & Salemink, E. (2017). Imagine the bright side of life: A randomized controlled trial of two types of interpretation bias modification procedure targeting adolescent anxiety and depression. Plos One, 12(7), 1–24. https://doi.org/10.1371/journal.pone.0181147

de Winter, J. C. F., Gosling, S. D., & Potter, J. (2016). Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data. Psychological Methods, 21(3), 273–290. https://doi.org/10.1037/met0000079

Ehlers, A. (1999). Posttraumatische belastungsstörung. Hogrefe Göttingen.

Ehlers, A., & Clark, D. M. (2000). A cognitive model of post-traumatic stress disorder. Behaviour Research and Therapy, 38(4), 319–345. https://doi.org/10.1016/S0005-7967(99)00123-0

Ehring, T., Ehlers, A., & Glucksman, E. (2006). Contribution of cognitive factors to the prediction of post-traumatic stress disorder, phobia and depression after motor vehicle accidents. Behaviour Research and Therapy, 44(12), 1699–1716. https://doi.org/10.1016/j.brat.2005.11.013

Ehring, T., Ehlers, A., & Glucksman, E. (2006). Contribution of cognitive factors to the prediction of post-traumatic stress disorder, phobia and depression after motor vehicle accidents. Behaviour Research and Therapy, 44(12), 1699–1716. https://doi.org/10.1016/j.brat.2005.11.013

Foa, E. B., Steketee, G., & Orsillo, S. M. (1999). The posttraumatic cognitions inventory (PTCI): Development and validation. Psychological Assessment, 11(3), 303–314. https://doi.org/10.1037/1040-3590.11.3.303

Fox, J., & Weisberg, S. (2019). An R companion to applied regression (third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Fox, J., & Weisberg, S. (2019). An R companion to applied regression (third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gómez de La Cuesta, G., Schweizer, S., Diehle, J., Young, J., & Meiser-Stedman, R. (2019). The relationship between mal-adaptive appraisals and posttraumatic stress disorder: A meta-analysis. European Journal of Psychotraumatology, 10(1), 1620084. https://doi.org/10.1080/20008198.2019.1620084

Hilbert, A., Tuschen-Cafler, B., Karwautz, A., Niederhofer, H., & Munsch, S. (2007). Eating disorder examination questionnaire (EDE-Q). Diagnostica, 53(3), 144–154. https://doi.org/10.1027/1015-5759/a000093

Hirsch, C. R., Meeten, F., Gordon, C., Newby, J. M., Bick, D., & Moulds, M. L. (2020). Repetitive negative thinking and interpretation bias in pregnancy. Clinical Psychology in Europe, 2(4), 1–17. https://doi.org/10.32872/cpe.v2i4.3615

Hirsch, C. R., Meeten, F., Krahé, C., & Reeder, C. (2016). Resolving ambiguity in emotional disorders: The nature and role of interpretation biases. Annual Review of Clinical Psychology, 12(1), 281–305. https://doi.org/10.1146/annurev-clinpsych-021815-093436

Klein, B., Ehlers, A., & Glucksman, E. (2007). Early predictors of chronic post-traumatic stress disorder in assault survivors. Psychological Medicine, 37(10), 1457–1467. https://doi.org/10.1017/S0033291707001006

Klein, B., Grey, N., Wild, J., Nussbeck, F. W., Stott, R., Hackmann, A., Clark, D. M., & Ehlers, A. (2013). Cognitive change predicts symptom reduction with cognitive therapy for posttraumatic stress disorder. Journal of Consulting and Clinical Psychology, 81(3), 383–393. https://doi.org/10.1037/a0031290

Krahé, C., Meeten, F., & Hirsch, C. R. (2022). Development and psychometric evaluation of a scrambled sentences test specifically for worry in individuals with generalized anxiety disorder. Journal of Anxiety Disorders. Advance online publication. https://doi.org/10.1016/j.janxdis.2022.102610

Krüger-Gottschalk, A., Knaevelsrud, C., Rau, H., Dyer, A., Schäfer, I., Schellong, J., & Ehring, T. (2017). The German version of the posttraumatic stress disorder checklist for DSM-5 (PCL-5): Psychometric properties and diagnostic utility. BMC Psychiatry, 17(1), 379. https://doi.org/10.1186/s12888-017-1541-6

Lee, J. S., Mathews, A., Shergill, S., & Yiend, J. (2016). Magnitude of negative interpretation bias depends on severity of depression. Behaviour Research and Therapy, 83, 26–34. https://doi.org/10.1016/j.brat.2016.05.007

Lewis, S. J., Arseneault, L., Caspi, A., Fisher, H. L., Matthews, T., Moffitt, T. E., Odgers, C. L., Stahl, D., Teng, J. Y., & Danese, A. (2019). The epidemiology of trauma and post-traumatic stress disorder in a representative cohort of young people in England and Wales. The Lancet Psychiatry, 6(3), 247–256. https://doi.org/10.1016/S2215-0366(19)30031-8

Lindgren, K. P., Kaysen, D., Wernitz, A. J., Gasser, M. L., & Teachman, B. A. (2013). Wounds that can’t be seen: Implicit...
trauma associations predict posttraumatic stress disorder symptoms. *Journal of Behavior Therapy and Experimental Psychiatry*, 44(4), 368–375. https://doi.org/10.1016/j.jbtep.2013.03.003

MacLeod, C., Koster, E. H. W., & Fox, E. (2009). Whither cognitive bias modification research? Commentary on the special section articles. *Journal of Abnormal Psychology*, 118(1), 89–99. https://doi.org/10.1037/a0014878

Mangiafico, S. (2021). *rcompanion: Functions to support extension education program evaluation*. https://CRAN.R-project.org/package = rcompanion

Marmar, C. R., Schlenger, W., Henn-Haase, C., Qian, M., Purchia, E., Li, M., Corry, N., Williams, C. S., Ho, C. L., Horesh, D., Karstof, K. I., Shaley, A., & Kulkia, R. A. (2015). Course of posttraumatic stress disorder 40 years after the Vietnam war. *Fortschritte Der Neurologie Psychiatrie*, 83(12), e17–e22. https://doi.org/10.1055/s-0041-110203

Rohrbacher, H., & Reinecke, A. (2014). Measuring change in depression-related interpretation bias: Development and validation of a parallel ambiguous scenarios test. *Cognitive Behaviour Therapy*, 43(3), 239–250. https://doi.org/10.1080/16506073.2014.919605

Rohrbacher, H., Reinecke, A., & Spätth, C. (2015). A psychometric evaluation of the German version of the quick inventory of depressive symptomatology (QIDS-SR16) in outpatients with depression. *Fortschritte Der Neurologie Psychiatrie*, 83(12), e17–e22. https://doi.org/10.1055/s-0041-110203
Author biographies

Felix Würtz is a PhD student at Ruhr-Universität Bochum. His research interests include the assessment and modification of cognitive biases and the investigation of their role in the aetiology of psychopathology.

Simon E. Blackwell is a postdoctoral researcher in the Department of Clinical Psychology and Psychotherapy, Mental Health Research and Treatment Center at Ruhr-Universität Bochum. His current research focuses on positive mental imagery and translational research methods.

Jürgen Margraf is Alexander von Humboldt-Professor of Clinical Psychology and Psychotherapy as well as co-director of the Mental Health Research and Treatment Center at Ruhr-Universität Bochum. His current research focuses on positive mental health, emotional disorders and psychological treatments.

Marcella L. Woud is a junior professor of Clinical Psychology and Experimental Psychopathology at Ruhr-Universität Bochum. Her work integrates experimental and clinical studies in the context of cognitive biases in anxiety, with a focus on dysfunctional appraisals and interpretation biases.