Journal of Anxiety Disorders 91 (2022) 102610

Contents lists available at ScienceDirect

Journal of Anxiety Disorders

journal homepage: www.elsevier.com/locate/janxdis

Development and psychometric evaluation of a scrambled sentences test specifically for worry in individuals with generalised anxiety disorder

Charlotte Krahé [*],[1], Frances Meeten [2], Colette R. Hirsch

Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

ARTICLE INFO

Keywords:
Scrambled sentences test
Interpretation bias
Cognitive processes
Worry
Anxiety
Generalised anxiety disorder

ABSTRACT

The tendency to draw negative conclusions from ambiguous information (interpretation bias) is prevalent across emotional disorders and plays a key role in the development and maintenance of pathological worry and anxious mood. Assessing interpretation bias using valid and reliable measures is central to empirical research. A commonly used measure of interpretation bias is the scrambled sentences test (SST), originally relating to depression. Given the association between interpretation bias and worry, we aimed to develop and psychometrically evaluate a new version of the SST with items pertaining to common worry domains for use in worry and anxiety research. In Studies 1–3 (analogue samples, combined $N = 288$), the new worry SST showed excellent construct validity (moderate-to-strong associations with worry and anxiety-related measures), and reliability (split-half and test-retest reliability). We confirmed construct validity in Study 4 ($N = 215$ individuals with generalised anxiety disorder). Furthermore, we demonstrated version specificity in analogue and clinical samples: the worry SST was associated with trait worry but not trait rumination, while the original depression SST largely showed the opposite pattern. Overall, the new worry SST is a psychometrically robust measure that may be especially useful for research into cognitive processes underpinning worry and anxiety.

1. Introduction

Ambiguity is part of daily life. For example, if someone laughs when you tell a joke, they could be laughing because it is funny, or because it is such a bad joke that they are laughing at you. People differ in their tendency to draw negative or positive conclusions when presented with ambiguous information, with a bias towards making negative interpretations prevalent in a range of mental health difficulties, including social anxiety (Chen, Short, & Kemps, 2020), generalised anxiety (Krahé, Whyte, Bridge, Loizou, & Hirsch, 2019), paranoia (Trotta, Kang, Stahl, & Yiend, 2020), and depression (Everaert, Podina, & Koster, 2017).

Excessive and uncontrollable worry is a form of repetitive negative thinking observed as a transdiagnostic symptom across anxiety disorders and depression (Ehring & Watkins, 2008), and is a core feature of generalised anxiety disorder (GAD; American Psychiatric Association, 2013). The tendency to generate negative interpretations (interpretation bias) plays a key role in the development and maintenance of pathological worry and anxious mood (Hirsch & Mathews, 2012; Hirsch, Meeten, Krahé, & Reeder, 2016; Mathews & MacLeod, 2005). Facilitating more benign interpretations in people with high levels of worry is associated with less subsequent worry and reduced anxiety compared to those in a control group, whose interpretations were not modified (e.g., Hirsch, Hayes, & Mathews, 2009; Hirsch et al., 2020; Hirsch et al., 2018).

Central to empirical research is the assessment of interpretation bias using robust measures with excellent reliability and validity. In Hirsch et al. (2016), we summarised commonly used methodologies (see also Schoth & Liossi, 2017). One widely used assessment is the scrambled sentences test (SST; Wenzlaff & Bates, 1998, 2000; see Würtz et al., 2022, for a review). This task presents participants with a series of six words that are not in order, and participants are instructed to use five of the words to make a grammatically correct sentence, which reveals a negative or a positive interpretation (see below for examples). The task is often completed within a short time period and under a cognitive load such that participants are asked to memorise a 6-digit string of numbers

[*] Correspondence to: Department of Primary Care and Mental Health, University of Liverpool, Eleanor Rathbone Building, Bedford Street South, Liverpool L69 7ZA, UK.

E-mail address: charlotte.krahe@liverpool.ac.uk (C. Krahé).

[1] Present address: Department of Primary Care and Mental Health, Institute of Population Health, University of Liverpool, Liverpool, UK.

[2] Present address: School of Psychology, University of Sussex, Falmer, Brighton, UK.

https://doi.org/10.1016/j.janxdis.2022.102610
Received 13 September 2021; Received in revised form 11 July 2022; Accepted 4 August 2022
Available online 6 August 2022

to prevent effortful executive control over responses and to allow latent biases to be observed (Schoth & Liossi, 2017; Viviani, Dommes, Bosch, Stingl, & Beschoner, 2018).

The SST was originally developed for use in relation to depression. Accordingly, the items consist of statements pertaining to depression symptoms (e.g., "*I usually feel very good/bad*") and depressive cognitions, including depressive rumination (e.g., "*people do/don't care about me*"; "*something/nothing is wrong with me*"). The SST has been widely used in the study of interpretation bias in depression (Everaert, Duyck, & Koster, 2014; Everaert, Tierens, Uzieblo, & Koster, 2013; Holmes, Lang, & Shah, 2009; Rude, Durham-Fowler, Baum, Rooney, & Maestas, 2010; Rude, Valdez, Odom, & Ebrahimi, 2003; Rude, Wenzlaff, Gibbs, Vane, & Whitney, 2002). The SST explains a unique proportion of variance in depression symptom severity, supporting prior literature and its use as a measure of interpretation bias in emotional disorders (O'Connor, Everaert, & Fitzgerald, 2021).

New versions of the SST have been developed with items pertaining to other clinical problems, such as symptoms of psychosis (e.g., Savulich, Shergill, & Yiend, 2017) and eating disorders (e.g., Brockmeyer et al., 2018). Yet, an SST version relating to worry content does not exist. This is despite interpretation bias being a key factor in maintaining pathological worry (Hirsch & Mathews, 2012) as well as a process common to repetitive negative thinking (Badra et al., 2017; Krahé et al., 2019). Importantly, in individuals with GAD and depression, worry and rumination predict shared variance in the measure of interpretation bias, but each also predicts unique additional variance (Krahé et al., 2019). Thus, although interpretation bias is associated with both depression and anxiety, a version of the scrambled sentences test for use in research that focuses on worry and/or anxious populations is warranted.

While there is high co-morbidity between anxiety and depression (e.g., Lamers et al., 2011), there is evidence to suggest that worry is greater in GAD than depression, and rumination is higher in depression than GAD (Krahé et al., 2019). Furthermore, the content of negative thoughts and the temporal focus differs between worry and rumination, which may in turn affect the content of biased processing. Worry focuses more on potential future threats and rumination on past or ongoing negative events (Watkins, Moulds, & Mackintosh, 2005). Therefore, it seems important to be able to capture cognitive biases in relation to aspects central to the emotional state or disorder in question (Hirsch et al., 2016). Indeed, recent research has demonstrated content specificity of interpretation bias, that is, a negative interpretation bias for ambiguous information pertaining to specific fears in children (Mobach, Rinck, Becker, Hudson, & Klein, 2019), and content specificity of interpretation bias to social situations in people with high social anxiety (Yu, Westenberg, Li, Wang, & Miers, 2019). Lastly, a recent systematic review, which included some of the data presented in this paper, found good convergent validity and internal consistency for the SST but concluded that adaptations are needed to enhance its specificity (Würtz et al., 2022). Accordingly, we set out to develop and validate a new version of the SST with items pertaining specifically to the cognitive component of anxiety, namely worry.

1.1. The current research

This paper presents four studies reporting the development and psychometric evaluation of a new worry SST (wSST) measure in the general population (Studies 1–3) and in individuals with GAD (Study 4). In Study 1, we generated items for the new measure and then examined its reliability and construct validity. In Study 2, we investigated the specificity of the new wSST to worry and anxiety by administering the wSST along with the established depression SST (dSST) measure. This allowed us to examine incremental validity of the wSST in predicting variance in levels of worry and anxiety over and above that explained by the dSST, as well as provide further psychometric evaluation of the existing dSST. In Study 3, we refined the items to present a final version and studied its test-retest reliability. In Study 4, we administered the final wSST and dSST to a large sample of individuals with GAD (with and without comorbid depression) to assess construct validity and specificity. Taken together, we aimed to develop a new, reliable, and valid measure of interpretation bias for use in studies focusing on worry and anxiety.

2. Study 1: Development and initial validation of the worry SST

In this first study, we developed and provided an initial psychometric evaluation of a scrambled sentences test with worry-related items. Two lists were constructed to enable multiple administrations of the task with novel materials each time; for example, to assess interpretation bias before and after an intervention, or under different conditions for the same participant e.g., with/without a cognitive load. Reliability and construct validity were assessed. In particular, we examined whether the tendency to make more positive interpretations, measured using the wSST, would be associated with lower levels of trait worry and anxiety.

2.1. Development of the worry SST

The SST (originally developed by Wenzlaff & Bates, 1998, 2000) involves re-ordering a set of randomly presented words to form grammatically correct sentences. Participants must use five of six words, presented in a 'scrambled' order, to form meaningful sentences. These sentences, by virtue of the words selected, can either be of negative or positive valence. For example, the words "*good to life is cruel me*" can be used to form a negative sentence ("*life is cruel to me*") or a positive sentence ("*life is good to me*"). Unscrambling sentences to form their positive or negative version is indicative of having made a positive or negative interpretation, respectively. The original SST comprised items relating to depression.

To develop SST items relating to worry, we examined common worry domains (Hirsch, Mathews, Lequertier, Perman, & Hayes, 2013; Tallis, Davey, & Bond, 1994), namely lack of confidence, aimless future, work/study competence, financial and physical threat, and relationships. Based on these, we created two separate SST lists, each comprising 20 unique items. Each list contained an equal number of items pertaining to the different worry domains, e.g., "*Approaching new people is fine/scary*" (lack of confidence), "*Everything will turn out fine/badly*" (aimless future), "*I am performing above/below expectations*" (work/study competence), "*I will/won't get into debt*" (financial), and "*I find maintaining relationships easy/difficult*" (relationships). We tried to construct items such that they could only be unscrambled in two ways: one positive, and one negative. Further, we ensured that disambiguating words (i.e., the ones that determine whether the sentence is negative or positive) were not always adjectives. The original items can be found in Supplementary Table 1.

Items were presented to participants alongside established self-report questionnaires (see below). In line with previous studies (Hirsch et al., 2018, 2020; Krahé et al., 2019), we calculated an SST 'positivity index' for each participant by coding sentences unscrambled in a negative or positive manner as zero or one, respectively, and then dividing the number of positively unscrambled sentences by the total number of sentences which participants unscrambled in a grammatically correct fashion. This yielded a ratio measure with scores ranging from zero to one; a score of one indicated that participants only formed positive sentences.[3]

[3] A negativity index could also easily be computed instead by dividing the number of *negatively* unscrambled sentences by the total number of grammatically correct sentences generated or by subtracting the positivity index from 1.

2.2. *Measures of worry and anxiety*

2.2.1. *Penn state worry questionnaire (PSWQ; Meyer, Miller, Metzger, and Borkovec, 1990)*

The PSWQ is a 16-item measure of trait worry comprising items such as, "*I am always worrying about something*", which are rated on a scale from 1 (*not at all typical of me*) to 5 (*very typical of me*) and summed (after reverse-scoring 5 items) to produce an overall score, with higher scores denoting greater worry. The PSWQ is extensively used and has been well-validated (Brown, Antony, & Barlow, 1992; Startup & Erickson, 2006; Stoeber & Bittencourt, 1998; Zlomke, 2009); α = .96 in the present sample.

2.2.2. *General anxiety questionnaire (GAD-7; Spitzer, Kroenke, Williams, and Löwe, 2006)*

The GAD-7 is a 7-item measure of anxiety in which symptoms such as, "*Feeling nervous, anxious, or on edge?*" are rated in reference to the last two weeks as having occurred "*not at all*" (coded zero), "*several days*" (coded 1), "*more than half the days*" (coded 2), or "*nearly every day*" (coded 3). Responses are summed to produce a total score, with higher scores denoting greater anxiety in the last two weeks (α = .93).

2.3. *Participants*

Ethical approval for this study and the following studies was obtained from the King's College London Research Ethics Committee. Participants for this study and Studies 2 and 3 were recruited via the online platform Amazon Mechanical Turk, and were invited to take part in a survey including questionnaires on mood, thinking style, and simple word- and number-based tasks.

The final sample comprised *N* = 99 participants, resident in the USA. Demographic information for all four studies is presented in Table 1.

Table 1
Demographic characteristics for studies 1-4.

		Study 1 (*N* = 99)	Study 2 (*N* = 92)	Study 3 (*N* = 97)[a]	Study 4 (*N* = 215)
Age - mean (SD), range		33.13 (10.45), 19-65	36.70 (11.96), 19-78	33.85 (10.09), 20-59	32.99 (10.99), 18-62
Gender (% female)		55.60	46.70	41.24	90.23
World region of origin (%)	North America	99	88.04	96.9	0
	South America	0	0	0	0
	Europe	1	1.09	1.03	100
	Asia	0	5.44	0	0
	Africa	0	0	0	0
	Oceania/ Australasia	0	1.09	0	0
	Information missing	0	4.34	2.06	0
Ethnicity (%)	White	70.71	79.35	80.41	81.40
	Asian / Pacific Islander	11.11	9.78	2.06	8.83
	Black or African American	10.10	9.78	10.31	2.33
	Hispanic or Latino	7.07	1.09	3.09	0
	Mixed: White, Black, Caribbean	1.01	0	4.12	6.04
	Other ethnic group	N/A	N/A	N/A	1.40

[a] *N* = 101 participants completed Study 3, but *n* = 3 were excluded because demographic details at time 1 and 2 did not match, and a further person failed to provide demographic information.

2.4. *Procedure*

Participants took part in an online survey. They first provided informed consent and demographic information and were then presented with the instructions for the SST (see Fig. 1). Participants then viewed a string of digits (cognitive load – either 6 or 7; see Supplementary Materials) and subsequently completed either SST list 1 or 2. Following completion of the SST, participants were prompted to write down the string of digits.

2.5. *Results*

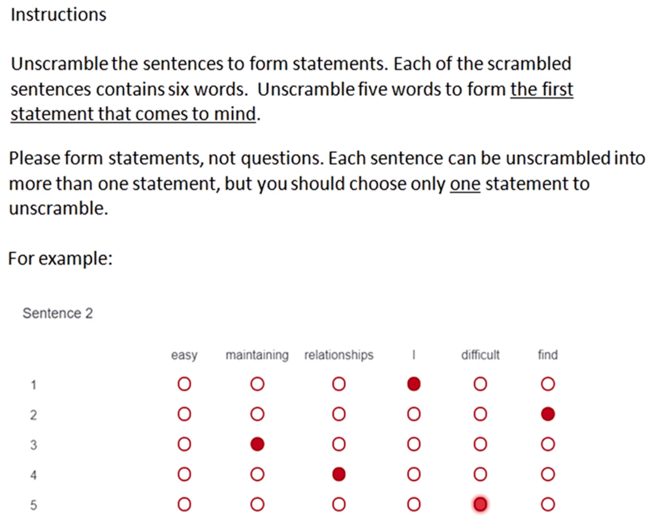
2.5.1. *Descriptive statistics and preliminary analyses*

Statistical analyses were conducted in Stata 13 (StataCorp, 2013). One participant was excluded from the analyses because they only formed one (out of 20) grammatically correct sentence. Thus, the final sample consisted of *N* = 98 participants. The mean number of grammatically incorrect or uncompleted sentences was *M* = .91 (*SD* = 1.39), indicating that very few items were not completed or completed in a grammatically incorrect manner. SST positivity indices (both lists) were slightly negatively skewed, as can be expected in a non-clinical sample. To account for the skewness in the analyses, we ran bootstrapped regression analyses (1000 replications), which do not place distributional assumptions on the data.

No differences on SST scores were found by item list, gender, age, or cognitive load (6 vs. 7 digits; see Supplementary Materials). Thus, we collapsed across load condition for the subsequent analyses. The standard load of six digits (see Wenzlaff & Bates, 1998) was used in Studies 2–4. Descriptive statistics for the SST and self-report questionnaires are presented in Table 2.

2.5.2. *Reliability*

Papers using the SST have generally focused on split-half reliability (e.g., O'Connor et al., 2021; Würtz et al., 2022) rather than reporting internal consistency such as Cronbach's α (though we have previously reported α on request; see Würtz et al., 2022). The use of Cronbach's α is controversial as it often underestimates test reliability (Dunn, Baguley, & Brunsden, 2014). Regarding the SST specifically, Cronbach's α may not be suitable as item coding is binary (zero or one); the Kuder-Richardson method may therefore be preferable. However, the SST also yields more 'missing' data than self-report questionnaires:



Instructions

Unscramble the sentences to form statements. Each of the scrambled sentences contains six words. Unscramble five words to form the first statement that comes to mind.

Please form statements, not questions. Each sentence can be unscrambled into more than one statement, but you should choose only one statement to unscramble.

For example:

Sentence 2

	easy	maintaining	relationships	I	difficult	find
1	○	○	○	●	○	○
2	○	○	○	○	○	●
3	○	●	○	○	○	○
4	○	○	●	○	○	○
5	○	○	○	○	●	○

Fig. 1. SST instructions and example worry SST item to illustrate how the scrambled sentences test was presented online in Studies 1–4. Participants were shown a neutral example from Wenzlaff and Bates (1998, 2000) as part of the instructions, but we here display one of our worry items from the main task.

Table 2
Descriptive statistics and correlations among measures in studies 1-4.

Study 1 (*N* = 98)	**Descriptive statistics**			**Correlations**										
	Measure	**Mean**	**SD**	SST	GAD-7	PSWQ								
	SST Worry	0.66	0.21	1										
	GAD-7	4.09	4.96	-.53**	1									
	PSWQ	46.08	16.96	-.57**	.69**	1								
Study 2 (*N* = 91)	**Descriptive statistics**			**Correlations**										
	Measure	**Mean**	**SD**	SST Total	SST Worry	SST Dep	GAD-7	PSWQ	PHQ-9	RRS				
	SST Total	0.73	0.23	1										
	SST Worry	0.70	0.24	.96**	1									
	SST Dep	0.77	0.25	.96**	.84**	1								
	GAD-7	4.32	4.77	-.62**	-.61**	-.59**	1							
	PSWQ	45.90	16.20	-.56**	-.57**	-.50**	.74**	1						
	PHQ-9	5.76	5.78	-.70**	-.68**	-.66**	.90**	.67**	1					
	RRS	40.10	15.42	-.73**	-.69**	-.71**	.76**	.65**	.80**	1				
Study 3 (*N* = 88)	**Descriptive statistics**					**Correlations**								
	Time 1 (*N* = 88)			**Time 2 (*N* = 53)**		**Time 1**								
	Measure	**Mean**	**SD**	**Mean**	**SD**	SST Total	SST Worry	SST Dep	GAD-7	PSWQ	PHQ-9	RRS		
	SST Total	0.72	0.29	0.73	0.31	1								
	SST Worry	0.70	0.27	0.71	0.31	.93**	1							
	SST Dep	0.73	0.34	0.75	0.34	.96**	.77**	1						
	GAD-7	4.98	6.15	4.55	5.99	-.66**	-.64**	-.61**	1					
	PSWQ	45.13	20.05	45.42	19.85	-.68**	-.64**	-.64**	.79**	1				
	PHQ-9	4.83	5.46	4.42	6.00	-.63**	-.56**	-.62**	.85**	.73**	1			
	RRS	42.22	17.11	40.45	18.00	-.64**	-.58**	-.62**	.83**	.78**	.80**	1		
Study 4 (*N* = 215)	**Descriptive statistics**							**Correlations**						
		GAD (*n* = 132)		**GAD + Dep (*n* = 83)**		**Total (*N* = 215)**		**Across groups**						
	Measure	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	SST Total	SST Worry	SST Dep	GAD-7	PSWQ	PHQ-9	RRS
	SST Total	0.47	0.20	0.35	0.20	0.42	0.21	1						
	SST Worry	0.42	0.21	0.34	0.21	0.39	0.22	.85**	1					
	SST Dep	0.51	0.25	0.35	0.24	0.45	0.26	.89**	.53**	1				
	GAD-7	15.01	3.38	16.25	3.36	15.49	3.42	-.22**	-.21**	-.19**	1			
	PSWQ	71.47	6.39	71.98	5.58	71.67	6.09	-.08	-.17*	-.01	.29**	1		
	PHQ-9	13.01	4.38	17.46	3.99	14.73	4.75	-.41**	-.34**	-.37**	.47**	.10	1	
	RRS	59.62	11.43	66.76	9.67	62.38	11.31	-.34**	-.23**	-.37**	.31**	.15*	.49**	1

Note. *Pearson correlation is significant at the 0.05 level (2-tailed); **Pearson correlation is significant at the 0.01 level (2-tailed). GAD = Generalised Anxiety Disorder; Dep = Depression; SST = Scrambled Sentences Test; GAD-7 = GAD-7 = Generalized Anxiety Disorder scale (measure of anxiety; PSWQ = Penn State Worry Questionnaire; SST Dep = Scrambled Sentences Test Depression items; PHQ-9 = Patient Health Questionnaire 9 (measure of depression); RRS = Ruminative Response Scale.

participants may construct grammatically incorrect sentences (cannot be scored) or may not finish all items (on timed tasks). While the number of grammatically incorrect or uncompleted items was very low in Studies 1–4 (see below), individual items may nevertheless have more 'missing' values than e.g., items on standard self-report questionnaires. Without imputing missing values, this can lead to difficulties computing internal consistency, especially if methods use casewise deletion. Therefore, as recommended by Parsons, Kruijt, and Fox (2019) for cognitive-behavioural measures, we focused on split-half reliability (Studies 1 and 4, i.e., initial and final measure) and test-retest reliability (Study 3 only) for the SST, though we report Cronbach's α for established self-report questionnaires for comparability with other research.

Split-half reliability (odd vs. even-numbered items) was high for both lists: Spearman-Brown Prophesy Reliability Estimate = .83 for list 1 and .85 for list 2. However, examining odd vs. even items is an arbitrary way of splitting the items. Therefore, we also calculated Guttman's λ4 measure of split-half reliability (Guttman's formula considers all possible splits to find the maximal λ4). While Cronbach's α can underestimate reliability, λ4 may overestimate reliability (Benton, 2015). We thus used Hunt and Bentler (2015)'s method, drawing a random series of locally optimal λ4 coefficients (1000 replications) and reporting the .5 quantile (median) alongside the maximal (1.0) λ4 value to provide a less upwardly biased estimate in our relatively small sample. Pairwise rather than casewise deletion was used to deal with missing values and maximise the amount of data included. Guttman's λ4 was excellent at .94 (.5 quantile = .90) for list 1 and .97 (.5 quantile = .93) for list 2.[4] Nevertheless, several items were dropped or replaced in Study 2 (see Supplementary Materials and below).

2.5.3. Construct validity

The SST positivity index was moderately negatively correlated with worry (PSWQ) and anxiety (GAD-7), indicating that a more positive interpretation bias was associated with lower worry and anxiety scores, as expected (see Table 2). Furthermore, we entered worry and anxiety as predictors into a regression model, with the SST positivity index as the outcome variable. As PSWQ and GAD-7 scores were significantly positively correlated (see Table 2), we mean-centred both variables to deal with multicollinearity issues. PSWQ score significantly predicted the SST positivity index ($b = -.005$, $SE = .002$, $p = .002$, 95% CIs [−.01; −.002]): higher PSWQ scores were associated with a lower SST positivity index, as expected. In addition, GAD-7 score also predicted the SST positivity index ($b = -.011$, $SE = .006$, $p = .044$, 95% CIs [−.02; −.0003]) in the same expected direction. The regression model accounted for 34.9% of the variance (adjusted R^2).

2.6. Discussion

Study 1 comprised the initial development and psychometric evaluation of a new scrambled sentences test for worry, consisting of two lists of items. Both SST lists had excellent split-half reliability. Moreover, higher levels of worry and anxiety were associated with a less positive interpretation bias.

3. Study 2: Further validation of the new worry SST and its specificity to worry and anxiety

The aim of Study 2 was to further revise and psychometrically evaluate the new wSST. We changed several items on the basis of the results of Study 1. Furthermore, to examine how specific the new SST was to worry and anxiety, we administered the wSST together with the established depression SST (dSST), and also assessed trait rumination and mood. We expected that a more positive interpretation bias as assessed by the new wSST would be more strongly associated with lower worry and anxiety than rumination and depression, and conversely, that a more positive interpretation bias as measured by the dSST would be more strongly related to lower rumination and depression symptoms than worry and anxiety. Although there was no ceiling effect in Study 1, we also generated and included neutral filler items (as previously used in some research on the SST; Everaert et al., 2014). These filler items were constructed to make statements that were not valenced (i.e., there was no ambiguity to resolve either positively or negatively). We included these items to make the overall task more opaque.

3.1. Scrambled sentences test (SST)

We revised and augmented the two item sets from Study 1 to each comprise 50 items: 20 worry-related items (adapted from Study 1; see below for changes), 20 depression-related items (taken from Wenzlaff & Bates, 1998, 2000) and 10 neutral filler items (created for the study; see Supplementary Materials).

3.1.1. Worry SST

Based on Study 1, we improved our item set by making the following changes: First, we replaced or amended items which inadvertently allowed both the positive and negative word to be used in the same sentence. For example, "*my goals will be achieved/unfulfilled*" could be arranged as "*unfulfilled goals will be achieved*" and was therefore changed. Second, we replaced or changed items which permitted more than one possible solution per valence. For example, we changed the item "*other people notice my faults/merits*", which could also be arranged as "*people notice my other faults/merits*". Third, we replaced an item which was consistently unscrambled in only a positive manner in Study 1. Fourth, we replaced an item for which a substantial proportion of participants did not make a grammatically correct sentence. Lastly, as participants might be influenced by how common and/or easy to read words are, we matched all disambiguating words for word length and frequency in the English language. The items used in Study 2 are presented in Supplementary Table 2.

3.1.2. Original depression SST (Wenzlaff & Bates, 1998, 2000)

The original scrambled sentences test includes 40 scrambled sentences relating to depressive and ruminative concepts, such as "*the future looks very bright/dismal*" or "*I am a worthwhile/worthless person*". All items were used and divided into two lists to be presented alongside the worry and filler items.

3.2. Measures of worry, anxiety, rumination, and depression

3.2.1. Trait worry and general anxiety

These constructs were assessed using the same measures as in Study 1 (PSWQ: α = .95; GAD-7: α = .93).

3.2.2. Ruminative response scale (RRS; Nolen-Hoeksema & Morrow, 1991)

The RRS is a 22-item measure of trait rumination, comprising items relating to the frequency of ruminative thoughts, such as, "*think about how you don't seem to feel anything anymore*", and rated on a scale from 1 (*almost never*) to 4 (*almost always*). Item responses are summed, with a higher total score denoting greater rumination (α = .96).

3.2.3. Patient health questionnaire 9 (PHQ-9; Kroenke & Spitzer, 2002)

The PHQ-9 measures symptoms of depression. It consists of nine items such as, "*Little interest or pleasure in doing things?*", which are rated as having occurred "*not at all*" (coded zero), "*several days*" (coded 1), "*more than half the days*" (coded 2), or "*nearly every day*" (coded 3) in the last two weeks. Responses are summed to produce a total score, with higher scores denoting greater depression severity (α = .90).

[4] One item was dropped due to too many missing values (matrix not positive definite) to compute Guttman's λ4.

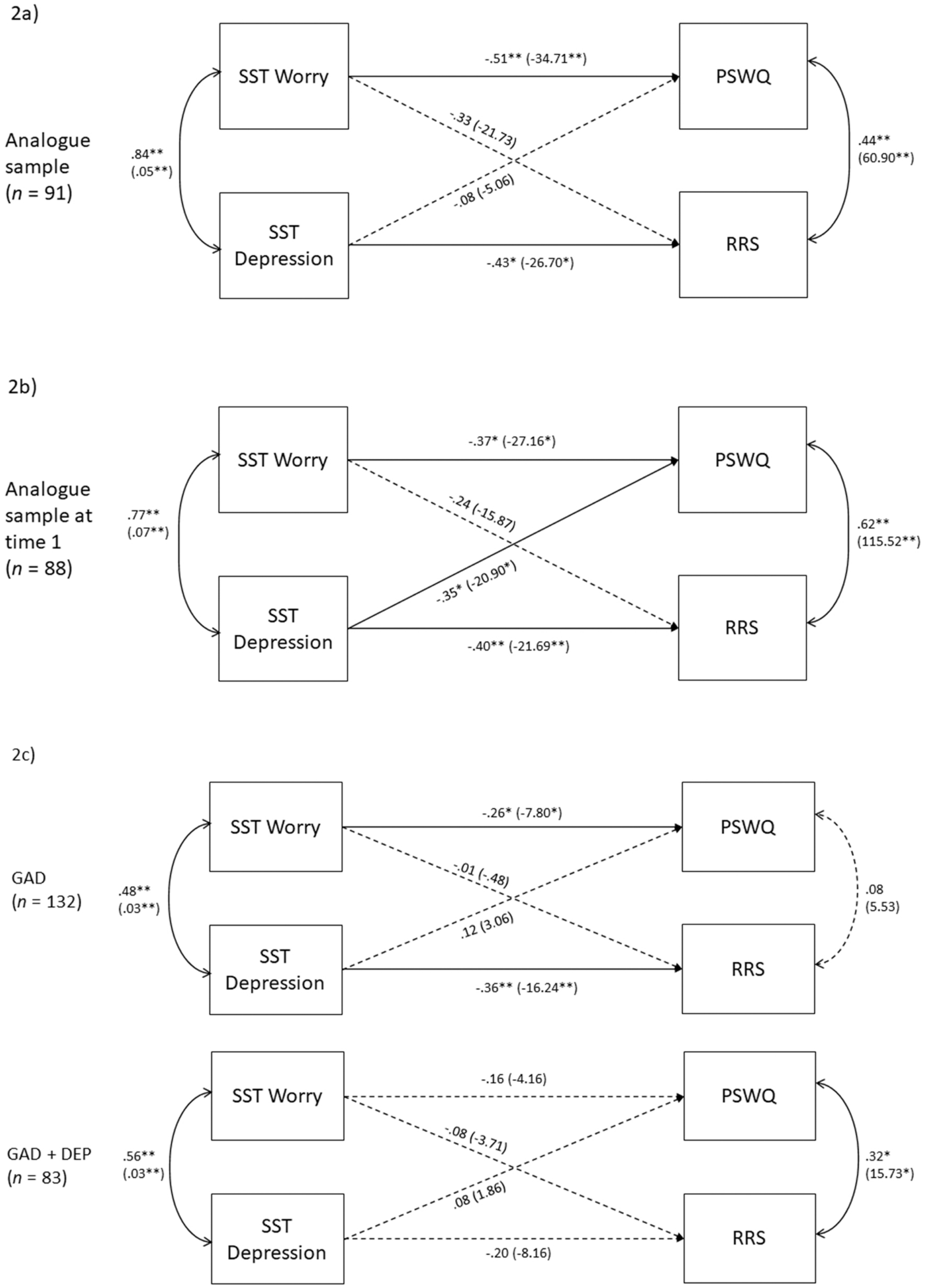


Fig. 2. Model results for Study 2 (2a), Study 3 (2b), and Study 4 (2c). Standardised β coefficients are shown, with unstandardised b coefficients in parentheses. Dashed lines indicate non-significant paths. Note. * indicates $p < .05$; ** indicates $p < .01$. SST = Scrambled Sentences Test; PSWQ = Penn State Worry Questionnaire; RRS = Ruminative Response Scale; GAD = Generalised Anxiety Disorder; DEP = Depression.

3.3. Participants

As in Study 1, participants were recruited via the online platform Amazon Mechanical Turk. A criterion was set to denote that participants who had completed Study 1 were not eligible to take part in this study. The final sample comprised $N = 92$ participants. All participants were resident in the USA and fluent in English (see Table 1).

3.4. Procedure

Participants provided informed consent and basic demographic information and were then presented with the SST instructions before being shown the string of six digits to hold in mind during the SST. Then, participants completed either SST list 1 or 2. Each list contained 20 worry items, 20 depression items, and 10 filler items (order randomised). After completing the SST, participants wrote down the string of digits.

3.5. Results

3.5.1. Descriptive statistics and preliminary analyses

Statistical analyses for this study and Studies 3 and 4 were conducted in Stata 16 (StataCorp, 2019). One participant was excluded from analyses because they did not form any grammatically correct sentences. Thus, the final sample consisted of $N = 91$ participants. The mean number of grammatically incorrect or uncompleted items was low ($M = 1.77$, $SD = 2.13$). Positivity indices did not vary by SST list, gender, or age (see Supplementary Materials for details); scores were collapsed across lists (see Table 2).

3.5.2. Reliability

In list 1, one worry item was negatively correlated with the other items (item-rest correlation $r = -.40$), while in list 2, six items were negatively correlated with the other items (item-rest correlations $r = -.11$ to $r = -.51$). Furthermore, on list 2, despite our efforts, one worry item was still solvable by using both the positive and negative words in a grammatically correct sentence, rendering it uncodeable. In addition, one depression item allowed possible variations in which the valence was less clear ("*most things make me happy / unhappy*" could be used to make "*things make me most happy*") and a further depression item had a large number of grammatically correct variations. In addition, one item in list 2 was accidentally presented twice. These issues were corrected and items dropped for Study 3, in which the total number of items was reduced (see below).

3.5.3. Construct validity and version specificity

Correlations between SST scores (overall SST, wSST, dSST) and questionnaire measures are presented in Table 2. As in Study 1, correlations were moderate to strong, with Pearson's r ranging from − .50 to − .73, indicating that a higher positivity index (i.e., a more positive interpretation bias) was associated with lower worry, anxiety, depression, and rumination scores.

To examine specificity, we constructed a path model with wSST and dSST scores as predictors and worry (PSWQ) and rumination (RRS) scores as outcomes, allowing correlations among SST versions and among symptom scores, and specifying all paths (i.e., running a fully saturated model). We estimated standard errors using bootstrapping (1000 replications). Model results are presented in Fig. 2. Notably, wSST predicted PSWQ but not RRS scores, while dSST predicted RRS but not PSWQ scores, indicating that both versions were specific to the constructs they were designed to capture.[5]

Table 3
Scrambled sentences test for worry – final item list.

List 1	List 2
I am improving/ruining my life	I find maintaining relationships easy/difficult
I am a pleasant/boring person	Others can see my merits/faults
I do/don't worry about money	I can/can't manage my finances
My living expenses are comfortable/tight	Everything will turn out fine/badly
I will/won't achieve my goals	I'm able/unable to support myself
I find the future exciting/scary	Approaching new people is fine/scary
I feel relaxed/tense around strangers	I am performing above/below expectations
I'm good/bad at making friends	My life will be fulfilling/unfulfilling
Finding a job is easy/hard	I will/won't earn enough money
I will/won't get into debt	I'm indifferent/worried about others' opinions

Note. Depression items not shown. Item order was randomised for presentation to participants.

3.6. Discussion

In Study 2, we further refined wSST items and explored the specificity of the new wSST. We found that both the new wSST and original dSST showed specificity. The wSST, designed to assess interpretations around worry content, was associated with trait worry, but not trait rumination scores. Conversely, the dSST, whose items relate to interpretations around depressed mood and rumination, was associated with trait rumination, but not worry. This finding indicates the appropriateness of using the SST version tailored to the thinking style or population being studied. When interested in worry, we demonstrate that using the new wSST is more relevant (both in face and incremental validity) than using the original dSST.

4. Study 3

In Study 3, we aimed to create the final wSST and assess its test-retest reliability. We reduced the item lists based on the findings of Study 2 and removed the filler items. We felt that filler items did not meaningfully influence the pattern of results and a shorter, more concise measure might be easier to administer (see Supplementary Materials for information on filler items). We created two item sets and tested the test-retest reliability of these sets by administering them two weeks apart. As a secondary aim, we again assessed convergent validity by including measures related to interpretation bias and broadened the scope to include measures capturing other cognitive processes related to worry (see Supplementary Materials).

4.1. Scrambled sentences test

We used two item sets of 20 items each, with each set comprising half worry-related and half depression-related items. The order of lists was counterbalanced across the two time points.

4.1.1. Worry SST items

Two lists of 10 items each comprised the final item sets. The final items in each list are presented in Table 3.

4.1.2. Depression SST items

As in Studies 1 and 2, we presented 10 depression items, taken from

[5] We repeated this analysis adding gender as a grouping variable. Constraining all parameters to be equal, model fit was excellent (CFI = .978, RMSEA = .064, SRMR = .164) indicating that results were unlikely to be moderated by gender.

Wenzlaff and Bates (1998, 2000).

4.2. Self-report questionnaires

Participants completed the PSWQ, RRS, GAD-7, and PHQ-9. Additionally, they completed the Attentional Control Questionnaire (Derryberry & Reed, 2002) and the Fatigue Severity Scale (Krupp, LaRocca, Muir-Nash, & Steinberg, 1989); see Supplementary Materials.

4.3. Participants

Participants were again recruited from Amazon Mechanical Turk. Participants who had completed Study 1 or Study 2 were not eligible to take part in this study. All participants were invited to complete the survey at two time points, two weeks apart. $N = 101$ participants completed the survey at time 1. Three participants were excluded because demographic data did not match ($n = 3$; e.g., male at time 1, female at time 2). All participants were resident in the USA and fluent in English (see Table 1 for demographic information at time 1).

4.4. Procedure

Participants provided informed consent, basic demographic information, and completed the self-report questionnaires. Then, they completed either SST list 1 or 2 (order counterbalanced across participants and time points). To keep timings equal across sessions, participants were given five minutes in which to unscramble as many sentences as possible, after which time the page moved on automatically. Participants were contacted again two weeks after they had initially completed the study. Thosewho participated at the second time point completed the SST list they had not seen at time 1.

4.5. Results

4.5.1. Descriptive statistics and preliminary analyses

Ten participants were excluded because they completed fewer than half of SST items at either/both sessions ($n = 10$). Thus, the final sample comprised $N = 88$ participants, who were included in analyses. Of these 88 participants, $n = 53$ completed the survey at both time points and comprised the test-retest sample.

The mean number of SST items that were either grammatically incorrect or uncompleted was again very low across lists (time 1: $M = 1.38$, $SD = 1.98$; time 2: $M = 1.30$, $SD = 2.04$; see Supplementary Materials for more information). Positivity scores were again unrelated to SST list and gender, but were associated with age, and there were no differences in demographic or questionnaire data or in SST scores between participants who only completed time 1 and those who completed both time points (see Supplementary Materials). Descriptive statistics are presented in Table 2.

4.5.2. Reliability

Test-retest reliability was evaluated in the $n = 53$ participants who completed the SST at both time points. Simple correlations between SST scores at time 1 and time 2 were $r = .92$ ($p < .05$) for the overall SST, $r = .82$ ($p < .05$) for the wSST, and $r = .86$ ($p < .05$) for the dSST. We computed mean intraclass correlation coefficients (ICCs[6]) in a two-way random-effects model (where "time" and "participant" were the factors). For the overall SST (both worry and depression items), the average ICC was .956 (95% confidence intervals .923 to .974), indicating excellent test-retest reliability. For the wSST (worry items only), the average ICC was .893 (95% confidence intervals .814 to .938) and for the dSST, the average ICC was .927 (95% confidence intervals .874 to .958). Thus, the new wSST showed excellent test-retest reliability, both on its own and in combination with the dSST.

4.5.3. Construct validity

Correlations between SST indices (overall SST, wSST, and dSST) and questionnaire measures at time 1 are presented in Table 2. Significant ($ps < .01$) negative moderate-to-strong correlations (Pearson's rs ranging from −.56 to −.68) were found between all three SST indices and all questionnaires, indicating that a more positive interpretation bias was related to lower symptom scores.

As in Study 2, we again explored the specificity of the new wSST at time 1 (see Supplementary Materials for model specifications). The wSST predicted PSWQ but not RRS scores, while dSST this time predicted both RRS and PSWQ scores (see Fig. 2). Thus, there was still specificity in this sample, though results were not quite as strong as in Study 2. We examined specificity further in a clinical sample (see Study 4).

4.6. Discussion

The final worry SST item set showed excellent test-retest reliability and convergent validity. As expected, we found that interpretation bias as measured by the new wSST was moderately to strongly associated with measures of worry and anxiety.

5. Study 4: validation of the worry SST in a sample of individuals with GAD

This final study sought to psychometrically evaluate the wSST in a clinical sample of individuals with GAD. Participants completed the wSST, dSST, PSWQ, RRS, GAD-7, and PHQ-9 self-report measures (see Study 3 for details).

5.1. Participants

The sample was drawn from a large-scale randomised controlled trial (Hirsch et al., 2021). Baseline (i.e., pre-intervention) data was obtained from $N = 221$ participants who met diagnostic criteria for GAD (with or without co-morbid depression) on the Structured Clinical Interview for DSM-5 Axis 1 Disorders (First, Williams, Karg, & Spitzer, 2015). Six participants were excluded due to missing or unusable data, leaving $N = 215$. Of these participants, $n = 132$ had a diagnosis of GAD, while $n = 83$ participants had a diagnosis of GAD with co-morbid depression. Inclusion and exclusion criteria for the trial are detailed in Hirsch et al. (2021) but included meeting diagnostic criteria for GAD, UK residency, fluency in English, being 18–65 years old, experiencing clinical levels of worry (i.e., total score of ≥ 62 on PSWQ) and clinical levels of anxiety (i.e., total score ≥ 10 on the GAD-7). Apart from their clinical status, notable differences to Studies 1–3 included country of residence (UK vs. USA) and gender (more female participants; see Table 1).

5.2. Procedure

Participants completed the SST and symptom measures as part of a baseline battery of measures administered prior to a cognitive-bias-modification intervention (Hirsch et al., 2021). They first completed demographic information, then the questionnaire measures and then the SST; time to complete the SST was constrained to five minutes to match Study 3.

5.3. Results

5.3.1. Descriptive statistics and preliminary analyses

The mean number of grammatically incorrect or uncompleted sentences was low across lists ($M = 1.77$, $SD = 2.35$; see Supplementary

[6] Values below 0.40 are considered poor reliability, values between 0.40 and 0.59 as fair reliability, 0.60–0.74 as good reliability, and values above 0.75 as excellent reliability (Fleis, Levin, & Paik, 2003).

Materials for information by list). Neither gender ($t(213) = -1.78$, $p = .077$, Hedges' $g = -.41$) nor age ($r = -.08$, *n.s.*) were associated with SST positivity index. However, SST positivity index varied by list, $t(213) = -4.40$, $p < .001$, Hedges' $g = -.60$ ($M = .39$, SD $= .19$ for list 1, $M = .48$, SD $= .21$ for list 2); thus, we controlled for list (see Section 5.3.3).

The two groups (i.e., GAD vs. GAD with comorbid depression) differed significantly in terms of SST scores, $t(213) = 4.21$, $p < .001$, Hedges' $g = -.59$: participants with GAD and depression displayed a less positive interpretation bias than participants with GAD only on the combined SST index (GAD group*:* $M = .47$, $SD = .20$; GAD with depression group: $M = .35$, $SD = .20$), as well as the separate wSST and dSST indices (see Supplementary Materials). SST scores in the clinical groups (see Table 2) were lower than those in the previous analogue samples (Studies 1–3), indicating that this clinical sample had a less positive interpretation bias.

5.3.2. *Reliability*

Guttman's λ4 measure of split-half reliability was .90 (.5 quantile = .85) and .89 (.5 quantile = .86) for the overall (worry plus depression items) lists 1 and 2, respectively. For the wSST, Guttman's λ4 = .76 (.5 quantile = .72) for list 1 and .73 (.5 quantile = .69) for list 2. For the dSST, Guttman's λ4 = .79 (.5 quantile = .73) for list 1 and .82 (.5 quantile = .76) for list 2. Thus, Guttman's λ4 indicated good to excellent split-half reliability.

5.3.3. *Construct validity and version specificity*

All SST indices were significantly negatively correlated with GAD-7, PHQ-9, and RRS scores, but interestingly, only the wSST score was significantly associated with the PSWQ score (see Supplementary Materials for correlations separately for each SST list). To examine version specificity, we again specified a path model (see Studies 2 and 3 for details). In this unconstrained model, we added group (GAD, GAD with depression) as a grouping variable and tested whether paths differed between groups (controlling for SST list). Results in the GAD group replicated Study 2: wSST predicted PSWQ, but not RRS scores, while dSST predicted RRS, but not PSWQ scores, as expected. In the GAD with depression group, the paths were non-significant (see Fig. 2). None of the paths differed significantly between groups (see Supplementary Materials, also for sensitivity analyses regarding gender; Supplementary Fig. 1).

5.4. *Discussion*

The aim of Study 4 was to further validate the new wSST in a clinical sample of individuals with GAD (either with or without comorbid depression). We replicated and extended findings from three separate analogue samples in this large clinical study to demonstrate the utility of the new wSST in worry and anxiety research. The new wSST was significantly correlated with symptom scores, supporting its construct validity. We again showed specificity for the separate wSST and dSST versions in the GAD only group: only the wSST was significantly associated with trait worry, while only the dSST was significantly associated with trait rumination (in an analysis controlling for the relative impact of the other version). This finding supports the utility of the new wSST measure in individuals with GAD and suggests that the dSST may be most useful when studying interpretation bias in relation to high levels of rumination and depression. It should be noted that we did not find specificity in the GAD with comorbid depression group, in which repetitive negative thinking may be more difficult to disentangle; here, the combined (worry plus depression items) SST may be most suitable.

Lastly, we found that individuals with both GAD and depression displayed a less positive interpretation bias than individuals with GAD only on all three SST indices. This group also had higher symptom scores, supporting the correlation between symptom severity and degree of interpretation bias (Krahé et al., 2019).

6. General discussion

The key aims of this work were to examine the reliability, validity, and specificity of a new scrambled sentences test version for worry (wSST). Overall, the wSST showed excellent test-retest reliability and good-to-excellent split-half reliability. Importantly, in line with theory (e.g., Hirsch & Mathews, 2012) and empirical evidence from other interpretation bias tasks (e.g., Butler & Mathews, 1983; Eysenck, Mogg, May, Richards, & Mathews, 1991), the measure had good construct validity: higher levels of worry and anxiety were associated with a less positive interpretation bias. Furthermore, the wSST was associated with trait worry, but not trait rumination scores, indicating that the wSST has good specificity. Taken together, our findings support the use of the new wSST as a reliable and valid measure of negative interpretation bias in the general population and in individuals where high levels of worry and anxiety are the primary concern (e.g., GAD).

While it can be used in cross-sectional and longitudinal research, the wSST is also envisaged to be vitally useful in intervention research. The causal relationship between interpretation bias and worry is well established in single-session studies (e.g., Feng et al., 2020; Hayes, Hirsch, & Mathews, 2010; Hirsch et al., 2009). Furthermore, interpretation bias training delivered using a multi-training session format over a number of weeks has been shown to have an impact on symptoms of anxiety and depression (Hirsch et al., 2021; Nieto & Vazquez, 2021; see also Fodor et al., 2020, for a meta-analysis). In the present studies, we developed two lists of SST items for use in pre-post intervention designs. The order of these lists can be counter-balanced across participants, avoiding participants having to complete the same items at both time points (which could lead to memory or practice effects). In this vein, we showed that there is no ceiling effect in the SST worry version (see Studies 1 and 2), though note that participants completed the task under cognitive load in all studies. Furthermore, based on findings from Study 3, we can be confident that *both* the dSST and wSST versions have excellent test-retest reliability (though this should be replicated in clinical samples), which is essential if these measures are used in intervention studies where the key aim is to modify interpretation bias. In such intervention studies, calculating standard error of measurement and minimal detectable differences (see Portney & Watkins, 2009) will also be highly useful for evaluating whether changes in SST scores reflect true change rather than random measurement error.

There are some limitations to this work. The present version of the SST requires those who complete it to have good knowledge of the English language and relies on participants forming grammatically correct sentences. We did not control for levels of literacy among participants or other possible confounding variables, such as working memory capacity and performance. This would be useful to include in future studies, especially as anxiety is related to poorer working memory capacity (Moran, 2016), and repetitive negative thinking is associated with difficulties deleting no longer relevant information from working memory (Zetsche, Bürkner, & Schulze, 2018). Furthermore, the new wSST measure may be influenced by socioeconomic factors in that financial worries may be more or less pertinent depending on individuals' circumstances. We tried to address this issue by drawing on a broad range of common worry domains in constructing the items. Nevertheless, measuring socioeconomic status and/or individuals' worry content could be useful in future research. Importantly, our samples were not ethnically diverse with 70–81% of our samples identifying as white. Evidence suggests that response patterns on measures of anxiety differ between different ethnic groups (Hambrick et al., 2010). The use of anxiety measures may therefore risk presenting biased results when used with ethnically diverse samples. This is because validation studies for anxiety measures have not traditionally included participants from a range of different ethnic groups. Thus, future research should examine the utility of the SST measures with more ethnically diverse samples. Lastly, we recruited participants from Amazon Mechanical Turk in Studies 1–3. While mixed views exist on obtaining data from this platform (Buhrmester, Kwang, & Gosling, 2011; Chmielewski & Kucker, 2020), we validated and tested the final measure in a large sample of participants with GAD (Study 4), who were not recruited using Mechanical Turk.

Overall, our findings support the utility of the worry SST as a reliable and valid measure of interpretation bias. This measure provides a distinct approach to measuring interpretation bias beyond the widely used recognition test, as it relies on faster and more automatic information processing, especially when completed under a cognitive load (Würtz et al., 2022). Contrary to interpretation bias measures without time limits or cognitive load, the SST assesses interpretations made without prolonged time for reflection, providing a more direct measure of implicit or latent biases. Moreover, compared to measures which ask participants to rank order different interpretations in relation to the extent with which they would come to mind (e.g., Amir, Foa, & Coles, 1998), a strength of the SST is that the task is oblique (different interpretations are not presented), and so it is less affected by demand characteristics. Thus, the SST provides a highly suitable way of assessing interpretation bias. We hope that in developing and psychometrically evaluating new worry SST items and in further validating established depression SST items, this paper contributes to ensuring methodological rigour in interpretation bias research, as well as providing further evidence for the link between worry, anxiety, and interpretation bias.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We are grateful to Jessica Whyte, Emese Nagy, Sabriye Ana Velizarova, and Weike Xia for their help with these studies. We also thank our experts by experience for valuable discussions around worry and anxiety, which strengthened the research. This work was supported by an MQ: Transforming Mental Health PsyIMPACT grant (number MQ14PP_84) to CH. CH received salary support from the National Institute for Health Research (NIHR). The views expressed are those of the authors and not necessarily those of MQ: Transforming Mental Health, the NHS, the National Institute for Health Research (NIHR), or the Department of Health.

Appendix A. Supplementary materials

Supplementary information associated with this article can be found in the online version at doi:10.1016/j.janxdis.2022.102610.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.

Amir, N., Foa, E. B., & Coles, M. E. (1998). Negative interpretation bias in social phobia. *Behaviour Research and Therapy, 36*(10), 945–957.

Badra, M., Schulze, L., Becker, E. S., Vrijsen, J. N., Renneberg, B., & Zetsche, U. (2017). The association between ruminative thinking and negative interpretation bias in social anxiety. *Cognition and Emotion, 31*(6), 1234–1242.

Benton, T. (2015). An empirical assessment of Guttman's Lambda 4 reliability coefficient. *Quantitative psychology pesearch* (pp. 301–310). Cham: Springer.

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*, 69–73.

Brown, T. A., Antony, M. M., & Barlow, D. H. (1992). Psychometric properties of the Penn State Worry Questionnaire in a clinical anxiety disorders sample. *Behaviour Research and Therapy, 30*(1), 33–37.

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.

Butler, G., & Mathews, A. (1983). Cognitive processes in anxiety. *Advances in Behaviour Research and Therapy, 5*(1), 51–62.

Chen, J., Short, M., & Kemps, E. (2020). Interpretation bias in social anxiety: A systematic review and meta-analysis. *Journal of Affective Disorders, 276*, 1119–1130.

Chmielewski, M., & Kucker, S. C. (2020). An MTurk crisis? Shifts in data quality and the impact on study results. *Social Psychological and Personality Science, 11*(4), 464–473.

Derryberry, D., & Reed, M. A. (2002). Anxiety-related attentional biases and their regulation by attentional control. *Journal of Abnormal Psychology, 111*(2), 225–236.

Dunn, T. J., Baguley, T., & Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology, 105*(3), 399–412.

Ehring, T., & Watkins, E. R. (2008). Repetitive negative thinking as a transdiagnostic process. *International Journal of Cognitive Therapy, 1*(3), 192–205.

Everaert, J., Duyck, W., & Koster, E. H. (2014). Attention, interpretation, and memory biases in subclinical depression: A proof-of-principle test of the combined cognitive biases hypothesis. *Emotion, 14*(2), 331–340.

Everaert, J., Podina, I. R., & Koster, E. H. W. (2017). A comprehensive meta-analysis of interpretation biases in depression. *Clinical Psychology Review, 58*, 33–48.

Everaert, J., Tierens, M., Uzieblo, K., & Koster, E. H. (2013). The indirect effect of attention bias on memory via interpretation bias: Evidence for the combined cognitive bias hypothesis in subclinical depression. *Cognition & Emotion, 27*(8), 1450–1459.

Eysenck, M. W., Mogg, K., May, J., Richards, A., & Mathews, A. (1991). Bias in interpretation of ambiguous sentences related to threat in anxiety. *Journal of Abnormal Psychology, 100*(2), 144–150.

Feng, Y.-C., Krahé, C., Meeten, F., Sumich, A., Mok, C. L. M., & Hirsch, C. R. (2020). Impact of imagery-enhanced interpretation training on offline and online interpretations in worry. *Behaviour Research and Therapy, 124*, Article 103497.

First, M. B., Williams, J. B. W., Karg, R. S., & Spitzer, R. L. (2015). *Structured clinical interview for DSM-5—research version (SCID-5 for DSM-5, research version; SCID-5-RV)*. Arlington, VA: American Psychiatric Association.

Fleis, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions*. Hoboken, NJ: John Wiley & Sons.

Fodor, L. A., Georgescu, R., Cuijpers, P., Szamoskozi, Ş., David, D., Furukawa, T. A., & Cristea, I. A. (2020). Efficacy of cognitive bias modification interventions in anxiety and depressive disorders: a systematic review and network meta-analysis. *The Lancet Psychiatry, 7*(6), 506–514.

Hambrick, J. P., Rodebaugh, T. L., Balsis, S., Woods, C. M., Mendez, J. L., & Heimberg, R. G. (2010). Cross-ethnic measurement equivalence of measures of depression, social anxiety, and worry. *Assessment, 17*(2), 155–171.

Hayes, S., Hirsch, C. R., & Mathews, A. (2010). Facilitating a benign attentional bias reduces negative thought intrusions. *Journal of Abnormal Psychology, 119*(1), 235–240.

Hirsch, C. R., Hayes, S., & Mathews, A. (2009). Looking on the bright side: Accessing benign meanings reduces worry. *Journal of Abnormal Psychology, 118*(1), 44–54.

Hirsch, C. R., Krahé, C., Whyte, J., Bridge, L., Loizou, S., Norton, S., & Mathews, A. (2020). Effects of modifying interpretation bias on transdiagnostic repetitive negative thinking. *Journal of Consulting and Clinical Psychology, 88*(3), 226–239.

Hirsch, C. R., Krahé, C., Whyte, J., Krzyzanowski, H., Meeten, F., Norton, S., & Mathews, A. (2021). Internet-delivered interpretation training reduces worry and anxiety in individuals with generalized anxiety disorder: A randomized controlled experiment. *Journal of Consulting and Clinical Psychology, 89*(7), 575–589.

Hirsch, C. R., Krahé, C., Whyte, J., Loizou, S., Bridge, L., Norton, S., & Mathews, A. (2018). Interpretation training to target repetitive negative thinking in generalized anxiety disorder and depression. *Journal of Consulting and Clinical Psychology, 86*(12), 1017–1030.

Hirsch, C. R., & Mathews, A. (2012). A cognitive model of pathological worry. *Behaviour Research and Therapy, 50*(10), 636–646.

Hirsch, C. R., Mathews, A., Lequertier, B., Perman, G., & Hayes, S. (2013). Characteristics of worry in generalized anxiety disorder. *Journal of Behavior Therapy and Experimental Psychiatry, 44*(4), 388–395.

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.

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

Hunt, T. D., & Bentler, P. M. (2015). Quantile lower bounds to reliability based on locally optimal splits. *Psychometrika, 80*(1), 182–195.

Krahé, C., Whyte, J., Bridge, L., Loizou, S., & Hirsch, C. R. (2019). Are different forms of repetitive negative thinking associated with interpretation bias in generalized anxiety disorder and depression? *Clinical Psychological Science, 7*(5), 969–981.

Kroenke, K., & Spitzer, R. L. (2002). The PHQ-9: A new depression diagnostic and severity measure. *Psychiatric Annals, 32*(9), 509–515.

Krupp, L. B., LaRocca, N. G., Muir-Nash, J., & Steinberg, A. D. (1989). The fatigue severity scale: Application to patients with multiple sclerosis and systemic lupus erythematosus. *Archives of Neurology, 46*(10), 1121–1123.

Lamers, F., van Oppen, P., Comijs, H. C., Smit, J. H., Spinhoven, P., van Balkom, A. J., & Penninx, B. W. (2011). Comorbidity patterns of anxiety and depressive disorders in a large cohort study: The Netherlands Study of Depression and Anxiety (NESDA). *Journal of Clinical Psychiatry, 72*(3), 341–348.

Mathews, A., & MacLeod, C. (2005). Cognitive vulnerability to emotional disorders. *Annual Review in Clinical Psychology, 1*, 167–195.

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the Penn State Worry Questionnaire. *Behaviour Research and Therapy, 28*(6), 487–495.

Mobach, L., Rinck, M., Becker, E. S., Hudson, J. L., & Klein, A. M. (2019). Content-specific interpretation bias in children with varying levels of anxiety: The role of gender and age. *Child Psychiatry and Human Development, 50*(5), 803–814.

Moran, T. P. (2016). Anxiety and working memory capacity: A meta-analysis and narrative review. *Psychological Bulletin, 142*(8), 831.

Nieto, I., & Vazquez, C. (2021). Disentangling the mediating role of modifying interpretation bias on emotional distress using a novel cognitive bias modification program. *Journal of Anxiety Disorders, 83*, Article 102459.

Nolen-Hoeksema, S., & Morrow, J. (1991). A prospective study of depression and posttraumatic stress symptoms after a natural disaster: The 1989 Loma Prieta Earthquake. *Journal of Personality and Social Psychology, 61*(1), 115–121.

O'Connor, C. E., Everaert, J., & Fitzgerald, A. (2021). Interpreting ambiguous emotional information: Convergence among interpretation bias measures and unique relations with depression severity. *Journal of Clinical Psychology, 77*(11), 2529–2544.

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science, 2*(4), 378–395.

Portney, L. G., & Watkins, M. P. (2009). *Foundations of clinical research: Applications to practice*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Rude, S. S., Durham-Fowler, J. A., Baum, E. S., Rooney, S. B., & Maestas, K. L. (2010). Self-report and cognitive processing measures of depressive thinking predict subsequent major depressive disorder. *Cognitive Therapy and Research, 34*(2), 107–115.

Rude, S. S., Valdez, C. R., Odom, S., & Ebrahimi, A. (2003). Negative cognitive biases predict subsequent depression. *Cognitive Therapy and Research, 27*(4), 415–429.

Rude, S. S., Wenzlaff, R. M., Gibbs, B., Vane, J., & Whitney, T. (2002). Negative processing biases predict subsequent depressive symptoms. *Cognition & Emotion, 16*(3), 423–440.

Savulich, G., Shergill, S. S., & Yiend, J. (2017). Interpretation biases in clinical paranoia. *Clinical Psychological Science, 5*(6), 985–1000.

Schoth, D. E., & Liossi, C. (2017). A systematic review of experimental paradigms for exploring biased interpretation of ambiguous information with emotional and neutral associations. *Frontiers in Psychology, 8*, 171.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine, 166*(10), 1092–1097.

Startup, H. M., & Erickson, T. M. (2006). The Penn State Worry Questionnaire (PSWQ). In G. C. L. Davey, & A. Wells (Eds.), *Worry and Its Psychological Disorders: Theory, Assessment and Treatment* (pp. 101–119). Chichester: Wiley.

StataCorp. (2013). *Stata statistical software: release 13*. College Station, TX: StataCorp LP.

StataCorp. (2019). *Stata statistical software: release 16*. College Station, TX: StataCorp LLC.

Stoeber, J., & Bittencourt, J. (1998). Weekly assessment of worry: an adaptation of the Penn State Worry Questionnaire for monitoring changes during treatment. *Behaviour Research and Therapy, 36*(6), 645–656.

Tallis, F., Davey, G. C., & Bond, A. (1994). *The worry domains questionnaire*. John Wiley & Sons.

Trotta, A., Kang, J., Stahl, D., & Yiend, J. (2020). Interpretation bias in paranoia: A systematic review and meta-analysis. *Clinical Psychological Science, 9*(1), 3–23.

Viviani, R., Dommes, L., Bosch, J. E., Stingl, J. C., & Beschoner, P. (2018). A computerized version of the scrambled sentences test. *Frontiers in Psychology, 8*, 2310.

Watkins, E., Moulds, M., & Mackintosh, B. (2005). Comparisons between rumination and worry in a non-clinical population. *Behaviour Research and Therapy, 43*(12), 1577–1585.

Wenzlaff, R. M., & Bates, D. E. (1998). Unmasking a cognitive vulnerability to depression: How lapses in mental control reveal depressive thinking. *Journal of Personality and Social Psychology, 75*(6), 1559–1571.

Wenzlaff, R. M., & Bates, D. E. (2000). The relative efficacy of concentration and suppression strategies of mental control. *Personality and Social Psychology Bulletin, 26*(10), 1200–1212.

Würtz, F., Zahler, L., Blackwell, S. E., Margraf, J., Bagheri, M., & Woud, M. L. (2022). Scrambled but valid? The scrambled sentences task as a measure of interpretation biases in psychopathology: A systematic review and meta-analysis. *Clinical Psychology Review, 93*, Article 102133.

Yu, M., Westenberg, P. M., Li, W., Wang, J., & Miers, A. C. (2019). Cultural evidence for interpretation bias as a feature of social anxiety in Chinese adolescents. *Anxiety, Stress, & Coping, 32*(4), 376–386.

Zetsche, U., Bürkner, P.-C., & Schulze, L. (2018). Shedding light on the association between repetitive negative thinking and deficits in cognitive control – A meta-analysis. *Clinical Psychology Review, 63*, 56–65.

Zlomke, K. R. (2009). Psychometric properties of internet administered versions of Penn State Worry Questionnaire (PSWQ) and Depression, Anxiety, and Stress Scale (DASS). *Computers in Human Behavior, 25*(4), 841–843.