Interpreting ambiguous emotional information: Convergence among interpretation bias measures and unique relations with depression severity

Cliodhna E. O’Connor¹ *, Jonas Everaert² *, & Amanda Fitzgerald³

¹ National Suicide Research Foundation, University College Cork, Ireland
² Department of Experimental Clinical and Health Psychology, Ghent University, Belgium
³ School of Psychology, University College Dublin, Ireland

* These authors contributed equally to this work.

Corresponding author: Cliodhna O’Connor, Cliodhna.oconnor@ucc.ie
Abstract

Objectives: This study aimed to examine the convergence among interpretation bias measures and their associations with depressive symptom severity. Research into interpretation biases employs measures of interpretation bias interchangeably, however, little is known about the relationship between these measures.

Method: Participants (N=82 unselected undergraduate students; 59 female) completed four computer-based interpretation bias tasks in a cross-sectional design study.

Results: Indirect measures, based on participants’ reaction times, were not correlated to each other and had poor split-half reliability. Direct measures were more strongly correlated with depressive symptoms than indirect measures, but only the Scrambled Sentences Task explained a reliable unique portion of variance in depressive symptoms.

Conclusions: Interpretation bias tasks may not measure the same cognitive process and may differ in the extent to which they are a cognitive marker of depression-linked interpretation bias. These findings help to improve the measurement of and theory underlying interpretation bias and depressive symptoms.

Keywords: interpretation bias, cognitive bias, depression, depressive symptoms
Introduction

Interpretation is a cognitive process central to resolving the ambiguous nature of daily life situations. For example, imagining an audience laughing at a speech, one person might interpret that as reassurance that they are engaging and entertaining, while another person might interpret that as being laughed at. The interpretation of such ambiguous situations likely shapes how people feel, think, and act. For instance, the person who interprets the aforementioned situation as being laughed at might experience feelings of embarrassment and shame, they might engage in negative self-talk ruminating on the situation and may be less likely to engage in public speaking in the future. Cognitive theories of depression posit that negatively biased processing in the interpretation of ambiguous situations may be a causal and maintaining factor in the emotional disturbances linked with depression (Clark et al., 1999; Ingram et al., 1999). It is hypothesised that individuals with more severe depressive symptoms have a tendency to infer more negative and fewer positive interpretations to explain ambiguous emotional information (Clark et al., 1999; Ingram et al., 1999; Wisco, 2009). Such interpretation biases are commonly regarded as proximal cognitive causes of depression and represent a target in cognitive-behavioral therapies (Clark et al., 1999; Derubeis et al., 2009) and cognitive training programs for depression (Fodor et al., 2020).

Although cognitive theory suggests a link between interpretation bias and depression, some reviews of the literature have concluded that empirical research has observed inconsistent findings in supporting this theory (Blanchette & Richards, 2010; Gotlib & Joormann, 2010). In an attempt to synthesise the available empirical evidence to support the cognitive theory that interpretation bias may contribute to depression, a comprehensive meta-analysis integrated 87 independent studies to estimate the overall effect size and identify key moderators (Everaert et al., 2017). The meta-analysis reported a medium overall effect size supporting the presence of the theorised interpretation biases in relation to depression. Specifically, this meta-analysis
reported that effect sizes for depression-linked interpretation biases are larger for studies using direct as opposed to indirect measures of interpretation bias (Everaert et al., 2017).

Direct vs indirect refers to the specific measurement procedures to capture interpretation bias within an experimental paradigm (Everaert et al., 2017). Direct measures require participants to report or endorse the emotional tone of their interpretations of ambiguous stimuli, whereas indirect measures involve behavioural (e.g., reaction times) and/or psychophysiological (e.g., startle reflexes) responses to emotional interpretations (Everaert et al., 2017). Whilst direct measures have higher face validity than indirect measures (i.e., direct measures capture the content of interpretations), indirect measures are less subject to response bias effects because they do not allow participants time to reflect on the material before being required to respond (Blanchette & Richards, 2010; Hirsch et al., 2016; Schoth & Liossi, 2017). Consequently, it is possible that methodological factors, such as the measure of interpretation bias being used, contribute in part to the inconsistent evidence base for interpretation biases in depression.

Within the context of the experimental paradigms, direct and indirect measurement procedures of interpretation bias may represent different theoretical constructs related to cognitive processing (Everaert, 2021; Hirsch et al., 2016). These theoretical constructs may relate to sub-processes in interpretation bias, specifically strategic cognitive processing and automatic cognitive processing. For direct measures, respondents report the content or the tone of their interpretation and for indirect measures, the content or tone of interpretations is captured by inferring it from behavioral performance. Given that direct measures involve self-report or endorsement of emotional tone within experimental conditions that typically allow the participant time to reflect on their response (Hirsch et al., 2016), this may characterize strategic cognitive processing and this may be a sub-process of interpretation bias (Everaert et al., 2017; Hirsch et al., 2016). On the other hand, indirect measures as typically employed within
experimental paradigms that measure behavioral or psychophysiological responses to emotional interpretations may represent more automatic processing (Bisson & Sears, 2007; Lawson et al., 2002; Teachman et al., 2012), and this may be another sub-process of interpretation bias (Everaert et al., 2017; Hirsch et al., 2016). Therefore, while inconsistent findings for the relationship between interpretation bias and depression may be a result of methodological factors, it is also possible that the theoretical constructs of strategic and automatic processing underlying these methods may be a factor in inconsistent findings.

A wide variety of cognitive tasks have been utilized in research to investigate the association between depression and biased interpretation of ambiguous emotional information (for reviews, see Blanchette & Richards, 2010; Everaert et al., 2017; Stuijfzand et al., 2017; Wisco, 2009). One of the most frequently used interpretation bias tasks is the Scrambled Sentences Test (SST; Wenzlaff, 1998). In the SST, participants are presented with scrambled sentences of six words (e.g., am loser a I winner born) which they need to unscramble using only five of the words. Each scrambled sentence can be resolved in one of two ways, in a positive manner or in a negative manner (e.g., I am a born winner/loser). The SST may be deemed a direct measure of interpretation bias (Everaert et al., 2017). Research with the SST has revealed that individuals reporting subclinical and clinical forms of depression unscramble more sentences in a negative manner, which is indicative of an interpretation bias linked to depression (Everaert et al., 2014; Hedlund & Rude, 1995; Rude et al., 2010; Sanchez et al., 2017; Sfärlea et al., 2019a; Viviani et al., 2018).

Another frequently used task to study interpretation biases is the presentation of a series of ambiguous scenarios (Berna et al., 2011; Rohrbacher & Reinecke, 2014; Voncken et al., 2003). One variant of this is the Interpretation Bias Questionnaire (Wisco & Nolen-Hoeksema, 2010). In the IBQ, participants are presented with vignettes describing ambiguous daily life situations. In response to each vignette, participants need to generate as many explanations that
come to mind and select the explanation that they consider the most plausible. In this way, the IBQ can measure biases in both the initial generation and subsequent selection of interpretations, thus, can yield two indices representing direct measurement of interpretation bias. Studies with this task have shown that individuals reporting higher depressive symptom levels generate more negative interpretations and select more negative interpretations as the most plausible explanation of an ambiguous scenario (Wisco & Nolen-Hoeksema, 2010, 2011).

Researchers have utilized the Homograph Interpretation Task (Grey & Mathews, 2000) to study interpretation biases in depression. In this task, participants are presented with homographs that have both a negative connotation and a benign/positive connotation (e.g., the word ‘patient’) and interpretations of these homographs are examined (e.g., being sick in hospital or being generous with time and care). This task results in an index based on reaction time to the homographs presented thus reflecting an indirect measurement of interpretation bias. Faster reaction time to negative connotations of homographs indicate a more negative interpretation bias. Current research suggests that individuals with depression endorse more negative interpretations of an ambiguous homograph (Hertel & El-Messidi, 2006; Mogg et al., 2006). Another study examining the relationship between the HIT and depression has indicated that individuals with greater self-reported depressive symptom severity displayed poorer ability to vividly imagine positive but not negative events from ambiguous homographs (Holmes et al., 2008).

Finally, an increasingly popular task for studying interpretation biases is the Word-Sentence Association Paradigm (Cowden Hindash & Amir, 2012). The WSAP presents participants with an ambiguous sentence (e.g., people always tell you to smile) followed by an unambiguous word that is negative (e.g., defective) or benign (e.g., loved). Participants must indicate whether they believe the ambiguous sentence and the unambiguous word are related by endorsing or rejecting the unambiguous word. The reaction times and endorsement/rejection
rates to the sentence-word combinations are used to create two indices of interpretation bias. The reaction time index is an indirect measure of interpretation bias with faster reaction time to the endorsement of negative material and the rejection of positive material indicating a more negative interpretation bias. The endorsement/rejection rate index are direct measures of interpretation bias. Research on depression with the WSAP revealed that individuals with higher depressive symptom levels endorsed more negative interpretations and they were also faster to endorse negative words as being related to an ambiguous scenarios (Beard et al., 2017; Cowden Hindash & Amir, 2012; Cowden Hindash & Rottenberg, 2015, 2017; Smith et al., 2016, 2018).

Although direct and indirect measures are commonly used in research on interpretation biases in depression, little is known about the relationship between various direct and indirect measures of widely used interpretation tasks. To date, there is only one study that has examined relationship between interpretation bias indexes drawn from different experimental paradigms. In that study (Lee et al., 2016), participants diagnosed with clinical depression completed several verbal interpretation bias tasks including the SST and the similarity rating task. The similarity ratings task has been particularly used in research on anxiety (Salemk et al., 2010). During this task, participants read descriptions of ambiguous events and rate positive and negative interpretations for similarity to the original description (Lee et al., 2016). Correlation analysis showed a moderate correlation between the SST and the similarity rating task. Interestingly, the analysis regressing depressive symptom severity on the interpretation bias scores showed that the similarity ratings task but not the SST explained a unique portion of the variance (though both tasks were correlated with depression severity). These initial observations indicate that both interpretation bias tasks may measure a similar but not identical construct and could have differential unique relations with depressive symptom severity.
Consistent with these initial findings, past research has indicated that alternative measures of the same construct, such as interpretation bias, are rarely perfectly convergent and this can result in inconsistent findings and conclusions in research (Carlson & Herdman, 2012). Although perfect convergence would be rare, we would expect measures of interpretation bias to correlate significantly and strongly (Pearson Product Moment Correlation >.7, Schober et al., 2018) and given that they propose to measure the same cognitive construct. However, little is understood about the level of convergence amongst measures of interpretation bias and it is possible that this could be contributing to inconsistent findings in the field. In addition, research has yet to examine the unique relations between direct and indirect measures of interpretation bias and depressive symptom levels. Answering these questions seems critically important to understand whether direct and indirect measures reflect the same underlying process of ambiguity resolution or whether they reflect sub-processes which contribute to the process of ambiguity resolution and to identify the best marker of interpretation biases in depression.

Interpretation bias research has been criticized for failing to report psychometric properties and for using measures of interpretation bias with potentially poor reliability and validity (Cristea et al., 2015; Parsons et al., 2019) and there have been increasing calls to routinely report the spilt half reliability of cognitive tasks. Convergent validity and reliability of interpretation bias measures are not widely reported, and the strength of interpretation bias measures should be considered in light of their psychometric properties (Everaert et al., 2017).

The present study

To further our understanding of the nature of interpretation biases in depression, the present study examined the interrelations among direct and indirect measures from diverse interpretation bias tasks (aim 1) and tested which measure had the strongest unique association with depressive symptom levels (aim 2). This study further extends prior work (Lee et al., 2016) on the convergence between interpretation bias tasks and unique relations with depressive
symptoms by focusing on alternative measures of interpretation bias to build upon our existing knowledge base.

In relation to the first aim, it was expected that moderate to strong correlations among the interpretation bias measures would occur between direct measures and between indirect measures. However, because direct and indirect measures involve different measurement procedures and may capture different aspects of interpretation bias (Everaert et al., 2017; Hirsch et al., 2016), it was expected that weaker correlations between direct measures and indirect measures would occur compared to correlations between direct measures and between indirect measures.

With respect to the second study aim, it was hypothesised that some interpretation bias measures may be better suited to represent depression-related interpretation biases, as reflected by stronger and/or unique relation with depressive symptom severity. As noted, different measures may reflect different processes contributing to interpretation bias and thereby have differential relations with depression (see also Lee et al., 2016). Based on research showing that depression is primarily marked by cognitive biases operating at later stages of information-processing (Armstrong & Olatunji, 2012), it was hypothesised that experimental conditions involving direct measurement procedures for interpretation tasks (which allow more strategic interpretations processes; Hirsch et al., 2016) may have stronger associations with depressive symptoms.

**Method**

**Participants**

Eighty-two unselected undergraduate students (59 female, 21 male; $M = 25.01$ years; $SD = 6.482$) were recruited from the SONA research participant pool at XXX (Sona Systems, n.d.). All participants spoke English fluently and enrolled in a university in Ireland. All participants provided informed consent. Participants who scored on the minimal to very severe ranges (i.e.,
all participants) were included in the study. Participants scored in the Beck Depression Inventory II (Beck, Steer, & Brown, 1996) as follows: minimal (58), mild (4), moderate (11) and severe (7). Twelve participants reported having been given a clinical diagnosis of a mental health disorder in the past (eight of those reported anxiety, four of those reported depression, one reported panic disorder, one reported bulimia nervosa, one reported eating disorder not otherwise specified and one reported bipolar I disorder). Participants were excluded if they were under the age of 18. The institutional review board approved the study protocol.

**Depressive symptom severity**

The Beck Depression Inventory II (Beck et al., 1996) was used to assess the presence and severity of depressive symptoms. Participants rate 21 items on a scale of 0 to 3, with a higher value reflecting greater severity of a symptom. Total scores can be categorised as minimal (0-13), mild (14-19), moderate (20-28), or severe (29-63). A review of 118 studies investigating the psychometric properties of the BDI-II in clinical and non-clinical samples revealed that the internal consistency was approximately 0.9 and the test-retest reliability ranged from 0.73 to 0.96 (Wang & Gorenstein, 2013). The internal consistency for this study was excellent (α = 0.94).

**Trait anxiety**

The State Trait Anxiety Inventory: Trait Subscale (STAI; Spielberger, Gorsuch, & Lushene, 1970) is a 20-item self-report measure which assesses levels of trait anxiety. Items in the trait subscale are rated on a 4-point scale (‘almost never’ to ‘almost always’), with the higher scores indicating higher levels of trait anxiety. The STAI is widely used for research purposes given its high levels of test-retest reliability (0.65-0.75) and internal consistency (0.86-0.95; Spielberger, Gorusch, Lushene, Vagg, & Jacobs, 1983). In this study, the internal consistency of the trait subscale was excellent (α = 0.95).

**Interpretation bias tasks**
**Word Sentence Association Paradigm (WSAP)** (Cowden Hindash & Amir, 2012). Each WSAP trial presents an ambiguous sentence (e.g. You carry a tray of food at a party) for 1000ms followed by an unambiguous word that is either negative (e.g. clumsy) or benign (e.g. graceful). Participants were required to judge whether the ambiguous sentence and the unambiguous word were related as quickly as possible and to indicate their judgement using the mouse. A ‘yes’ response indicated that participants thought the sentence and word pair were related and a ‘no’ response indicated that participants thought the sentence and word pair were not related. Participants completed 170 trials consisting of ambiguous sentence – unambiguous word pairs. All stimuli were presented in black Times New Roman 14pt font on a white background. Endorsement (yes) and rejection (no) responses of the relationship between the word and the sentence were recorded. In addition, reaction times served as an index of the perceived relatedness of the semantic material.

Relative bias score comparing negative vs. benign interpretations were computed for the endorsement/rejection rates and the reaction time indices. The interpretation bias index based on the endorsement/rejection rates was obtained by calculating the percentage of trials on which participants either endorsed a negative word or rejected a positive word following an ambiguous sentence. The bias index based on the reaction times was obtained by subtracting the mean response times for trials on which negative words were endorsed and positive words were rejected from the mean response times for trials on which negative words were rejected and positive words were endorsed. This method was chosen to obtain similar relative bias scores across the different interpretation tasks in this study and informed by meta-analytic research indicating relative scores may better capture depression-linked interpretation biases (Everaert, Podina, & Koster, 2017).

**Interpretation Bias Questionnaire** (Wisco & Nolen-Hoeksema, 2010). The IBQ presents 10 hypothetical ambiguous scenario descriptions, instructing participants to imagine
each scenario occurring to themselves. Participants need to write down as many interpretations that came to mind in response to ambiguous situations (to measure biases in interpretation generation) and then indicate which explanation they considered the ‘most likely’ (to measure biases in interpretation selection). All responses to the IBQ were coded by two independent coders. Trained coders unaware of participants’ anxiety/depressive symptom levels rated the negativity and positivity of all interpretations using 5-point Likert scales (Wisco & Nolen-Hoeksema, 2010). Coders demonstrated adequate interrater reliability for ratings of the responses (ICC = .939). Any discrepancies were resolved by a third coder. Separate interpretation bias indices were created for interpretation generation and selection (Wisco & Nolen-Hoeksema, 2010). An interpretation generation bias score was created by subtracting the positivity ratings from the negativity ratings for all interpretations generated across the scenarios. An interpretation selection bias scores was computed by subtracting the positivity ratings from the negativity ratings of the interpretations that were selected as the most likely explanation across the scenario descriptions.

Scrambled Sentences Test (Wenzlaff & Bates, 1998). The SST requires participants to unscramble sentences using five of the six displayed words to form grammatically correct and meaningful statements (e.g., looks the future bright very dismal). By reporting the unscrambled sentence that first comes to mind, every sentence is resolved in either a positive (e.g., the future looks very bright) or negative (e.g., the future looks very dismal) manner. Participants completed 60 scrambled sentences in total.

Similar to prior work with the task (e.g., Rude et al., 2002), a cognitive load procedure was implemented to prevent deliberate (e.g., social desirable) report strategies. At the end of the session, participants were prompted to input the six-digit number presented at the beginning. A negative interpretation bias was created by dividing the number of negative resolutions by the number of total grammatically correct resolutions each participant gave.
The version of the SST used in this study was presented to the participants via EPrime 2.0 (Pittsburgh: Psychology Software Tools Inc). Participants were presented with six scrambled words in the top half of the screen. They were required to click on the words in the order they wanted to unscramble them in. When they clicked on a word, the word was presented in the lower half of the screen in their desired order.

**Homograph Test** (Grey & Mathews, 2000). The homograph test (HIT) in this study was adapted from Grey and Mathews (2000) and involved the presentation of 20 homographs with both a threat-related meaning and non-threat related meaning. On each trial of the task, participants were presented with a homograph (e.g. patient) and were then presented with a fragmented word (e.g., s_ck/k_nd) which when completed would be threat-related (e.g. sick) or non-threat-related (e.g. kind). Participants were instructed to press the space bar as soon as they were aware what the missing letter was. Once participants pressed the space bar, they were prompted to type the missing letter in the space shown. On detecting a response, or after 32 seconds, if no response was detected, a message was displayed prompting them to locate and press the missing letter key on the keyboard. A feedback message then appeared on the screen indicating correct or wrong response. Ten homographs were paired with threat-related meaning and the other ten homographs were paired with non-threat-related meaning. Stimuli were presented in black Times New Roman 14pt font on a white background. Reaction times were recorded to index interpretation bias. An interpretation bias scores was calculated by subtracting the mean reaction times for correct responses to threat-related words from the mean of reaction time for correct responses to non-threat-related words.

**Procedure**

All measures were presented to participants on a 13” laptop screen with a mouse attached on their dominant side either through Qualtrics (Qualtrics, Provo, UT) or Eprime 2.0 (Pittsburgh: Psychology Software Tools Inc) depending on the interpretation bias task.
Participants were seated approximately 50cm from the laptop screen throughout the study. Participants started with a questionnaire capturing demographic information, followed by the BDI-II. Participants then completed the measures of interpretation bias (WSAP, IBQ, SST, HIT). Participants were given the opportunity to take a small break between tasks. The order of the questionnaire was counter-balanced across participants manually by the researcher. The session lasted approximately 45 minutes.

**Data preparation and analysis**

Prior to addressing the study aims, all variables were inspected to identify outliers using the interquartile range rule. For each study variable, the interquartile range (IQR) was computed and multiplied by 1.5. This value was then added to the third quartile (Q3) and subtracted from the first quartile (Q1). Values outside these cutoffs were considered outliers. A total of 2.90% of all data was identified as extreme values. To reduce the impact of the outliers, the extreme values were replaced by cutoff values of Q3+1.5xIQR (for extremely high values) or Q1-1.5xIQR (for extremely low values).

Moreover, incomplete data was obtained for two participants. Little’s MCAR test ($\chi^2(30) = 21.56$, p=.869) indicated that missingness may be completely at random. Missing values for these participants were then estimated using the R-package MICE (Azur, Stuart, Frangakis, & Leaf, 2011; van Buuren & Groothuis-Oudshoorn, 2011). The imputation model included all study variables as predictors. The default method of predictive mean matching was used to create 5 completed datasets after 10 iterations.

To address the study aims, correlational analyses were performed to examine the convergence among interpretation bias task indices. Next, a multiple regression model was fitted to investigate unique associations among interpretation bias indices and depressive symptom levels. This regression model included all interpretation bias task indices as independent variables and depressive symptom severity scores as dependent variables. All
statistical analyses were conducted on each of the five imputed datasets and the results were pooled to obtain summary estimates.

**Results**

Means, standard deviations, correlations among the study variables are reported in Table 1.

**Aim 1: Convergence among interpretation bias task indices**

The absolute values of the pooled correlation coefficients varied from .001 to .543, indicating no to moderately strong associations among interpretation bias indices. Moderately strong correlations emerged for bias indices stemming from the same interpretation bias task, with .543 for IBQ indices ($p<.001$) and .460 for the WSAP ($p<.001$). With respect to relations between bias indices from different tasks, the correlation analysis showed that the HIT bias index was not significantly related to any of the other bias indexes (all $p$’s>.05). By contrast, the SST bias index had the strongest correlations with other biases indices. The SST bias index was significantly correlated with IBQ selection bias ($r=.530$, $p<.001$), IBQ generation bias ($r=.421$, $p<.001$), and WSAP Endorsement/Rejection bias ($r=.469$, $p<.001$). There was a marginally significant correlation between the SST index and the WSAP reaction time bias index ($r=.218$, $p=.053$). Finally, the IBQ generation bias was significantly correlated with the WSAP Endorsement/Rejection bias ($r=.292$, $p=.008$) and marginally correlated with the WSAP reaction time bias index ($r=.217$, $p=.05$). The IBQ selection bias had a significant correlation with the WSAP Endorsement/Rejection bias ($r=.400$, $p<.001$), but not with the WSAP reaction time bias index ($r=.154$, $p=.194$).

**Aim 2: Associations between interpretation biases and depressive symptom severity**
Correlational analyses (see Table 1) showed that depressive symptom severity was significantly related to IBQ selection bias ($r=.418$, $p<.001$), IBQ generation bias ($r=.329$, $p=.002$), SST bias ($r=.800$, $p<.001$), and WSAP endorsement bias ($r= 387, p< .001$). The WSAP reaction time bias index was not significantly related to depressive symptom severity scores ($r=.162, p=.143$). The results of pooled regression models showed that the interpretation bias indices explained a significant amount of the variance in depressive symptom severity scores, adjusted $R^2=.649$, 95%CI[.507, .759]. Inspecting the regression coefficients (see Table 2), the results showed that the SST bias index and the HIT bias index were associated with depressive symptom severity when including all interpretation bias indices in the regression model. This suggests that the SST and HIT bias indexes are uniquely related to depressive symptom severity. The other bias indexes did not explain a unique portion of the variance in depressive symptom severity scores. Similar correlational patterns were observed between trait anxiety and interpretation bias indices.

**Split-half reliability of interpretation bias task indices**

Following increasing calls to routinely report the split-half reliability of cognitive tasks (Cristea et al., 2015; Parsons et al., 2019), the present study computed the split-half reliability using the adjusted Spearman–Brown prophecy formula (Brown, 1910; Spearman, 1910). The reliability indices were obtained by correlating scores on even-numbered items or trials with scores on odd-numbered items or trials. The split-half reliability for the SST was $r=.945$. For the HIT, a split-half reliability of $r=.258$ was observed. The split-half reliability for the WSAP was $r=.268$ for the bias index based on reaction times and $r=.592$ for the bias index based on endorsement/rejection rates. Finally, the split-half reliability for the IBQ was $r=.758$ for interpretation generation bias and $r=.560$ for the interpretation selection bias.
Discussion

This study aimed to advance our understanding of the nature of interpretation biases in depression by examining the convergence among interpretation bias measures and their associations with depressive symptom severity in an unselected sample. Notably, this study observed no to moderately strong correlations among interpretation bias indices that were derived from commonly used interpretation bias tasks. This finding suggests that bias indices derived from the Interpretation Bias Questionnaire (Wisco & Nolen-Hoeksema, 2010), Scrambled Sentences Test (Wenzlaff & Bates, 1998), Homograph Interpretation Task (Grey & Mathews, 2000), and Word-Sentence Association Paradigm (Cowden Hindash & Amir, 2012) may not all measure the same underlying process or component of interpretation bias. As anticipated from previous studies (Everaert et al., 2017; Lee et al., 2016; Sfärlea et al., 2019b), moderately strong correlations emerged among the direct measures; IBQ selection bias, IBQ generation bias, the SST bias, and WSAP Endorsement/Rejection bias indices. Indirect measures based on reaction times (from the WSAP and HIT) were not correlated. Correlations between direct and indirect measures were small, with exception of the moderate correlation between the WSAP reaction time and Endorsement/Rejection bias indices (which is likely driven by shared method factors). This pattern of findings is consistent with the notion that direct vs. indirect measurement procedures derived from interpretation bias tasks may represent sub-processes contributing to interpretation bias (Everaert et al., 2017; Hirsch et al., 2016).

It is possible that response factors associated with completing the measures contributed to the findings. Deliberate response strategies such as demand characteristics and features related to depressive symptoms may drive responses to tasks that require participants to report or endorse content of emotional interpretations of the ambiguous material such as the IBQ. For instance, individuals with depressive symptoms may consciously indicate a more negative interpretation bias and this could be driven by demand characteristics rather than interpretation
bias (Gotlib & Joormann, 2010). Depressed individuals may also respond in a negative manner because of factors related to anhedonia which results in loss of motivation to participate and by default, they choose the most negative interpretation for the ambiguous information presented to them (Everaert et al., 2017). These strategies are unlikely to be utilized in indirect measures given that they are associated with more strategic processes. It is important to note that some interpretation bias measures, namely the SST and WSAP, are less susceptible to response biases due to their design. The SST implements a cognitive load procedure and instructs participants to report the first interpretation that comes to mind, and the WSAP prompts participants to respond quickly and imposes a time limit. With such task requirements, it is less likely that responding on a given task would be influenced by deliberate response strategies. Therefore, response strategies involved in completing tasks may contribute to but do not provide a comprehensive explanation of the pattern of findings observed in this study.

The most plausible perspective on the observed pattern of correlations between direct and indirect measures in this study may reside in the nature of the indirect measures, specifically reaction time indices (Cristea et al., 2015; Parsons et al., 2019). Relatively poor split-half reliability was found for bias indices based on reaction time, whereas acceptable to good reliability was found for direct bias measures. While reaction times have been used commonly in research on cognitive biases, there are increasing concerns about their psychometric properties. In particular, researchers have pointed out that aggregate indexes based on differences between reactions times on different trial types have low reliability for the measurement of individual differences (Hedge et al., 2018). This differential reliability could be a source of the limited convergence among measures that are supposed to reflect the same construct.

Regardless of the sources driving differences in correlations in this study, the substantial range of correlations between various measures of interpretation bias present a challenge in
examining and understanding findings in this field in general. Research suggests that even when
convergent validity among measures is strong ($r = .90$), up to 5% of findings derived from these
measures may differ (Carlson & Herdman, 2012). In the present study, correlations between
measures of interpretation bias ranged from -.12 to .54 (or .29 to .54, when excluding measures
based on reaction time), indicating at best, moderate convergent validity. Carlson and Herdman
(2012) suggest that alternative measures of the same construct with convergent validity less
than $r = .50$ cannot be substituted for the desired measure. This research supports the idea that
inconsistent findings in interpretation bias research and interpretation bias modification
research could be due to the lack of convergent validity that exists between measures. This has
implications for the use of these measures in interpretation bias (modification) research and
future research designs should address this limitation by improving the validity of these
measures using factorial analyses.

Consistent with theoretical models (Beck, 1976; Clark et al., 1999; Ingram et al., 1999)
and previous research, this study observed that depressive symptom severity scores are related
to various interpretation bias measures. Depressive symptom levels were correlated with a bias
in generating and selecting interpretations (IBQ), unscrambling sentences negatively (SST), as
well as endorsing more negative and rejecting more positive interpretations (WSAP). This
pattern of correlations is in line with findings from previous studies investigating interpretation
biases in depression whereby it was reported that smaller effect sizes occur for studies on
depression that use indirect measures than direct measures (Everaert et al., 2017). Interestingly,
when accounting for potential shared variance, the present study observed that only two bias
indices explained a unique portion of variance in depression symptoms (the SST and HIT).
While the HIT finding may be less reliable because of its low split-half reliability, the SST
finding is particularly interesting and suggests that the SST bias index may be the best marker
of interpretation biases in depression.
The SST may tap into features that are present in the other interpretation bias tasks administered in this study. The SST bias index showed the strongest correlations with the other bias measures of both the IBQ and WSAP. By using a cognitive load procedure and instructing participants to report the first sentence that comes to mind, the SST may capture processes that are involved in the activation and selection of plausible interpretations of ambiguity. This process may play an important role in quickly endorsing/rejecting negative/positive meanings in the WSAP and both generating and selecting interpretations in the IBQ. As such, the SST may capture both automatic and controlled processes contributing to interpretation bias by relying on meaning activation and selection. Of note, adding a cognitive load procedure to the SST may be critically important in capturing depression-related interpretation biases. Previous research has observed differential findings based on the delivery of the SST with and without cognitive load (Rude et al., 2003). Moreover, the SST with cognitive load is more effective in impacting on interpretation bias change (Bowler et al., 2012).

This finding that the SST was uniquely associated with depressive symptom severity runs counter to the finding by Lee et al. (2016) in which the SST did not contribute independently to a model of interpretation bias measures and depression. The difference between the findings of the studies may be attributed to the selection of other interpretation bias tasks and thereby the bias indices added to the statistical model. This study focused on interpretation bias tasks that are commonly-used in research on depression, but did not consider all interpretation bias tasks in the literature. Future research could explore a wider range of interpretation bias measures to understand the associations between measures of interpretation bias and depression in a sample of participants ranging from normal to very severe levels of depression. Further research could also incorporate measures of anxiety symptoms given that correlational patterns between trait anxiety and interpretation bias measures observed in the
The present study resembled the correlations patterns between depressive symptoms and interpretation bias measures.

The observations with regard the HIT deserve some elaboration. The bias index from the HIT was not significantly correlated with any of the other bias indices. A potential explanation for this is that the materials used in the HIT may be more targeted at threat-related interpretation (e.g., choke) in comparison to the other interpretation bias measures and perhaps threat is an important aspect of interpretation bias that needs to be considered with depression. Notably, the HIT also shared a unique association with the BDI-II when compared to other measures of interpretation bias. Considering these two findings simultaneously, this indicates that the HIT could be measuring a cognitive process which is different from the understanding of interpretation bias observed in the other tasks, however, is in some way contributing to depressive symptoms. However, again, the findings should be interpreted in light of the poor reliability of the task. The idea of using words with more than one meaning (one threat-related and one benign or positive) seems as though it would present an ideal opportunity for assessing negative biased processing, although the limited number of potential homographs in the English language that would have one threat-related and one benign/positive connotation presents a significant problem.

The findings of this study may have implications for cognitive training programs targeting interpretation processes. Interpretation bias modification interventions differ with respect to the experimental paradigm used to modify interpretation bias (e.g. WSAP; Möbius et al., 2015; ambiguous scenario training; de Voogd et al., 2017) as well as task to measure changes in the bias as a result of the training. The current study findings suggest that the nature of the experimental paradigm used to measure the bias, and potentially its training variant, may be contributing to differential findings in the context of depression (e.g., some interpretation bias modification programmes do not transfer to other measures of interpretation bias (Hertel...
& Mathews, 2011). It is possible that various paradigms measure interpretation processes that are differentially related to depression. Therefore, cognitive bias modification research may benefit from utilizing experimental training paradigms that target interpretation processes that are uniquely and strongly related to depressive symptoms. The observations of this study indicate that interpretation processes as measured by the SST could be a promising target.

The findings of this study should be interpreted in light of some limitations. Namely, this research was conducted on a sample of university students, which limits the generalizability of the present findings. It is possible that this specific age group did not complete some cognitive tasks such as the HIT with the same understanding of some words as other cultural groups (e.g., ‘stole’ in relation to ‘fur’). Before any experimental research is conducted on a specific sample (e.g., university students in Ireland), the measures should be validated for that sample (Ziegler, 2014).

The average depressive symptom scores indicate that there were overall minimal depressive symptom levels. Although there was some variability in symptom severity scores, caution should be given in generalizing the present findings to individuals experiencing severe and clinical levels of depression. This study observed considerable variability in the interpretation bias indices, allowing this study to examine individual differences in how people interpret ambiguity. Future research should be conducted in samples of individuals with sub-clinical and clinical levels of depression as well as individuals in remission from depressive episodes to further understand the nature of interpretation biases in depressive symptoms.

Another limitation concerns the interpretation bias tasks that were included in this study. Though the IBQ, SST, WSAP, and HIT are among the most used tasks, other methods and tasks exist and deserve consideration in future research aiming to understand the nature of interpretation biases in depression. Indirect measures of interpretation bias allow us to investigate interpretation bias without the potential impact of deliberate response strategies and
studies which incorporate a range of tasks in assessing interpretation are likely to be more robust (Hirsch et al., 2016). Further research should aim to improve the reliability and validity of the indirect measures used in this study and other existing indirect measures (e.g., semantic priming tasks (Bisson & Sears, 2007; Lawson & MacLeod, 1999; Mogg et al., 2006) and affective startle reflexes (Allen et al., 1999; Forbes et al., 2005; Kaviani et al., 2004; Lawson et al., 2002)) and to explore the potential for novel indirect measures (e.g., combining cognitive and biological methods to measure event-related potentials (Moser et al., 2008). This could enhance the comprehensiveness of interpretation bias research (Hirsch et al., 2016).

Despite these limitations, this study advances the understanding of the nature of interpretation biases in depression theoretically and methodologically. The present investigation is the first to examine the interrelation among direct and indirect interpretation bias tasks and their association with depressive symptoms. It was found that direct measures of interpretation bias were moderately correlated. The relationship between indirect measures of interpretation bias was weak and could be explained by their limited split-half reliability. The SST bias index was uniquely related to depression and may be the best marker of depression-linked interpretation biases. Theoretically, this suggests that the cognitive processing involved in the SST may be associated with depressive symptoms.
Acknowledgements

This The authors would like to acknowledge Dr. Alexandra Cowden-Hindash and Dr. Anne-Wil Kruijt for providing materials to conduct this research.

Conflict of Interest

The authors declare they have no competing interests.

Funding

This work was supported by the Irish Research Council (Scholarship ID number: GOIPG/2015/3968).

References


https://doi.org/10.1016/j.jad.2004.04.007


https://doi.org/10.1016/j.brat.2016.05.007


https://doi.org/10.1016/j.jbtep.2015.03.002


https://doi.org/10.1016/j.brat.2005.10.008


https://doi.org/10.1177/2515245919879695


Table 1. Descriptive statistics and correlations among study variables.

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>M (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. BDI-II</td>
<td>.81&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.418&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.329&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.800&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.197</td>
<td>.162</td>
<td>.387&lt;sup&gt;c&lt;/sup&gt;</td>
<td>10.84 (10.81)</td>
</tr>
<tr>
<td>2. STAI</td>
<td>.41&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.37&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.64&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.09&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.14&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.46&lt;sup&gt;c&lt;/sup&gt;</td>
<td>45.54 (9.94)</td>
<td></td>
</tr>
<tr>
<td>3. IBQ selection</td>
<td>.543&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.530&lt;sup&gt;c&lt;/sup&gt;</td>
<td>-.001</td>
<td>.154</td>
<td>.400&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.036 (0.90)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. IBQ generation</td>
<td>.421&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.037</td>
<td>.217&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.292&lt;sup&gt;a&lt;/sup&gt;</td>
<td>-0.65 (0.81)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. SST</td>
<td>.023</td>
<td>.218&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.469&lt;sup&gt;c&lt;/sup&gt;</td>
<td>0.22 (0.22)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. HIT</td>
<td>-.060</td>
<td>-.120</td>
<td>114.20 (537.96)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. WSAP-RT</td>
<td>.460&lt;sup&gt;c&lt;/sup&gt;</td>
<td>-159.80 (571.13)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. WSAP-End/Rej</td>
<td>0.37 (0.15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* <sup>a</sup>p<.05, <sup>b</sup>p<.01, <sup>c</sup>p<.001; higher scores indicate more negative interpretation bias; BDI-II=Beck Depression Inventory – II; IBQ=Interpretation Bias Questionnaire; SST=Scrambled Sentences Test; HIT=Homograph Interpretation Task; WSAP-RT=Word-Sentence Association paradigm – Reaction Time bias index; WSAP-End/Rej=Word-Sentence Association paradigm – Endorsement/Rejection rate bias index.

Table 2. Pooled results from regression models predicting depressive symptoms

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.83</td>
<td>2.50</td>
<td>0.33</td>
<td>.741</td>
</tr>
<tr>
<td>IBQ selection</td>
<td>0.21</td>
<td>1.15</td>
<td>0.18</td>
<td>.858</td>
</tr>
<tr>
<td>IBQ generation</td>
<td>-0.07</td>
<td>1.11</td>
<td>-0.06</td>
<td>.950</td>
</tr>
<tr>
<td>SST</td>
<td>38.80</td>
<td>4.13</td>
<td>9.39</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>HIT</td>
<td>10.54</td>
<td>3.88</td>
<td>2.72</td>
<td>.008</td>
</tr>
<tr>
<td>WSAP-RT</td>
<td>-0.22</td>
<td>1.42</td>
<td>-0.16</td>
<td>.876</td>
</tr>
<tr>
<td>WSAP-End/Rej</td>
<td>3.69</td>
<td>6.20</td>
<td>0.60</td>
<td>.553</td>
</tr>
</tbody>
</table>

*Note.* B=pooled regression coefficient; SE=Standard error of pooled regression coefficient.