HHS Public Access

Author manuscript

J Abnorm Psychol. Author manuscript; available in PMC 2017 August 01.

Published in final edited form as:

J Abnorm Psychol. 2016 August ; 125(6): 840–851. doi:10.1037/abn0000184.

Unreliability as a Threat to Understanding Psychopathology: The Cautionary Tale of Attentional Bias

Thomas L. Rodebaugh[a],
Washington University in St. Louis

Rachel B. Scullin,
Washington University School of Medicine

Julia K. Langer,
Washington University in St. Louis

David J. Dixon,
Washington University School of Medicine

Jonathan D. Huppert,
The Hebrew University of Jerusalem

Amit Bernstein,
Haifa University

Ariel Zvielli, and
Haifa University

Eric J. Lenze
Washington University School of Medicine

Abstract

The use of unreliable measures constitutes a threat to our understanding of psychopathology, because advancement of science using both behavioral and biologically-oriented measures can only be certain if such measurements are reliable. Two pillars of NIMH's portfolio – the Research Domain Criteria (RDoC) initiative for psychopathology and the target engagement initiative in clinical trials – cannot succeed without measures that possess the high reliability necessary for tests involving mediation and selection based on individual differences. We focus on the historical lack of reliability of attentional bias measures as an illustration of how reliability can pose a threat to our understanding. Our own data replicate previous findings of poor reliability for traditionally-used scores, which suggests a serious problem with the ability to test theories regarding attentional bias. This lack of reliability may also suggest problems with the assumption (in both theory and the formula for the scores) that attentional bias is consistent and stable across time. In contrast, measures accounting for attention as a dynamic process in time show good reliability in our data.

[a]Corresponding author. Department of Psychology, Washington University in Saint Louis: 1 Brookings Drive, CB 1125 Psychology Building, St. Louis, MO 63130, phone: 314-935-8627, rodebaugh@wustl.edu, Fax: 314-935-7588.
Thomas L. Rodebaugh, Department of Psychology; Rachel B. Scullin, Eric J. Lenze, and David J. Dixon, Department of Psychiatry, Washington University School of Medicine. Jonathan D. Huppert, Department of Psychology, Hebrew University. Amit Bernstein and Ariel Zvielli, Department of Psychology, University of Haifa. Julia K. Langer is now at the Minneapolis Veterans Affairs Health Care System. The contents do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

The field is sorely in need of research reporting findings and reliability for attentional bias scores using multiple methods, including those focusing on dynamic processes over time. We urge researchers to test and report reliability of all measures, considering findings of low reliability not just as a nuisance but as an opportunity to modify and improve upon the underlying theory. Full assessment of reliability of measures will maximize the possibility that RDoC (and psychological science more generally) will succeed.

Keywords

attentional bias; reliability; anxiety disorders; methodology; RDoC; target engagement; personalized medicine

Numerous theorists and researchers have argued that because mental disorders defined by purely descriptive diagnostic criteria (American Psychiatric Association, 2013) remain poorly understood, we should look elsewhere in our attempts to delineate the factors that underlie them (Engstrom & Kendler, in press; Kihlstrom, 2002). Accordingly, the National Institute of Mental Health (NIMH) has favored a research program based on Research Domain Criteria (RDoC; Cuthbert & Kozak, 2013). The ambitious RDoC project aims to develop an empirical taxonomy of mechanisms sub-serving psychopathology. Accordingly, in RDoC materials, purportedly psychopathology-related mechanisms have been assembled as rows in a matrix, with the matrix columns representing observable indicators in different forms of measurement, from genes, to brain responses, to behaviors, to self-report. The effort therefore relies on precise delineation of key mechanisms sub-serving psychopathology and psychometrically robust indicators reflecting these processes.

We believe that the RDoC project will succeed or fail in part according to what extent measures possessing *reliability* can be found for the proposed indicators. In illustrating our point, we focus on one cell from this emerging empirical taxonomy of psychopathology: *attentional bias toward threat*. Attentional bias toward threat has been researched a great deal, particularly in the area of anxiety disorders. Meta-analyses of various measures of attentional biases have suggested that anxiety disorders are associated with an attentional bias toward threat (Bar-Haim, Lamy, Pergamin, Bakermans-Kranenburg, & van Ijzendoorn, 2007). Such attentional biases are further proposed to be vulnerabilities for anxiety and the development of anxiety-related disorders (Hirsch et al., 2011; Mathews & MacLeod, 1985; Mathews & MacLeod, 1994; Van Bockstaele et al., 2014). In a potentially related finding, functional polymorphisms in the serotonin transporter gene show a similar pattern, with purported risk alleles being related to bias toward threat in comparison to non-risk alleles (Pergamin-Hight, Bakermans-Kranenburg, van Ijzendoorn, & Bar-Haim, 2012). Attentional bias is thus of interest both as a potential risk marker (Perez-Edgar et al., 2010) and as a potential focus of intervention to treat anxiety disorders (Amir, Beard, Burns, & Bomyea, 2009; Hazen, Vasey, & Schmidt, 2009; Krebs, Hirsch, & Mathews, 2010). On the strength of such research, attentional bias toward threat would indeed seem to have a natural home as a cell in the RDoC matrix.

Interest and findings regarding attentional bias have gone well beyond anxiety disorders. Research suggests that depression (and risk for depression) produces altered attention

patterns that can be detected using tasks similar to those used in the study of anxiety (Joormann, Talbot, & Gotlib, 2007). Researchers have also investigated the possibility of attentional bias playing a role in other disorders, such as pain disorders (Dear, Sharpe, Nicholas, & Refshauge, 2011). All of these findings suggest that attentional bias is a basic construct underlying a range of psychopathology (Harvey, Watkins, & Mansell, 2004). Researchers have often implied that attentional bias and its behavioral task measurement may be particularly useful because it is closer to more obviously biological processes than to self-report, largely because it is thought to be relatively automatic and involuntary (cf. McNally, 1995, for a review and critique of such claims).

However, there is a major problem with many of the methods used to examine attentional bias: Most studies of attentional bias do not report any psychometric information for the measure used, and those that do typically report poor reliability (we review these studies below).[1] As a case example, we will use the most common method for assessment of attentional bias (cf. Bar-Haim, et al., 2007): the dot probe task (Mathews, MacLeod, & Tata, 1986) and the scores of global aggregated mean attentional bias toward threat derived from it. We will refer to this typically-used index as the *global attentional bias score* through the remainder of this paper.

Our purpose in this paper is to describe the problem that poor reliability poses for the understanding of psychopathology, using attentional bias as an example. Moreover, in pursuing solutions to this problem, we find evidence of the potential to improve both measurement and theory. We first illustrate the problem and its implications and then explore possible solutions, which include: (a) solve typical reliability problems, (b) consider that one's theory may be incorrect, and (c) investigate alternative methods. Importantly, attentional bias is just one example of the issues we raise: All cells in the RDoC matrix, including both behavioral tests and biological markers, deserve scrutiny, and *all* are susceptible to potential problems with reliability. To have a component of a measurement matrix with an unreliable method for measuring one cell will be an impediment to understanding the entire row of that matrix, and, by implication, may point to an Achilles' heel of the entire RDoC project. In fact, researchers have already put the global attentional bias score to uses that require excellent reliability, underscoring the problems that result if the scores are, in fact, not reliable.

Examples of Problems Resulting from a Lack of Reliability

Low reliability is a particular threat to two methodological issues of great interest: Selection based on individual differences and mediation analyses. These two issues underlie the main pillars of the NIMH's clinical translational focus, namely the RDoC project and the move towards target engagement studies in intervention development and testing. More broadly, these issues are essential to the concept of personalized or precision medicine (e.g., Pencina

[1]We assume that the reader will agree with at least one of the following statements: (a) Reliability is important in the ways described in typical treatments of classical test theory; or (b) Reliability as dealt with or computed via classical test theory is at worst a philosophically incoherent attempt to get at the more important issue of validity. For more regarding the former, see, e.g., Nunnally and Bernstein (1994); for more regarding the latter, see, e.g., Borsboom (2005). Adopting either viewpoint, lack of reliability is an inherent problem when it threatens valid uses of scores; we will depict such threats below.

& Peterson, 2016) and personalized treatment more generally. As discussed below, all attempts to either test mediation or identify individuals with certain characteristics require very high measurement reliability.

The possibility of using information about individual differences to select specific participants (e.g., for specific treatments) represents a promising means of applying psychopathology research. For example, Calamaras and colleagues used attentional bias scores at pretreatment to identify subgroups of participants who had biases toward vs. away from threat, and found that these biases were reduced after treatment (Calamaras, Tone, & Anderson, 2012). Unfortunately, confidence in the interpretation that treatment caused these changes requires measures with excellent reliability and stability, and what level of reliability should be considered *good* or *excellent* must be tied to the intended use of the score. As discussed by Nunnally and Bernstein (pages 264–265, 1994, among others), making decisions about individuals requires a particularly high level of reliability, higher than required for detecting an association between an individual difference variable and an outcome, or for detecting a difference between groups. The example provided by Calamaras et al. (2012) involves using individual scores to sort participants into groups. Only quite high reliability would justify confident decisions about individuals based on scores (e.g., Nunnally and Bernstein suggest that internal consistency greater than .90 would be preferable). Calamaras et al., with appropriate caution, speculated that their findings might reflect stabilization of attentional bias over the course of treatment. In the absence of good reliability, however, such findings could also simply reflect a statistical artifact (e.g., random error combined with regression to the mean).

To illustrate that the same finding could be due to poor reliability and not an intervention effect, we turned to our own data, which is described in detail in the supplementary material for this paper (Study 1). For our current purposes, it is only necessary to know that we selected the participants who had the most extreme global bias scores at the first of two study visits between which no interventions occurred; thus, a plausible explanation for any shift in bias scores could be regression to the mean. As shown in Figure 1, the overall impression is that the clear difference between the groups at Visit 1 vanishes, quite similarly to the change over time seen by Calamaras and colleagues. Of course, one could generate other explanations for this pattern of scores, and we cannot demonstrate conclusively here that the effect seen by Calamaras et al. is definitely due to the combination of random error and regression to the mean. Our point is simply that selecting participants based on extreme, unreliable, and unstable scores will tend to produce apparent change across time even when no substantive change has actually occurred.

Tests of mediation of intervention effects via attentional bias modification is another important example of research requiring high reliability (e.g., Amir et al., 2009). Tests of mediation are important to the RDoC effort because they allow a demonstration of causal chains. For example, a test of mediation is necessary to fully test a proposition such as: certain genetic profiles cause particular brain development patterns that, in turn, cause self-report symptoms. The RDoC matrix implies such mediation tests, as well as similar mediation tests within the realm of treatment (i.e., that treatments work on symptoms because they engage certain processes). Moving beyond studies of RDoC, intervention

studies have had an increasing focus on *target engagement* (Paul et al., 2010; Simon, Niphakis, & Cravatt, 2013), or determining specific mechanisms (i.e., the processes engaged) of pharmacological and psychosocial treatments. When an intervention shows target engagement, it is demonstrable that the mechanism thought to be involved in the treatment has been activated by the treatment and can therefore be proposed to be a mechanism of change. Accordingly, target engagement is essentially an initial step toward testing mediation. As with mediation, if attentional bias cannot be reliably measured, then by definition target engagement cannot be reliably demonstrated. The problem is then compounded for the eventual test of mediation, because such tests require very high reliability (see below). The endeavor to demonstrate target engagement is thus also highly vulnerable to threats of poor reliability.

Based on simulation studies, it has been recommended that individual measures used in mediation analyses have reliabilities of .90 or greater (Hoyle & Kenny, 1999; Hoyle & Robinson, 2004) unless latent variables can be estimated (which we have not seen done in the attentional bias literature). Notably, it has been suggested that several recent failures to detect effects on symptoms of modifying direction of attentional bias should be considered failures of manipulation, not failures of treatment, because the authors did not demonstrate that their intervention successfully manipulated the direction of global attentional bias (Clarke, Notebaert, & MacLeod, 2014; MacLeod & Clarke, 2015). That is, the implication is that the *treatment* did not fail per se; rather, the *researchers* failed to implement the treatment effectively, because if they had done so, then attentional bias would change. Certainly the general sentiment that one should conduct manipulation checks (and test for mediation when feasible) is an admirable one. However, the contention that is inherent to target engagement research – that attentional bias modification cannot be found to be successful unless researchers show that change in attentional bias mediated change in clinical measures – is predicated on the idea that there is a reliable way to assess attentional bias.

One might wonder if the problem with testing mediation is the fact that at least three time points and three measures should be involved in a proper test of mediation (see Cole & Maxwell, 2003, for a review). Perhaps a piecemeal test of whether an intervention has changed attentional bias would be less troubled by problems with reliability. However, if there is no reliable way to assess direction of attentional bias (at one time point), then the assertion that researchers should be able to demonstrate that they *changed* direction of biases cannot follow logically. Demonstrating change, along with testing mediation and selecting individuals based on scores, is yet a third instance in which high reliability is usually essential. This is because *change* in bias (i.e., a different score) should be expected to be less reliable than a single estimate of attentional bias under most conditions (see our discussion of difference scores below).

To summarize, the particular problems poor reliability causes for selecting individuals and testing mediation have not only led to difficulties for attentional bias research, but also have implications for the RDoC endeavor more generally. Without excellent reliability, personalized treatment is impossible to pursue effectively. Similarly, without excellent reliability, one cannot adequately test mediation, and it becomes impossible to determine how the cells in the matrix relate to each other causally. It seems worth emphasizing that

poor reliability can make it challenging to conduct any research endeavor; here we focus specifically on research aims that require the highest levels of reliability. The use of measures with only minimally acceptable reliability, for example, can lead to a frustrating situation: Many studies (e.g., with some failures to replicate) might demonstrate that an individual difference can predict outcomes, yet attempts to select participants based on that individual difference would have an unacceptably high failure rate. The solution to the problem is obtaining excellent reliability.

The Poor Reliability of Global Attentional Bias Scores Relying on Aggregated Means

We have thus far focused on the problems that poor reliability can cause, and only mentioned that the global attentional bias score has not shown good reliability. The reader may wonder, however, exactly what previous studies on this topic have shown. We have reviewed all of the studies that we could find ($n = 13$) that either focus on the reliability of global attentional bias scores, or at least report this reliability. Most studies of the reliability of global attentional bias scores show unacceptable reliability. This has been true when the task has used anxiety-related stimuli across different stimulus durations, clinical versus nonclinical participants, and words versus visual stimuli (including faces) (Schmukle, 2005; Staugaard, 2009). A more recent study reported a similar lack of test-retest stability in an fMRI context (Britton et al., 2013). Another found no stability of attentional bias indices from the dot probe, whereas an index derived from the evoked response potential was quite stable but lacked a relationship with anxiety (Kappenman, Farrens, Luck, & Proudfit, 2014). A study of attentional bias modification found no initial reliability of the global attentional bias score, although reliability began to approach a good level as participants completed many (i.e., several thousands) of dot-probe trials during training sessions (Enock, Hofmann, & McNally, 2014). We do not consider this study to demonstrate good reliability for the global attentional bias score because the initial assessment (i.e., what would be used for most assessment purposes) showed no evidence of reliability. Yet another study found no evidence for acceptable internal consistency of attentional bias indices in undergraduate participants high and low in social anxiety (Waechter, Nelson, Wright, Hyatt, & Oakman, 2014). Another study focusing on attentional bias toward pain stimuli similarly concluded that reliability of the resulting indices was low (Dear, et al., 2011).

In seeming contrast to the findings recounted above are six empirical results suggesting potentially acceptable reliability, although three of these only found reasonable reliability via modifications to standard practice. In one study, a split-half reliability of the global attentional bias index was reported as $r = .45$ (Bar-Haim et al., 2010). In a second, split-half reliability for the index was reported as $r = .44$ in the sample overall, ranging from .09 to .59 across study groups and trial types (Waechter & Stolz, 2015). A single study has suggested that an alternative way of processing the data produced by the dot-probe task may provide at least modest reliability for the traditional score (Price et al., 2015). A trio of studies focused on trial-level bias signal (TL-BS) indices and reported evidence of modest to good reliability; the authors of these studies, however, found no evidence of acceptable reliability for the global attentional bias score (e.g., Amir, Zvielli, & Bernstein, in press; Schäfer et al.,

in press; Zvielli et al., 2015). We will describe TL-BS indices further as an example of a potential solution to the problem of poor reliability, but for now maintain our attention on the standard scores.

The efforts by Price et al. (2015) deserve specific comment. These researchers performed an admirable service to the field in examining a wide variety of ways that handling outliers and other aspects of the underlying reaction time data might improve reliability of the standard attentional bias score. The core premise of their work was that problematic reliability may be in large part accounted for by data cleaning issues to which response time data may be particularly sensitive. Accordingly, they primarily focused on whether certain trial types might be more reliable (e.g., trials in which the participants must respond to stimuli at the top versus the bottom of the screen), as well as how to handle extreme values in response times. Although they found a method that produced better reliability in their data, it is important to realize that the method produced: (a) good reliability (intra-class correlation coefficient = .65) in a single study and adequate reliability in two other studies only when multiple dot-probe tasks were included in calculation of reliability and (b) uniformly poor reliability when a single dot-probe task session was included in calculations. (Price and colleagues also report on a measure of variance of attentional bias, but here we focus on the global attentional bias score.) As Price and colleagues note, most studies and most clinical applications are more likely to focus on a single dot-probe task session, rather than multiple sessions across time combined in a single index. Thus, Price et al. did find a potential method for obtaining better reliability, but this way does not correspond to how attentional bias scores are typically calculated or used. Their findings thus provide no reassurance that previous studies using the dot probe task produced reliable attentional bias scores.

Further Empirical Demonstration of the Poor Reliability of the Global Attentional Bias Score

In Study 1 of the online supplementary material, we present in detail our attempts to assess and improve the reliability of global attentional bias scores from the dot probe task with words. To aid the reader, we will summarize some key points from our study here. We assessed 24 participants with a current anxiety disorder, using two dot probe task sessions, approximately three weeks apart. Notably, a total sample of 20 is sufficient for good power (.90) to detect borderline adequate reliability (reflected in a correlation of .60; Faul, Erdfelder, Buchner, & Lang, 2009). Thus, our study was sufficiently powered. However, we should note that our data are only provided as a demonstration of the ubiquity of findings that are common in the literature: We by no means suggest that our study is the most well-powered demonstration of these common findings.

We measured bias over 1216 trials, which is more trials than is typical, in part because adding items (here, trials) is a common method for improving reliability of aggregated mean scores, in which variability reflects measurement error (cf. Price et al., 2015). The number of trials also permitted calculation of bias scores by word pair (i.e., participants responded to 16 trials for each word pair, allowing calculation of attention bias for each word). We were thus able to test internal consistency using Cronbach's alpha with specific-word bias scores

as items, in addition to estimates of reliability. These methodological design features should improve reliability. We found that all global attentional bias scores derived from the dot probe task demonstrated unacceptable reliability across the sample as a whole, and no specific subsample of participants showed acceptable reliability or stability, whether defined as internal consistency, split-half reliability, or test-retest stability. Furthermore, we were unable to show acceptable test-retest stability for the entire task or the first 100 or 200 trials of the task. A summary of these results is given in Table 1, and the supplemental material provides detailed information about the tests conducted. As can be seen from Table 1, in many cases estimates of internal consistency, split-half reliability, and test-retest stability were negative; in most cases not even the upper limit of the 95% confidence interval approached good reliability. This was also true when we focused on the index that appeared most promising in Price et al. (2015): Although by some indices that index showed improved reliability and stability (with some estimates approaching .45), in no case was reliability good; indeed, in many cases these estimates were also negative.

Potential Solutions for the Problems of Poor Reliability

We hope it is clear from the above that the global attentional bias score shows poor reliability and that this fact raises concerns for its use in either RDoC- or target-engagement-related research. Our impression is that the score is not alone in this predicament, and that psychopathology researchers would benefit from measures that have been improved to the point of having excellent reliability. We will evaluate below three potential classes of solution to this issue.

Solution 1: Solve common problems of reliability

Under the broad topic of avoiding common problems of reliability, researchers have already tried multiple strategies. These have included varying test length, because longer measures generally produce better reliability unless they exhaust the patience of participants (cf. Nunnally and Bernstein, 1994). Researchers have also sought better items (i.e., better stimuli—in our case, stimuli participants found very negative) with no resulting demonstration of excellent reliability. A remaining common reliability problem is the use of *difference scores*: The fact that the global attentional bias score is a difference score is one reliability bugbear that has thus far proven unavoidable. We have not described in detail how the global attentional bias score is actually calculated, but it is important to do so for the remainder of our discussion. Global attentional bias scores consist of a linear composite of four mean (or, in some cases, median) reaction times that differ based on the position of the probe and whether trials are congruent or incongruent. Figure 2 demonstrates how the task unfolds and may be helpful in interpreting the following description.

As shown in Figure 2, the participant must respond to the probe, and as such the probe is in different position in different trials; for example, in our data the probe was either toward the top or bottom of the screen as demonstrated in Figure 2. In addition, the probe may replace the threatening stimulus, therefore being *congruent* with threat, or may replace the nonthreatening stimulus, therefore being *incongruent* with threat. In Figure 2, if we take the word *threat* as a threat, the trial depicted is congruent: The probe replaces the threatening

word. If the probe had replaced the neutral word (*strand*) instead, it would be incongruent. The global score involves averaging response times to four trial types: The four combinations of the probe's location and congruence versus incongruence. When response times are faster, on the average, to congruent trials, the participant is said to have an attentional bias toward threat.

The global attentional bias score is therefore a difference score. As Nunnally and Bernstein (1994; p. 270) note, difference scores are *linear composites* in which one element is subtracted from another. Linear composites more generally include scores created by adding together multiple measures; a difference score involves the same (i.e., linear) process, but subtraction instead of addition. Linear composites of any type possess reliability based on the individual scores, plus the level of correlation between the two scores. Thus, *adding* two measures of depression together will typically produce higher reliability than either of the two measures possesses individually, because their individual reliability is compounded by their correlation with each other (i.e., when the correct reliability formula is applied, which is notably not Cronbach's alpha; cf. Nunnally & Bernstein, 1994). When two measures are *subtracted*, their individual reliability acts in favor of overall reliability, but their degree of positive correlation acts *against* overall reliability. That is, subtracting one depression score from another will typically produce a difference score that is less reliable than either score on its own. Accordingly, many researchers are familiar with the notion that difference scores possess questionable reliability in many typical cases.

Consider, for example, the reliability estimate for the subtraction of one reaction time score from another, when the split-half reliability of each reaction time score is .45, and the correlation between the two reaction time scores is the somewhat lower .25. We will assume that the scores have been transformed into *z* scores only because it makes the calculation simpler. The reliability of the composite would be .27, worse than the reliability of each reaction time. However, internal consistency of .95 for each reaction time score and a correlation of .80 between those scores would produce good reliability (.75), and any correlation between the scores lower than .80 would increase that estimate. This general pattern can be extrapolated to any case in which there are two measures with certain reliabilities and a given correlation between them, and we present the overall impact on reliability of these factors in Figure 3.[2] The R code used to generate the figure is available in the supplemental material.

As can be seen from the figure, it is not difficult to produce reliable difference scores when correlations between measures are low. Take the upper left panel, for example: If the internal consistency of both measure is .80 or higher, reliability of the composite will be quite good. As the correlation increases, however, the need for both individual measures to be reliable also increases. Finally, as depicted in the bottom right panel, it becomes impossible to produce reliable difference scores from measures when their high internal consistencies are matched by a high positive correlation between measures. (Note, however, that, as implied above, if the correlation between measures is .90, reliability of the composite can still be

[2]We are indebted to Scott Baldwin, who supplied the initial R code for creating this figure.

achieved, but only if the individual measures demonstrate reliability > .90; such values are not shown in the figure.)

The important point is that *difference scores are not inherently unreliable, but they do demand high reliability of individual measures relative to the correlation between the measures*: Ideally, the correlation between the measures would be very negative (or at least far from very positive) and the individual measures would each have high reliability. Because correlations between similar measures are rarely very negative, one might therefore wonder whether it would be better to find a way to compute bias scores without involving a difference score.

We started this section with the note that difference scores are a common problem for reliability, which makes it reasonable to seek an attentional bias score that does not involve computing a difference. However, although we fully support solving common problems for reliability, we can state with some confidence that *difference scores are not the essential problem with the global attentional bias score.* On the surface, this statement might seem at odds with the above. Shouldn't the fact that typical bias scores involve subtraction mean they must have poor reliability? The answer is no, not if either (a) the scores being subtracted do not correlate too strongly in a positive direction or (b) if the scores being subtracted each have excellent reliability to begin with.

The typical theory regarding attentional bias implies both (a) and (b). The proposition is that people with bias toward threat will tend to respond more quickly when a probe replaces a threatening stimulus, and less quickly when the probe replaces a nonthreatening stimulus (i.e., when the threat is present elsewhere on the screen). If such attentional bias is present and consistent, this should mean that participants should be faster at responding to threat *to the extent that they are slower to respond to nonthreatening stimuli when threatening stimuli are also present*. That is, the correlation between the two types of scores should be low, and perhaps even negative, at least to the extent that attentional bias causes differences in response times. However, in all the data we are familiar with, the correlation is not negative but is instead positive, and typically at least moderate, which contradicts the underlying theory. In our data from Study 1 in the supplementary material, the four reaction times that contribute to the traditional bias score correlated highly and positively at each time point. In fact, when those four values at each time point are included in a correlation matrix, the *smallest* correlation (across time points) was .82. Referring to Figure 3 shows clearly that good reliability will be hard to come by for a difference score computed from these data. In the data from Study 2 from the supplementary material, the correlations (from our neutral condition, which represents the typical dot probe task) had a wider range, but remained positive; the correlations that would most negatively impact the reliability of the difference score showed one low correlation ($r = .12$, $p = .636$) and three correlations that ranged from moderate to very high (rs ranging from .45 to .95, ps < .06). This situation is less dire than that seen in Study 1, but still not ideal for computing a difference score.

This point deserves particular emphasis. The fact that the attentional bias formula contains a subtraction should not, *in theory*, lead to poor reliability *because the attentional bias theory typically implies stability of bias, which in turn implies a weak or negative correlation*. The

fact that the resulting correlation is not negative and not typically weak (and is instead typically at least moderate and frequently quite high) suggests problems not only for the attentional bias score's reliability, but also for the notion that attentional bias toward threat is stable.

Empirical demonstration of the effect of sustained attentional bias on reliability—The mathematical demonstration above would obviously be more convincing with empirical results. We therefore set out to test the implication that consistent attentional bias would produce global scores with acceptable reliability. Study 2 in the supplementary material demonstrated that the presence of participants with a directed attentional bias affects reliability of global attentional bias scores. The full procedure is described in the supplemental material, but, to briefly recap, we asked participants to either complete a dot-probe task under standard instructions (n = 19), purposefully attend to negative faces (n = 20), or purposefully avoid negative faces (n = 22). Our data show that under standard instructions, the global score shows the poor split-half reliability that would be expected given the data reviewed above (.14). In contrast, within both groups that were instructed to respond consistently to the negative faces, reliability was good (split-half reliability > .81). Thus, when participants showing a consistent, instructed bias were included, reliability was good. As shown in the supplemental material, this improved reliability is accompanied by a decrease in the overall positive correlation between measures *and* an increase in the reliability of the individual measures.

Importantly, the global attentional bias score showed superior reliability when conditions mimicked the situation proposed by the original theory: Consistency of attentional bias across trials. This outcome follows mathematically from the fact that the presence of consistent bias produces fewer threats to the reliability of a difference score. The fact that such scores are not typically reliable implies problems with any theory of *consistent* attentional bias toward threat. In that case, the problem would not be the difference score, but something about the assumption that bias toward threat is either present at all or consistent across time.

Potential solution 2: Consider that one's theory may be incorrect

Broadly speaking, after eliminating all known threats to reliability, the remaining possibility, however unpleasant, must be the truth: Some element of the proposed relationship between behavior and measure is actually incorrect. In the case of the global attentional bias scores, the above argument suggests that the notion that attentional bias is stable may be incorrect.

Indeed, this possibility has already been argued by Zvielli and colleagues (2015). They proposed that that attentional bias may be better understood as a *process expressed in time*. Specifically, they proposed that attentional bias is not static, but a dynamic process often expressed in fluctuating, phasic bursts, often towards *and* then away from motivationally-relevant stimuli over time. Accordingly, the authors proposed a novel computational procedure—Trial-Level Bias Signal (TL-BS) scores—designed to estimate attentional bias as a dynamic process in time using existing experimental task data (e.g., obtained via the dot probe). TL-BS scores are described in detail by Zvielli and colleagues, but, in brief, they are

computed through a type of running-window calculation, repeatedly estimating trial-level attentional bias levels and direction (towards, away) by subtracting temporally contiguous pairs of trial types (e.g., congruent and incongruent) response times (RTs). Notably, the TL-BS scores are still difference scores, but they are calculated in a manner that accounts for the possibility that attentional bias unfolds, from moment-to-moment, *across time*. Zvielli et al. (2015) initially computed indices for five theoretically-interpretable features of the TL-BS scores. These include indices of mean and peak levels of attentional bias towards and attentional bias away from target stimuli (e.g., threat), as well as overall temporal variability of bias.

TL-BS reliability in our data—In our data, described at length in Study 1 of the supplementary material, these TL-BS indices demonstrated good reliability overall (e.g., the majority of indices ≥ .65 and many > .85). As reported in Table 2, TL-BS indices showed good to excellent internal consistency and stability for the most part, with the exception of the peak TL-BS scores, which showed good internal consistency (αs > .75) but only modest stability (ICCs < .50). Because the TL-BS indices (particularly the mean toward, mean away, and variability indices) provide good to excellent reliability, we submit that the proposed dynamic process perspective on attentional bias provides one plausible way forward for the field that requires more study. Although we focus on reliability here, it should be noted that multiple studies have demonstrated multiple forms of validity of TL-BS scores as well (Amir et al., in press; Schäfer et al., in press; Yuval, Zvielli, & Bernstein, in press; Zvielli et al., 2015).

Issues with TL-BS scores—Of course, TL-BS scores are not a panacea. In our Study 1 data, we observed high inter-correlations between the features of bias temporal dynamics (TL-BS parameters), making it unclear whether the indices can or should be interpreted separately in our sample. Whether this is a problem or a feature of the TL-BS indices depends on whether or not the correlation of the indices is a faithful representation of the underlying phenomena. Zvielli et al (2015) described various limitations in the application of the TL-BS. For example, certain design features of tasks, such as presentation of multiple different stimulus durations or multiple stimulus conditions within a block limit capacity to faithfully estimate bias at the trial-level. Perhaps such issues should not be surprising given that the task was not developed with TL-BS scores in mind.

Nor should we look to a single panacea in studying any complex psychological phenomenon. The finding that TL-BS scores show better reliability is suggestive that attentional bias is a dynamic process and not a stable trait. Our better understanding of the phenomena under study will ideally lead to new tasks (and new indices for old tasks) that are specifically designed to capture this element of attention over time. Alternative hypotheses that conflict with the assumptions behind TL-BS scores should, in turn, lead to alternative tasks and substantive tests between competing theories. For example, the fact that the global (aggregated mean) bias score typically shows poor reliability does not eliminate the possibility that there might be other ways of calculating a global bias score that is reliable. In our view, TL-BS scores for the dot-probe are just one part of what must be a broader effort to articulate precisely how behavior and scores on attention tasks are related: We agree with

Borsboom's (2006) assertion that, ideally, the development of a model of how scores relate to attributes would be the first step in developing a new measure, not an afterthought.

Solution 3: Investigate alternative methods (i.e., "No, not the dot probe!")

We submit that consistent findings of poor reliability should lead not only to careful reflection on the theory of the phenomenon studied and how that relates to the measure, but should also lead to consideration of whether an entirely new measure is needed. Indeed, many researchers who are aware of the reliability issues for the global attentional bias score have suggested moving to other tasks.

However, thus far the reality has been less encouraging in terms of reliability estimates. We will focus on eye tracking indices as one example, although there are certainly several that have been proposed. Waechter and colleagues (2014) found that eye movement indices assessed over an entire five second trial had good reliability, but noted that reliability for more commonly used indices (involving shorter periods of time) were much less reliable (low to modest reliability). Similarly, Price et al. (2015) examined eye tracking indices and found modest reliability of single sessions at best; for averages of two administrations, they found good, although still not excellent reliability (around .70) for some indices. To be clear, modest or acceptable reliability is better than low reliability; the point is that without further improvement it appears unlikely that these indices will be useful for tests focusing on mediation or selecting individuals. Amir et al. (in press) found TL-BS parameters computed for eye-tracking demonstrated moderate to high levels of split-half reliability (e.g., all above .53). In contrast, global eye tracking measures (i.e., the ones typically used), showed poor split-half reliability (e.g., all below .40). It seems plausible that assessments using eye tracking could also benefit from attending to the dynamic expression of bias over time.

In our experience, when researchers discover the literature regarding poor reliability of the global attentional bias score, they often consider moving to a measure that appears more objective or seems more biological, possibly under the assumption that removing subjectivity of response will increase reliability. At the risk of putting too fine a point on it, we want to emphasize the fact that no amount of apparent objectivity of a measure absolves us from grappling with the possibility that our theories or choice of measure may simply be mistaken. Apparently objective measures might seem to assess alluringly *real* qualities and therefore not require proof of reliability. Our experience is that those who work extensively with such objective measures are well aware that this is not the case. For example, one might hope that moving directly to neuroimaging tasks might side-step reliability issues in behavioral tasks, but, instead, imaging experts assert that reliability is just as essential to the validity of imaging task data as any other measure (Barch & Mathalon, 2011; Barch & Yarkoni, 2013). Indeed, imaging tasks can initially show poor reliability that can be improved, but only when researchers assess reliability and focus on maximizing it: Friedman and colleagues (2008) describe assessing and finding ways to improve (initially poor) reliability across research sites of an imaging study.

As may be obvious, the fact that measures that are often thought of as more objective also have reliability challenges is part of the reason that our concerns about reliability are strongly related to RDoC. None of the methods of assessment described in RDoC are

immune to concerns about reliability of measurement. No measure yet developed for psychopathology, whether it involves paper and pencil, electrodes, chemicals, or large magnets, lacks human input on at least some aspect of the assessment process and are therefore subject to questions of reliability. It may be useful to note that the statistics involved in standard reliability indices *originated in attempts to deal with errors in measuring objective qualities*, such as the position of planets (see Borsboom, 2005, for a review). If anything, the classical test theory concept of reliability is *more* suited to issues regarding the measurement of objective qualities than subjective self-report.

Recommendations

Regarding the dot probe task as a measure of attention bias, our recommendation is not to simply abandon it. Instead, we propose that researchers investigate the nature of attentional processing of emotional information across a variety of tasks, with careful attention to what the reliability of those measures tell us about our underlying theories of attention in the context of emotional material. For example, the TL-BS scores could be reported alongside the traditional global attentional bias score, with reliability given for all indices, including novel indices derived from the dot probe and new scores from novel tasks.

Our recommendations run broader than the dot-probe task or attentional bias alone. We have framed our discussion partially in regard to RDoC, and the RDoC initiative emphasizes the use of multiple measures: We strongly advocate for the use of multiple measures at multiple levels of analysis, with reliability consistently reported for all measures. We advocate for this in part because of familiar axioms regarding the fact that poor reliability limits the validity of measures. However, we also hope that researchers will not simply consider reliability as a technical threshold that must be passed, but also potentially as one means to evaluate the validity of the underlying theory. As suggested by Borsboom, Mellenbergh, and van Heerden (2005), it is nonsensical to speak of reliably measuring something that is not being measured. If a measure is not reliable and cannot be made reliable through familiar means, despite the fact that the measure corresponds well with what is theoretically being measured, it may well be that the lack of reliability suggests that the underlying theory is in need of revision. This is precisely the point argued by Zvielli et al (2015) regarding the dynamic process perspective on attentional bias. Careful consideration of what is being measured, such that the quantification of a measure is consistent with the studied phenomenon, is essential to produce scores that have meaningful reliability (as well as validity, for that matter; cf. Borsboom, 2006; Borsboom et al., 2005).

We hope we have been persuasive that issues of reliability are crucial to the study of psychopathology as a whole. Unreliability need not be only threatening: Instead, it can be a spur to refining theory and developing measures that comport well with theory. Indeed, we look forward to the future of attentional bias research with considerably more optimism than when we started writing this paper, and hope to see additional research that reports reliability for a variety of measures, at least some of which capture the potentially dynamic nature of the phenomenon. More broadly, we encourage researchers to test and report reliability of their tasks and measures, fearlessly, in an attempt to overcome the threat of unreliability to the entire field of psychopathology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was funded in part by MH070547 and the Taylor Family Institute for Innovative Psychiatric Research to Eric Lenze, MH090308 to Thomas Rodebaugh, a grant from the Milton Rosenbaum Foundation to Jonathan Huppert, and UL1 RR024992 to Washington University. Development of the MacBrain Face Stimulus Set was overseen by Nim Tottenham and supported by the John D. and Catherine T. MacArthur Foundation Research Network on Early Experience and Brain Development. Please contact Nim Tottenham at tott0006@tc.umn.edu for more information concerning the stimulus set. The authors have no competing interests in this research. Thanks to David Balota for providing some guidance regarding reaction time reliability, as well as to Deanna Barch for insight regarding neuroimaging and reliability.

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders: DSM-5. 5. Washington, D.C: American Psychiatric Association; 2013.

Amir I, Zvielli A, Bernstein A. (De)coupling of covert-overt attentional bias dynamics. Emotion. in press.

Amir N, Beard C, Burns M, Bomyea J. Attention modification program in individuals with generalized anxiety disorder. Journal of Abnormal Psychology. 2009; 118:28–33. [PubMed: 19222311]

Amir N, Beard C, Taylor CT, Klumpp H, Elias J, Burns M, Chen X. Attention training in individuals with generalized social phobia: A randomized controlled trial. Journal of Consulting and Clinical Psychology. 2009; 77:961–973. DOI: 10.1037/a0016685 [PubMed: 19803575]

Balota DA, Yap MJ, Cortese MJ, Hutchison KA, Kessler B, Loftis B, … Treiman R. The English Lexicon project. Behavior Research Methods. 2007; 39:445–459. [PubMed: 17958156]

Bar-Haim Y, Holoshitz Y, Eldar S, Frenkel TI, Muller D, Charney DS, … Wald I. Life-threatening danger and suppression of attention bias to threat. The American Journal of Psychiatry. 2010; 167:694–698. DOI: 10.1176/appi.ajp.2009.09070956 [PubMed: 20395400]

Bar-Haim Y, Lamy D, Pergamin L, Bakermans-Kranenburg MJ, van Ijzendoorn MH. Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. Psychological Bulletin. 2007; 133:1–24. [PubMed: 17201568]

Barch DM, Mathalon DH. Using brain imaging measures in studies of procognitive pharmacologic agents in schizophrenia: psychometric and quality assurance considerations. Biological Psychiatry. 2011; 70:13–18. DOI: 10.1016/j.biopsych.2011.01.004 [PubMed: 21334602]

Barch DM, Yarkoni T. Introduction to the special issue on reliability and replication in cognitive and affective neuroscience research. Cognitive, Affective & Behavioral Neuroscience. 2013; 13:687–689. DOI: 10.3758/s13415-013-0201-7

Borsboom, D. Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. New York, NY: Cambridge University Press; 2005.

Borsboom D. The attack of the psychometricians. Psychometrika. 2006; 71:425–440. DOI: 10.1007/s11336-006-1447-6 [PubMed: 19946599]

Borsboom D, Mellenbergh GJ, van Heerden J. The Concept of Validity. Psychological Review. 2004; 111:1061–1071. DOI: 10.1037/0033-295x.111.4.1061 [PubMed: 15482073]

Britton JC, Bar-Haim Y, Clementi MA, Sankin LS, Chen G, Shechner T, … Pine DS. Training-associated changes and stability of attention bias in youth: Implications for Attention Bias Modification Treatment for pediatric anxiety. Developmental Cognitive Neuroscience. 2013; 4:52–64. DOI: 10.1016/j.dcn.2012.11.001 [PubMed: 23200784]

Calamaras MR, Tone EB, Anderson PL. A pilot study of attention bias subtypes: Examining their relation to cognitive bias and their change following cognitive behavioral therapy. Journal of Clinical Psychology. 2012; 68:745–754. DOI: 10.1002/jclp.21875 [PubMed: 22610950]

Clarke PJF, Notebaert L, MacLeod C. Absence of evidence or evidence of absence: Reflecting on therapeutic implementations of attentional bias modification. BMC Psychiatry. 2014; 14doi: 10.1186/1471-244x-14-8

Cole DA, Maxwell SE. Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. Journal of Abnormal Psychology. 2003; 112:558–577. DOI: 10.1037/0021-843x.112.4.558 [PubMed: 14674869]

Cuthbert BN, Kozak MJ. Constructing constructs for psychopathology: The NIMH research domain criteria. Journal of Abnormal Psychology. 2013; 122:928–937. DOI: 10.1037/a0034028 [PubMed: 24016027]

Dear BF, Sharpe L, Nicholas MK, Refshauge K. The psychometric properties of the dot-probe paradigm when used in pain-related attentional bias research. The Journal of Pain. 2011; 12:1247–1254. DOI: 10.1016/j.jpain.2011.07.003 [PubMed: 21982721]

Engstrom EJ, Kendler KS. Emil Kraepelin: Icon and reality. American Journal of Psychiatry. in press.

Enock PM, Hofmann SG, McNally RJ. Attention bias modification training via smartphone to reduce social anxiety: A randomized, controlled, multi-session experiement. Cognitive Therapy and Research. 2014; 38:200–216. DOI: 10.1007/s10608-014-9606-z

Faul F, Erdfelder E, Buchner A, Lang AG. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behavior Research Methods. 2009; 41:1149–1160. DOI: 10.3758/BRM.41.4.1149 [PubMed: 19897823]

First, MB.; Spitzer, RL.; Gibbon, M.; Williams, J. Structured Clinical Interview for DSM-IV Axis I Disorders-Patient Edition (SCID-I/P, Version 2.0). New York: Authors; 1996.

Frenkel, TI.; Bar-Haim, Y. Israeli Database of Emotional Facial Expressions. Department of Psychology, Tel Aviv University; 2006.

Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover GH, … Potkin SG. Test-retest and between-site reliability in a multicenter fMRI study. Human Brain Mapping. 2008; 29:958–972. DOI: 10.1002/hbm.20440 [PubMed: 17636563]

Harvey, AG.; Watkins, E.; Mansell, W. Cognitive behavioural processes across psychological disorders. New York: Oxford University Press; 2004.

Hazen RA, Vasey MW, Schmidt NB. Attentional retraining: A randomized clinical trial for pathological worry. Journal of Psychiatric Research. 2009; 43:627–633. DOI: 10.1016/j.jpsychires.2008.07.004 [PubMed: 18722627]

Hirsch CR, MacLeod C, Mathews A, Sandher O, Siyani A, Hayes S. The contribution of attentional bias to worry: Distinguishing the roles of selective engagement and disengagement. Journal of Anxiety Disorders. 2011; 25:272–277. DOI: 10.1016/j.janxdis.2010.09.013 [PubMed: 20980126]

Hoyle, RH.; Kenny, DA. Sample size, reliability, and tests of statistical mediation. In: Hoyle, RH., editor. Statistical strategies for small sample research. Thousand Oaks, CA: Sage; 1999. p. 195-222.

Hoyle, RH.; Robinson, JC. Mediated and moderated effects in social psychological research. In: Sansone, C.; Morf, CC.; Panter, AT., editors. Handbook of Methods in Social Psychology. Thousand Oaks, CA: Sage; 2004. p. 213-233.

Lund K, Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments & Computers. 1996; 28:203–208.

Joormann J, Talbot L, Gotlib IH. Biased processing of emotional information in girls at risk for depression. Journal of Abnormal Psychology. 2007; 116:135–143. DOI: 10.1037/0021-843x.116.1.135 [PubMed: 17324024]

Kappenman ES, Farrens JL, Luck SJ, Proudfit GH. Behavioral and ERP measures of attentional bias to threat in the dot-probe task: Poor reliability and lack of correlation with anxiety. Frontiers in Psychology. 2014; 5 ArtID: 1368.

Kihlstrom, JF. To honor Kraepelin...: From symptoms to pathology in the diagnosis of mental illness. In: Beutler, LE.; Malik, ML., editors. Decade of behavior. Washington, DC: American Psychological Association; 2002. p. 279-303.p. xiiip. 331

Krebs G, Hirsch CR, Mathews A. The effect of attention modification with explicit vs. minimal instructions on worry. Behaviour Research and Therapy. 2010; 48:251–256. DOI: 10.1016/j.brat.2009.10.009 [PubMed: 19926075]

MacLeod C, Clarke PJF. The attentional bias modification approach to anxiety intervention. Clinical Psychological Science. 2015; 3:58–78. DOI: 10.1177/2167702614560749

MacLeod C, Mathews A. Anxiety and the allocation of attention to threat. The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology. 1988; 40(4-A):653–670. [PubMed: 3212208]

Mathews A, MacLeod C. Selective processing of threat cues in anxiety states. Behaviour Research and Therapy. 1985; 23:563–569. DOI: 10.1016/0005-7967(85)90104-4 [PubMed: 4051929]

Mathews A, MacLeod C. Cognitive approaches to emotion and emotional disorders. Annual Review of Psychology. 1994; 45:25–50. DOI: 10.1146/annurev.ps.45.020194.000325

McNally RJ. Automaticity and the anxiety disorders. Behaviour Research and Therapy. 1995; 33:747–754. DOI: 10.1016/0005-7967(95)00015-p [PubMed: 7677712]

Nunnally, JD.; Bernstein, IH. Psychometric theory. 3. New York: McGraw-Hill; 1994.

Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, Schacht AL. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature reviews Drug discovery. 2010; 9:203–214. DOI: 10.1038/nrd3078

Pencina MJ, Peterson ED. Moving from clinical trials ot precision medicine: The role for predictive modeling. Journal of the American Medical Association. 2016; 315:1713–1714. [PubMed: 27115375]

Perez-Edgar K, Bar-Haim Y, McDermott JM, Gorodetsky E, Hodgkinson CA, Goldman D, … Fox NA. Variations in the serotonin-transporter gene are associated with attention bias patterns to positive and negative emotion faces. Biological Psychology. 2010; 83:269–271. DOI: 10.1016/j.biopsycho.2009.08.009 [PubMed: 19723555]

Pergamin-Hight L, Bakermans-Kranenburg MJ, van Ijzendoorn MH, Bar-Haim Y. Variations in the promoter region of the serotonin transporter gene and biased attention for emotional information: A meta-analysis. Biological Psychiatry. 2012; 71:373–379. DOI: 10.1016/j.biopsych.2011.10.030 [PubMed: 22138391]

Price RB, Kuckertz JM, Siegle GJ, Ladouceur CD, Silk JS, Ryan ND, … Amir N. Empirical recommendations for improving the stability of the dot-probe task in clinical research. Psychological Assessment. 2015; 27:365–376. DOI: 10.1037/pas0000036 [PubMed: 25419646]

R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015. https://www.R-project.org/

Schäfer J, Bernstein A, Zvielli A, Höfler M, Wittchen HU, Schönfeld S. Attentional bias dynamics predict posttraumatic stress symptoms: A prospective longitudinal study among soldiers. Depression and Anxiety. in press.

Schmukle SC. Unreliability of the dot probe task. European Journal of Personality. 2005; 19:595–605. DOI: 10.1002/per.554

Simon GM, Niphakis MJ, Cravatt BF. Determining target engagement in living systems. Nature Chemical Biology. 2013; 9:200–205. DOI: 10.1038/nchembio.1211 [PubMed: 23508173]

Staugaard SR. Reliability of two versions of the dot-probe task using photographic faces. Psychology Science. 2009; 51:339–350.

Van Bockstaele B, Verschuere B, Tibboel H, De Houwer J, Crombez G, Koster EHW. A review of current evidence for the causal impact of attentional bias on fear and anxiety. Psychological Bulletin. 2014; 140:682–721. DOI: 10.1037/a003483 [PubMed: 24188418]

Waechter S, Nelson AL, Wright C, Hyatt A, Oakman J. Measuring attentional bias to threat: Reliability of dot probe and eye movement indices. Cognitive Therapy and Research. 2014; 38:313–333. DOI: 10.1007/s10608-013-9588-2

Waechter S, Stolz JA. Trait anxiety, state anxiety, and attenional bias to threat: Assessing the psychometric properties of response time measures. Cognitive Therapy and Research. 2015; 39:441–458. DOI: 10.1007/s10608-015-9670-z

Wendland JR, Martin BJ, Kruse MR, Lesch KP, Murphy DL. Simultaneous genotyping of four functional loci of human SLC6A4, with a reappraisal of 5-HTTLPR and rs25531. Molecular Psychiatry. 2006; 11:224–226. DOI: 10.1038/sj.mp.4001789 [PubMed: 16402131]

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

Yuval K, Zvielli A, Bernstein A. Attentional bias dynamics and posttraumatic stress in survivors of violent conflict and atrocities: New directions in clinical psychological science of refugee mental health. Clinical Psychological Science. in press.

Zvielli A, Bernstein A, Koster EHW. Temporal dynamics of attention bias. Clinical Psychological Science. 2015; 3:772–788.

General Scientific Summary

Researchers often fail to report whether their measurements related to mental disorders give precise and consistent measurements even when what they measure should be consistent and stable over time. We focus on one such measure, the global attentional bias score, to illustrate the problems caused by poor reliability, and moreover how assessing and improving the reliability of our measures can improve our theories and ultimately advance our knowledge.

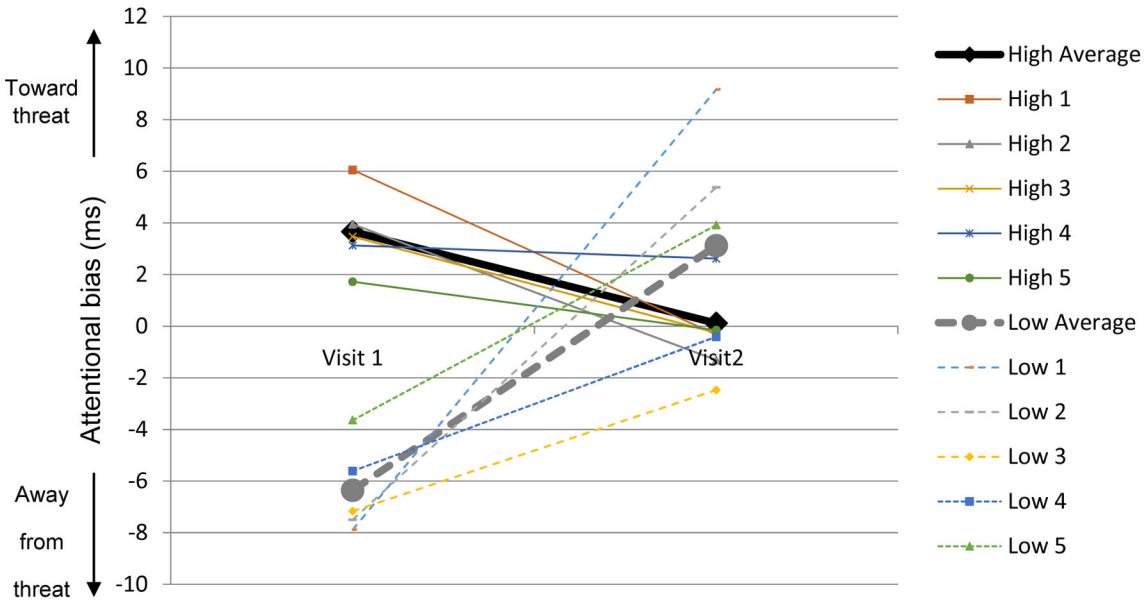

Figure 1.

Demonstration of what would occur if participants in our Study 1 were selected based on extreme bias scores at Visit 1. Shown are the participants with the five highest and five lowest global attentional bias scores in Study 1 of the supplemental material, based on Visit 1, and their global attention bias scores at both visits. Thin lines represent individual cases; thick lines represent the average within each group. Arguably due to low stability of the global attentional bias score, participants show a regression to the mean overall (all high scores and all low scores trend toward zero, with some flipping sign); Calamaras et al. (2012) showed the same pattern and speculated it might be due to treatment, but here participants received no treatment.

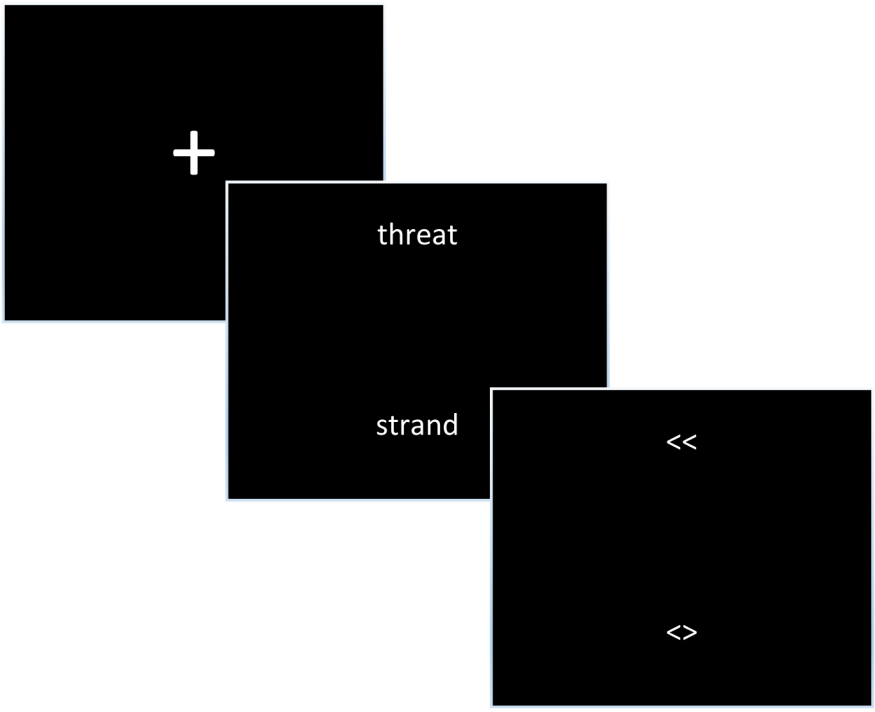

Figure 2.

General dot probe task example. A fixation cross (upper left) is replaced by two stimuli, here toward the top and bottom of the screen. The stimuli are then replaced by the probe. Here, the participant must find the probe that represents two arrows pointing in the same direction and respond with the left button. Among the many variations in the literature include the stimuli (e.g., faces, among other pictures), the type of probe (e.g., E versus F; a simple dot), and whether participants have to discriminate between the probe and a distractor (here, ≪ is the probe and <> is a distractor; in some studies there is no distractor).

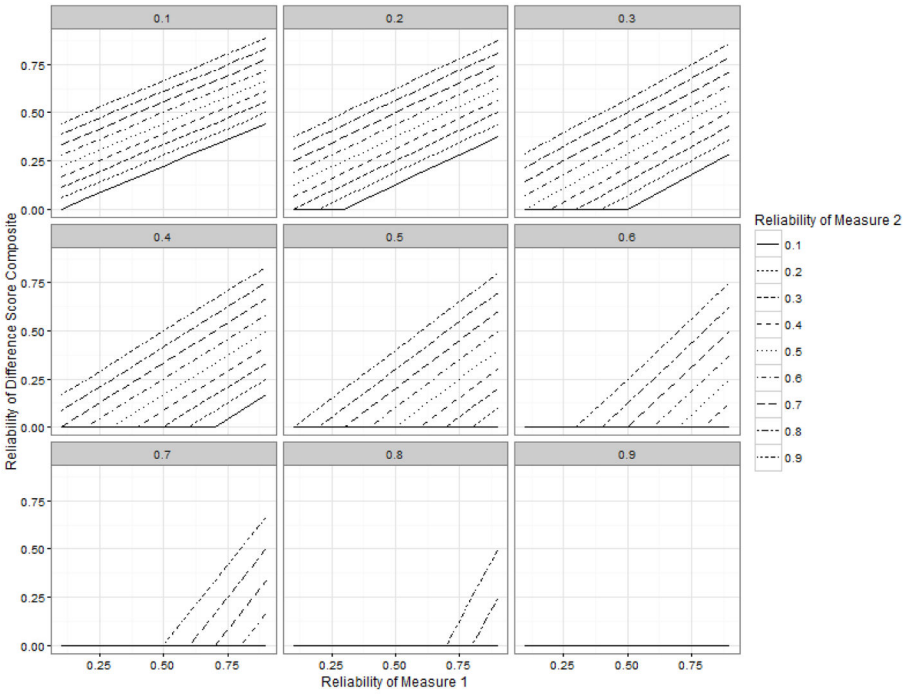

Figure 3.

Demonstration of the impact of a positive correlation between measures on a difference score composite. As in the text, it is assumed that the measures are standardized (i.e., *z* scores). Each panel represents a correlation between the two measures, from .10 to .90. The reliability of the hypothetical measure 1 is given along the x axis, and the reliability of a hypothetical measure 2 is given by line type. The top line in each panel is always a reliability of .90, with descending line types representing .80, .70, etc. This is because when measure 2 has a reliability of .9, the reliability of the composite is always higher than when measure 2 has a reliability of .8, etc. The reliability of a resulting difference score is given along the y axis. In some cases some (or all) line types do not appear because the reliability of the difference score would be estimated to be at or below zero. This plot was created with the R package ggplot2 (Wickham, 2009) with the help of Scott Baldwin.

Table 1

Reliability for Traditional Global Bias Scores from Study 1

Test and Data	Overall Anxiety Disorders Sample (*n* =24)	Highest Subgroup (Group Name)
Test-Retest Stability		
100 Trials (Price Method)	ICC = −.12 (CI: −.51, .29)	ICC = .13, (CI = −.38, .56) (Anxiety disorder, no current mood disorder)
200 Trials (Price Method)	ICC = .08 (CI: −.35, .47)	ICC = .18, (CI = −.32, .59) (Anxiety, no current mood disorder)
First Half (Price Method)	ICC = −.13 (CI: −.51, .29)	ICC = −.04, (CI = −.60, .50) (Panic disorder)
Second Half (Price Method)	ICC = −.39 (CI: −.70, .03)	ICC = −.38, (CI = −.76, .18) (Generalized social anxiety disorder)
All Trials (Price Method)	ICC = −.36 (CI: −.70, .06)	ICC = .13, (CI = −.46, .62) (Panic disorder)
All Trials (Bias for Each Word)	ICC = −.30 (CI: −.60, .10)	ICC = .08, (CI = −.43, .56) (Panic disorder)
All Trials (Most Negative Words)	ICC = .06 (CI: −.30, .43)	ICC = .47, (CI = −.003, .79) (Panic disorder)
Internal Consistency		
Visit 1 (Bias for Each Word)	α = −.32 (CI: −1.20, .33)	α = −.25 (CI: −1.39, .52) (Generalized social anxiety disorder)
Visit 2 (Bias for Each Word)	α = −.48 (CI: −1.46, .25)	α = −.37 (CI: −1.62, .48) (Generalized social anxiety disorder)
Split Half Reliability		
Visit 1 (Bias for Each Word)	*r* = −.38 (CI: −.52, .25)	*r* = −.16 (CI: −2.10, .57) (Anxiety, no current mood disorder)
Visit 2 (Bias for Each Word)	*r* = −.06 (CI: −1.46, .54)	*r* = .32 (CI: −.79, .75) (Anxiety, no current mood disorder)
Visit 1 (Price Method)	*r* = .29 (CI: −.64, .69)	*r* = .42 (CI: −.21, .65) (Anxiety disorder, no current mood disorder)
Visit 2 (Price Method)	*r* = −.58 (CI: −2.66, .32)	*r* = −.36 (CI: −3.24, .56)

Note. In many cases, negative correlations mean that the assumptions of the reliability test were violated, resulting in numbers within confidence intervals that are outside the theoretical parameters. The confidence intervals are presented as computed to provide the clearest picture of the data. ICC = Intraclass Correlation Coefficient (two-way random for absolute agreement, single measure); CI = 95% confidence interval. Subgroups tested for reliability included all participants, participants with an anxiety disorder with no current mood disorder, participants with panic disorder with or without agoraphobia and agoraphobia without panic disorder, and participants with generalized social anxiety disorder. Each method of computing the bias score is described in full in the supplemental material. Further information is available in Study 1 of the supplemental material.

Table 2

Full Reliability for Trial Level Bias Scores from Study 1

Reliability Type	Mean Toward	Mean Away	Peak Toward	Peak Away	Variability
Test-Retest Stability					
Block 1	ICC = .86 (CI: .48, .95)	ICC = .68 (CI: .40, .85)	ICC = .49 (CI: .14, .74)	ICC = .20 (CI: −.23, .56)	ICC = .90 (CI: .74, .96)
Block 4	ICC = .73 (CI: .48, .88)	ICC = .74 (CI: .49, .88)	ICC = .47 (CI: .11, .73)	ICC = .15 (CI: −.25, .51)	ICC = .79 (CI: .57, .90)
First 152 trials	ICC = .66 (CI: .36, .84)	ICC = .51 (CI: .14, .75)	ICC = .26 (CI: −.15, .60)	ICC = .22 (CI: −.21, .57)	ICC = .83 (CI: .65, .92)
Last 152 trials	ICC = .64 (CI: .34, .83)	ICC = .54 (CI: .18, .78)	ICC = .37 (CI: −.01, .66)	ICC = .16 (CI: −.27, .53)	ICC = .69 (CI: .41, .85)
Internal Consistency					
Visit 1	α = .93 (CI: .88, .97)	α = .96 (CI: .92, .98)	α = .77 (CI: .60, .89)	α = .82 (CI: .68, .91)	α = .97 (CI: .95, .99)
Visit 2	α = .92 (CI: .87, .96)	α = .95 (CI: .91, .98)	α = .82 (CI: .69, .91)	α = .79 (CI: .63, .89)	α = .97 (CI: .94, .98)
Split-Half Reliability					
Block 1, Visit 1	*r* = .83 (CI: .61, .93)	*r* = .87 (CI: .71, .95)	*r* = .31 (CI: , 59, .70)	*r* = .56 (CI: −01, .81)	*r* = .95 (CI: .88, .98)
Block 1, Visit 2	*r* = .68 (CI: .27, .86)	*r* = .92 (CI: .81, .96)	*r* = .59 (CI: .05, .82)	*r* = .65 (CI: .19, .85)	*r* = .88 (CI: .72, .95)

Note. ICC = Intraclass Correlation Coefficient (two-way random for absolute agreement, single measure); CI = 95% confidence interval. Each method of computing the bias score is described in full in the supplemental material. Test-retest always refers to stability across study visits. Internal consistency always refers to the alpha coefficient for the split-halves of the blocks within the visit in question (i.e., the 8 split-halves of the 4 blocks per visit). Further information is available in Study 1 of the supplemental material.