CANADIAN PSYCHOLOGICAL ASSOCIATION
SOCIÉTÉ CANADIENNE DE PSYCHOLOGIE

Canadian Psychology / Psychologie canadienne

ISSN: 0708-5591

2020, Vol. 61, No. 4, 281–288
http://dx.doi.org/10.1037/cap0000236

Psychological Measurement and the Replication Crisis: Four Sacred Cows

Scott O. Lilienfeld
Emory University and University of Melbourne

Adele N. Strother
Emory University

Although there are surely multiple contributors to the replication crisis in psychology, one largely unappreciated source is a neglect of basic principles of measurement. We consider 4 sacred cows—widely shared and rarely questioned assumptions—in psychological measurement that may fuel the replicability crisis by contributing to questionable measurement practices. These 4 sacred cows are: (a) we can safely rely on the name of a measure to infer its content; (b) reliability is not a major concern for laboratory measures; (c) using measures that are difficult to collect obviates the need for large sample sizes; and (d) convergent validity data afford sufficient evidence for construct validity. For items a and d, we provide provisional data from recent psychological journals that support our assertion that such beliefs are prevalent among authors. To enhance the replicability of psychological science, researchers will need to become vigilant against erroneous assumptions regarding both the psychometric properties of their measures and the implications of these psychometric properties for their studies.

Public Significance Statement
This article outlines four widely held but erroneous measurement assumptions that may adversely affect the accuracy and replicability of psychological findings. The effects of questionable measurement practices stemming from these assumptions are discussed, and new data bearing on the prevalence of these assumptions in academic journals are presented. In addition, this article offers several potential remedies that researchers and journals can implement to improve the measurement of psychological constructs.

Keywords: psychological measurement, replication crisis, questionable measurement practices, laboratory measures, discriminant validity

The much-decried replication crisis in psychology is as much an opportunity for self-reflection and self-correction as it is for self-flagellation (Nelson, Simmons, & Simonsohn, 2018). This crisis encourages us to pause, take, a deep breath, and reconsider our standard means of doing business in psychological science (Asendorpf et al., 2013; Lilienfeld & Waldman, 2017). Over the past decade, considerable attention has been accorded to p-hacking, file drawering of negative or inconclusive results, hypothesizing after results are known, and other questionable research practices (Fiedler & Schwarz, 2016; John, Loewenstein, & Prelec, 2012; Tackett & Miller, 2019) as sources of the replication crisis. Nevertheless, with few exceptions, scant attention has been directed to another likely cause of this crisis: problematic measurement.

As Flake and Fried (2019) observed, the field of psychological science at large has been characterized by what can be described as a measurement schmeasurement attitude (see also Hughes, 2018). Many researchers pay little heed to the psychometric properties of their measures, cavalierly neglecting them or taking them for granted. Perhaps paralleling this state of not-so-benign neglect, survey data suggest that psychometric issues, including test construction and scaling, are receiving inadequate attention in psychology graduate programs (Aiken, West, & Millsap, 2008); this educational gap may be even more pronounced for graduate programs in neuroscience/biopsychology than in other psychological subdisciplines (Schwartz, Lilienfeld, Meca, & Sauvigné, 2016). Anecdotally, the first author of this article has recently heard from several young scholars on the academic job market that hiring committees in many major psychology departments have openly expressed disinterest in recruiting professors with assessment expertise.

This attitude is short sighted because questionable measurement practices (QMPs; Flake & Fried, 2019) may be a largely unappreciated contributor to replication failures in psychology. If researchers rely on measures with questionable reliability, construct validity, or both, they should not be surprised to find that their findings are erratic across samples. In this commentary, we consider four sacred cows—widely shared and rarely questioned assumptions—in the domain of psychological measurement that may fuel

This article was published Online First August 13, 2020.

Scott O. Lilienfeld, Department of Psychology, Emory University, and School of Psychological Sciences, University of Melbourne; Adele N. Strother, Department of Psychology, Emory University.

Correspondence concerning this article should be addressed to Scott O. Lilienfeld, Department of Psychology, Emory University, Room 473, 36 Eagle Row, Atlanta, GA 30322. E-mail: slilien@emory.edu

the replication crisis. Many of these assumptions may in part undergird QMPs.[1]

The lessons imparted by QMPs are hardly new (e.g., Epstein & O'Brien, 1985). At the same time, as Hegel (1823) reminded us, one of the main lessons we learn from history is that we do not learn from history. Well over 6 decades ago, Cronbach (1954) half-jokingly wrote of the alien worlds of psychometrics and clinicia, observing that many psychotherapists and clinical researchers were largely unfamiliar with the rigorous psychometric criteria demanded by measurement experts. Hence, even as the sacred cows we present may strike some readers as old news, they need to be reiterated afresh with each upcoming generation of psychological scholars. Indeed, in his closing comments, Cronbach wrote that "psychometric missions to clinicia must continue" (p. 270). We hope that our brief commentary will serve as a modest but constructive excursion in this regard.

Sacred Cow #1: We Can Safely Rely on the Name of a Measure to Infer Its Content

One of the first principles that every introductory psychology student learns about measurement is that validity is truth in advertising: A valid index is true to its name and measures what it purports to measure (Lilienfeld, Lynn, & Namy, 2018). A corollary of this principle is that to infer the meaning of a measure, we must appraise its amassed construct validity evidence by considering its convergent and discriminant correlates within a nomological network (an interlocking set of predictions regarding a construct's associations with observable measures, the association of this construct with other constructs, and the observed measures' associations with other measures; Cronbach & Meehl, 1955; Garber & Strassberg, 1991) rather than to rely on the test developer's ex cathedra pronouncements regarding its content.

Still, time and again, psychological researchers have fallen prey to the jingle fallacy, the error of assuming that two or more phenomena, such as two or more psychological measures, are identical merely because they bear the same name (Thorndike, 1904).[2] As a consequence of this fallacy, two investigators testing a substantive hypothesis of depression using two different measures of this construct may arrive at different results. In turn, they may mistakenly attribute this inconsistency to shortcomings or boundary conditions in the substantive hypothesis rather than to differences in the content of the depression measure itself.

Indeed, an analysis of seven widely used measures of clinical depression revealed that they span 52 signs/symptoms, with 40% of these signs/symptoms being unique to only one measure (Fried, 2017). Broadly comparable and in some cases larger levels of content heterogeneity extend to measures of many other psychological disorders, including bipolar disorder, schizophrenia, obsessive-compulsive disorder, eating disorders, and autism spectrum disorder (Newson, Hunter, & Thiagarajan, 2020). Similarly, in the psychopathic personality (psychopathy) literature, many measures of this construct differ markedly in their content coverage, with some featuring extensive representation of boldness but others featuring little or no representation of it (Lilienfeld, Watts, Francis Smith, Berg, & Latzman, 2015).

Pronounced differences in coverage of mental disorders are potentially problematic because the principle of content validity requires that a measure samples adequately from the universe of content relevant to the construct (Haynes, Richard, & Kubany, 1995). If it does not, researchers and practitioners may draw inferences regarding an unrepresentative reflection of the clinical phenomenon of interest. Furthermore, substantial differences in content coverage across measures of the same disorder are likely to yield heterogeneous phenotypes across studies, thereby impeding the search for shared etiological influences (Newson et al., 2020). Put somewhat differently, it is difficult to pin down an underlying cause of a shifting target.

The problem at hand goes well beyond the jingle fallacy, however. Many psychological measures contain substantial amounts of construct-irrelevant variance, that is, variance stemming from sources other than the target construct (Messick, 1995). For example, the personal distress subscale of the most widely used self-report measure of empathy, namely the Interpersonal Reactivity Index, tends to be negligibly correlated with other empathy measures and does not load onto a higher-order empathy dimension (Murphy et al., 2020). These findings notwithstanding, many investigators continue to use the personal distress scale as a partial proxy for empathy, often combining it with the Interpersonal Reactivity Index Empathic Concern Scale to form an affective empathy composite. As a consequence, inconsistencies in the empathy literature may stem in part from some investigators' reliance on measures that detect dispositions, such as emotional distress, that are at best marginally relevant to empathy.

In some cases, the problem posed by construct-irrelevant variance is subtler yet arguably more ubiquitous. Many and arguably most life event scales, especially those featuring extensive coverage of potentially controllable stressful life events (e.g., loss of a job, conflict with spouse), are substantially contaminated by such personality traits as negative emotionality (NE). NE is a broad disposition that reflects the extent to which individuals experience unpleasant emotions of many kinds, including anxiety, hostility, guilt, and alienation (Watson & Clark, 1984). This contamination probably arises from two sources. First, the trait anxiety component of NE relates to and probably influences how individuals interpret and react to ambiguous stimuli because trait anxiety is linked to sensitivity to cues of threat (Barsky, Thoresen, Warren, & Kaplan, 2004; Brett, Brief, Burke, George, & Webster, 1990). Second, the interpersonal features (e.g., hostility) of NE relate to and probably influence individuals' risk of exposure to negative life events (Manuck & McCaffery, 2010) because chronically irritable and angry individuals often evoke unwelcome reactions from others. Consistent with these findings, scales containing

[1] Our list of sacred cows is by no means exhaustive. For example, we do not address widely shared but erroneous assumptions regarding the interpretation of specific statistics, such as the presumption that Cronbach's alpha is an adequate index of test homogeneity (Flake, Pek, & Hehman, 2017; Sijtsma, 2009).

[2] We would be remiss not to also mention the jangle fallacy, the error of assuming that two or more psychological phenomena, such as two measures, are different merely because they bear different names (Kelley, 1927). For example, hundreds of studies in the 1960s and 1970s interpreted scores on a measure of repression-sensitization as specific to the construct of repression (e.g., Byrne, Golightly, & Sheffield, 1965). Nevertheless, later studies showed that this measure is merely one of many alternative indicators of negative emotionality and is more or less interchangeable with numerous measures of trait anxiety and emotional maladjustment (Watson & Clark, 1984).

measures of potentially controllable life events are moderately heritable (Bemmels, Burt, Legrand, Iacono, & McGue, 2008), with much of this genetic variance being mediated by personality traits, including NE (Saudino, Pedersen, Lichtenstein, McClearn, & Plomin, 1997). Hence, because of their overlap with NE, many measures of life events almost certainly detect considerably more than life events per se. Furthermore, some discrepancies in the literature regarding the direct or moderating relations between life events and psychopathology (e.g., Monroe & Reid, 2009) may be due to differences across life events measures in their saturation with NE.

In turn, this body of research raises questions regarding the increasingly popular practice of interpreting scores on measures of adverse childhood experiences (ACEs), which include items assessing emotional abuse, neglect, homelessness, parental mental illness, and the like, as pure indicators of traumatic event exposure among children (e.g., Jones, Nurius, Song, & Fleming, 2018). Many authors further interpret the well-replicated correlations between ACEs and mental and physical health outcomes as directly causal (e.g., Hughes et al., 2017). These inferences are unwarranted. Among other things, commonly used ACE checklists may in part reflect a host of extraneous variables distinct from childhood trauma per se, including familial poverty and both environmental and genetic risk for psychopathology (Coyne, 2017; see also Anda, Porter, & Brown, 2020; Kelly-Irving & Delpierre, 2019).

In sum, researchers should never rely exclusively on the names of psychological measures to infer their content. They should instead carefully review the nomological network of external correlates (Cronbach et al., 1955) surrounding these measures to ascertain for themselves whether these measures are performing as advertised.

Sacred Cow #2: Reliability Is Not a Major Concern for Laboratory Measures

As Epstein (1979) noted more than 4 decades ago, psychology at large has tended to valorize laboratory measures, presuming that they are inherently more scientific than more easily collected measures, such as questionnaires or behavioural observations. As a consequence, many researchers have accorded insufficient consideration to basic psychometric principles, especially reliability, when using and interpreting laboratory measures (see also Block, 1977). In the words of one author team, "researchers [using laboratory measures of information processing] have been granted psychometric free rein that would probably never be extended to researchers using other measures, such as questionnaires" (Vasey, Dalgleish, & Silverman, 2003, p. 84).

Indeed, there is every reason to believe that laboratory indicators are subject to the same psychometric limitations as other psychological indicators. In fact, laboratory indicators may often be less reliable than most other measures because of their high levels of situational uniqueness (Epstein, 1979; Lilienfeld & Treadway, 2016). For example, scores on these measures may be influenced by a host of transient situational variables of little or no relevance to the constructs they are intended to detect, such as fatigue, inattention, demand characteristics, order effects, the precise phrasing of instructions, the perceived attitude of the research assistant administering the measure, and so on.

Because classical test theory posits that validity is limited by the square root of reliability (Meehl, 1986), the low reliability of many laboratory measures is likely to impose marked constraints on their construct validity.[3] Furthermore, because of low reliability, the association between measured variables becomes a biased (inaccurate) estimate of the association between their respective constructs. In addition, because regression to the mean is exacerbated when reliability is low, low reliability leads to unstable estimates of the magnitudes of statistical effects and boosts the risk of replication failures (Streiner, 2016).

A cautionary tale in this regard derives from research on the dot-probe task, which has been widely used to measure attentional biases in anxiety and anxiety disorders (Bar-Haim et al., 2007). Although the dot-probe task was administered in hundreds of studies to test the hypothesis that individuals with marked levels of anxiety are hypersensitive to threat cues (Kappenman, Farrens, Luck, & Proudfit, 2014), few investigators had examined its reliability. When they belatedly did so, they discovered that this task displayed only marginal internal consistency and test-retest reliability (Chapman, Devue, & Grimshaw, 2019; Kruijt, Parsons, & Fox, 2019; Schmukle, 2005; Staugaard, 2009). The poor reliability of this task may partially account for the numerous replication failures in this literature (e.g., Asmundson & Stein, 1994; Everaert, Mogoaşe, David, & Koster, 2015; Parsons, Kruijt, & Fox, 2019; Wenzel & Holt, 1999).

One might suspect that the large-scale neglect of reliability considerations extends to a wide variety of laboratory measures in addition to the dot-probe task. As a preliminary test of this conjecture, the authors of this article scoured the *Method* sections of all empirical articles published in the 2019 edition of the *Journal of Abnormal Psychology* (arguably one of the two flagship journals in the field of psychopathology). To provide a rough gauge of comparison of laboratory with nonlaboratory measures, we limited ourselves to articles that included both (a) one or more laboratory measures and (b) one or more nonlaboratory (e.g., self-report, psychiatric interview, behavioural observation) measures. We erred on the side of providing an overly liberal (generous) estimate of the extent to which the authors reported on the reliability of their measures, giving them credit for doing so if they reported at least one form of reliability (internal consistency, test-retest, interrater) for any of their laboratory or nonlaboratory measures. Of the 34 articles coded (74%), 25 reported no data (either in the main text or Supplemental Materials) on any form of reliability of their laboratory measures. The corresponding figure for nonlaboratory measures was 17 (50%); this difference yielded a χ^2 value of 3.99 ($p < .05$). Pending replication and extension to other journals, of course, these findings offer strongly suggestive evidence that many or most authors in the recent psychopathology literature do not routinely report basic reliability data on their measures and provisional evidence that this reporting problem may be even more marked for laboratory than for nonlaboratory measures.

[3] To be sure, this statement is something of an oversimplification, given that there are multiple forms of reliability (internal consistency, test-retest, interrater) and multiple forms of validity nested within the broader concept of construct validity (e.g., content, criterion). Furthermore, contemporary testing standards emphasize that reliability and validity are not attributes of a measurement instrument per se but rather indices of how well a test performs under specific conditions and in specific settings.

In addition, a longstanding anomaly in the personality and psychopathology literatures has been the low or at best modest correlations between self-report and laboratory indicators of numerous psychological constructs, including impulsivity (Cyders & Coskunpinar, 2012; Sharma, Markon, & Clark, 2014), cognitive empathy (Murphy & Lilienfeld, 2019), physical aggression (Muntaner et al., 1990), creativity (Park, Chun, & Lee, 2016), and emotional intelligence (Brackett, Rivers, Shiffman, Lerner, & Salovey, 2006). Many authors have been tempted to attribute these low associations to the shortcomings of self-report measures, such as their undue reliance on insight (Nisbett & Wilson, 1977) or their susceptibility to social desirability response artifacts (Paulhus, 2017).

Although there may be some truth to this interpretation, it is also likely that at least some of the fault lies with laboratory measures as well. Because they were developed largely to detect the effects of short-term experimental manipulations (e.g., aggression-inducing stimuli), many or most of these measures were designed to maximize within-person variance and minimize between-person variance, a phenomenon termed the reliability paradox (Hedge, Powell, & Sumner, 2018). As a consequence, laboratory measures are often poorly suited to detect stable individual differences in personality traits. In addition, whereas most self-report measures are typical performance measures, which assess how people generally behave in everyday life, most laboratory measures are maximal performance measures, which assess how people behave when they are pushed to perform their best (see Cronbach, 1960 for a discussion of the typical-maximal performance distinction). For example, whereas questionnaire measures of emotional intelligence assess individuals' longstanding levels of self-esteem, empathy, emotional regulation, and the like, laboratory measures of emotional intelligence assess the extent to which individuals are capable of emotion recognition, cognitive empathy, and other skills when pushed to their limits. Although maximal performance measures can be enormously useful for detecting individuals' aptitudes, they are often mismatched for detecting individuals' enduring patterns of behavior under routine conditions. For example, individuals with marked psychopathic personality traits may be able to detect subtle facial expressions of emotion (e.g., sadness) when asked to do so in a laboratory setting but may fail to detect such expressions in everyday life, given their insufficient motivation to do so. This discrepancy is not a flaw of the laboratory test per se. Still, it may be tempting to misinterpret scores on this and many otherwise informative laboratory measures as reflecting typical performance.

Parsons et al. (2019) recently furnished readers with a set of user-friendly guidelines, along with a tutorial and R code, for calculating the internal consistencies of laboratory-based indices. We concur with Parsons et al. that journal editors and reviewers should routinely ask researchers to provide reliability data on laboratory-based measures and insist that researchers provide compelling justifications for any failures to do so.

Sacred Cow #3: Using Measures That Are Difficult to Collect Obviates the Need for Large Sample Sizes

Small sample sizes in psychological research boost the risk of type I errors, increase the imprecision of effect size estimates, and result in large standard errors (and accompanying large confidence intervals). Having served as editor and associate editor of two major psychopathology journals (*Clinical Psychological Science* and *Journal of Abnormal Psychology*, respectively), the first author of this article frequently encountered a curious justification from authors who submitted articles based on extremely small samples, often *N*s of 15 or less per cell. This justification appeared to be invoked most frequently by authors of cognitive and affective neuroscience articles. In essence, their rationale can be paraphrased as follows: "Our data were extremely time and labor intensive to collect, so our sample size was necessarily limited."

As an action editor on such articles, I reacted to this defense with decidedly mixed emotions. On the one hand, I was sympathetic—and remain sympathetic—to the formidable pragmatic challenges involved in collecting human functional brain imaging and other neuroscience data, which require extensive equipment, data collection, and data processing. On the other hand, is it unclear how, if at all, such practical difficulties should be weighed when evaluating an article's methodological rigor. As psychometrician Frederick Lord (1953) reminded us, the data do not know where they came from. A sample size of 10 is still a sample size of 10, regardless of whether one obtained it using functional magnetic resonance imaging or a self-report questionnaire. The same principles of statistical power hold in both cases.

These caveats notwithstanding, the statistical power of investigations in human neuroscience remains low on average and is lower than in most other domains of psychology (Button et al., 2013; Carter, Tilling, & Munafò, 2017; Grabitz et al., 2018; Turner, Paul, Miller, & Barbey, 2018). In one analysis of cognitive neuroscience research, the mean statistical power to detect small effects was only .14, meaning that many null results in this literature may be false negatives (Szucs & Ioannidis, 2017).

Nevertheless, contrary to what many authors appear to assume, low statistical power does not merely boost the risk for type II errors. Instead, positive findings emanating from underpowered studies are more likely than findings emanating from adequately powered studies to be type I errors (false positives), a statistical phenomenon termed the winner's curse (Algermissen & Mehler, 2018; Button et al., 2013). Still, many authors in the neuroscience literature justify their small sample sizes on the grounds that earlier similar studies yielding positive results relied on comparably small samples (Goodhill, 2017). This reasoning is flawed, however, because it overlooks the possibility that the previous findings were false-positives.

The broader problem we have highlighted is not unique to the neuroscience literature, however. By virtue of their focus on statistically rare populations, such as individuals with dissociative disorder, survivors of suicide attempts, or direct witnesses of a traumatic event (e.g., the September 11, 2001, terrorist attacks), many articles in the psychopathology literature are similarly underpowered to detect all but large effects (Tackett et al., 2017). In addition, owing at least in part to the expense and time intensiveness of data collection, many or most studies in the infancy literature are characterized by small samples. For example, many studies in the infant looking-time literature rely on samples of eight to 12 participants per cell, rendering findings difficult to replicate (Oakes, 2017; see also Bergmann et al., 2018 for a discussion of low statistical power in infant and child language acquisition research).

Fortunately, the neuroscience and infancy literatures are beginning to grapple with the challenges posed by low-powered studies by developing collaborative protocols shared across multiple laboratories (e.g., the ManyBabies Project; Frank et al., 2017; Poline et al., 2012). We strongly encourage similar collaborative efforts in other domains in which they are feasible, such as many studies of relatively rare clinical phenomena (Tackett et al., 2017). In the meantime, it is incumbent on investigators who rely on small samples to qualify the strength of their conclusions accordingly, especially when communicating the implications of their findings to the media and their academic colleagues.

Sacred Cow #4: Convergent Validity Data Afford Sufficient Evidence for Construct Validity

In their classic article, Campbell and Fiske (1959) introduced the now-familiar distinction between convergent and discriminant validity and argued that the latter principle is essential for appraising the construct validity of psychological instruments: "For the justification of novel trait measures, for the validation of test interpretation, or for the establishment of construct validity, discriminant validation as well as convergent validation is required. *Tests can be invalidated by too high correlations with other tests from which they were intended to differ*" (p. 81; emphasis added).

Insufficient consideration of discriminant validation may be one insufficiently appreciated source of replication failures. If investigators assume that their measure detects target construct X when it in fact primarily detects construct Y, they may obtain negative results. These negative results may, however, be due to the task impurity (Miyake & Friedman, 2012) of the measure rather than to an inadequacy in their substantive hypothesis.

Nevertheless, even a casual inspection of the *Method* sections of articles in the personality psychology and psychopathology literatures reveals that many or most of them appear to accord negligible or even no attention to the discriminant validity of their measures (Lilienfeld, 2004). In the psychopathology domain, this neglect is especially problematic, given the extensive covariation among most measures of psychopathology (Borsboom, Cramer, Schmittmann, Epskamp, & Waldorp, 2011; Krueger & Markon, 2006; Lilienfeld, Waldman, & Israel, 1994). As a consequence of this covariation, an investigator can mistakenly conclude that a measure detects disorder X when it detects disorder Y to an equal or greater extent.

For example, the first author of this article recently reviewed an article in which the authors had derived a new, ad hoc index of psychopathy from preexisting items in their data set. As validational support for this measure, they reported that it correlated approximately $r = .30$ in their sample with a well-established psychopathy measure. Setting aside the unimpressive magnitude of this convergent validity correlation, this association is difficult or impossible to interpret without any accompanying discriminant validity evidence, which the authors did not report. It is entirely possible, for example, that their new measure correlated even more highly with measures of constructs that are overlapping with but theoretically separable from psychopathy, such as substance use disorder (Smith & Newman, 1990), antisocial personality disorder (Hare, Hart, & Harpur, 1991), or narcissistic personality disorder (Miller et al., 2010). If so, their central conclusions, which implied that their findings were largely or entirely specific to psychopathy, would have been erroneous.

To test the hypothesis that many or most authors deemphasize discriminant validity evidence relative to convergent validity evidence, we examined the *Method* sections of all empirical articles published in 2019 in *Psychological Assessment*, which is often regarded as the flagship measurement journal in the fields of psychopathology and cognitive assessment. We coded whether the author had reported evidence for the convergent (or concurrent) validity, discriminant (or divergent) validity, or both of the measures administered in their study. We again adopted a liberal (generous) criterion, counting the evidence as positive if the authors mentioned data regarding the convergent or discriminant validity of any the measures in their study. For studies of categorical measures of psychiatric diagnoses, we counted data on sensitivity as evidence for convergent validity and data on specificity as evidence for discriminant validity. Nevertheless, if authors referred only to generic evidence of validity (e.g., "has well-established construct validity"; "is a well-validated instrument") or to adequate psychometric properties (e.g., "good psychometric properties") of their measures with no reference to specific convergent or discriminant validity data, we did not regard their *Method* sections as offering evidence for either convergent or discriminant validity.

Of the *Method* sections of the 69 articles coded, 45 (65%) provided no specific evidence for the convergent or discriminant validity of any of their measures. Of the 24 reporting specific validity information for one or more of their measures, 10 (42%) presented both convergent and discriminant validity data and 14 (58%) presented convergent validity data only. No articles presented discriminant validity data only. These findings, although limited to 1 year of one journal, are consistent with our hypothesis that when characterizing the psychometric properties of measures, discriminant validity tends to receive short shrift at large as well as short shrift relative to convergent validity.

Although we have focused on the psychopathology and personality literatures here, the same neglect of discriminant validity appears to be endemic to many if not most other psychological domains. To take merely one example, many articles examining the correlates of specific mental capacities, such as spatial, arithmetic, or verbal ability, neglect to account for their substantial covariation with measures of general mental ability (Schmidt, 2017). As a consequence, they often imply misleadingly that their results are attributable to these specific abilities rather than to global intelligence. In sum, to address this widespread problem, we recommend that editors and journal reviewers require authors to provide at least as much information regarding their measures' discriminant validity as for their convergent validity.

Concluding Thoughts

It is tempting to take the psychometric properties of our measures for granted. But doing so can contribute to serious errors when interpreting findings and to inconsistent results across studies. As discussed earlier, researchers can take several steps to counteract questionable measurement practices.

First, when selecting a measure, researchers should carefully examine its nomological network of external correlates to infer its content rather than rely on the measure's name alone. Second, researchers who administer laboratory measures should not assume that their reliabilities are adequate but should instead use recently published formulas (Parsons et al., 2019) for calculating

and reporting such reliability. Researchers who are unable to provide these data should present a compelling rationale for their exclusion. Third, researchers who work in fields in which small sample sizes are common (e.g., human neuroscience, infant research, psychopathology) should explore collaborative options that allow them to pool data to increase statistical power. Conversely, when reporting findings from studies with small sample sizes, researchers should qualify the strength of their conclusions accordingly, and journals should encourage this practice. Fourth, journal editors and reviewers should require that researchers report discriminant validity data—and not merely convergent validity data—for all measures. By attending to the recommendations we have outlined here, researchers can hopefully avoid falling prey to these and other sacred cows, avoid questionable measurement practices (Flake & Fried, 2019; Hughes et al., 2017), and ensure that their results and conclusions are grounded more firmly in the basic science of psychological measurement.

Résumé

Bien qu'il soit certain que de nombreux facteurs contribuent à la crise de la reproductibilité en psychologie, l'un d'entre eux, largement méconnu, est la négligence des principes de base de la mesure. Nous examinons quatre principes « intouchables » de la mesure en psychologie – des hypothèses largement diffusées et rarement remises en question – qui, en rendant les pratiques de mesure discutables, peuvent alimenter la crise de la reproductibilité. Ces quatre intouchables sont les suivants : (A) nous pouvons nous fier en toute confiance au nom d'une mesure pour en déduire le contenu; (b) la fiabilité n'est pas une préoccupation majeure pour les mesures en laboratoire; (c) le recours à des mesures qui sont difficiles à recueillir écarte le besoin d'échantillons de taille plus importante; (d) des données convergentes sur la validité constituent des éléments de preuve suffisants de la validité conceptuelle. Pour les éléments a et d, nous fournissons des données provisoires issues de revues de psychologie récentes qui soutiennent notre affirmation selon laquelle de telles croyances prévalent parmi les auteurs. Afin d'améliorer la reproductibilité de la science de la psychologie, les chercheurs devront être vigilants face aux suppositions erronées concernant les propriétés psychométriques de ces mesures et aux répercussions de ces propriétés psychométriques pour leurs études.

Mots-clés : mesure psychologique, crise de la reproductibilité, pratiques de mesure discutables, mesures en laboratoire, validité discriminante.

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist, 63,* 32–50. http://dx.doi.org/10.1037/0003-066X.63.1.32

Algermissen, J., & Mehler, D. M. A. (2018). May the power be with you: Are there highly powered studies in neuroscience, and how can we get more of them? *Journal of Neurophysiology, 119,* 2114–2117. http://dx.doi.org/10.1152/jn.00765.2017

Anda, R. F., Porter, L. E., & Brown, D. W. (2020). Inside the adverse childhood experience score: Strengths, limitations, and misapplications. *American Journal of Preventive Medicine*. Advance online publication. http://dx.doi.org/10.1016/j.amepre.2020.01.009

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., . . . Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119. http://dx.doi.org/10.1002/per.1919

Asmundson, G. J., & Stein, M. B. (1994). Selective processing of social threat in patients with generalized social phobia: Evaluation using a dot-probe paradigm. *Journal of Anxiety Disorders, 8,* 107–117. http://dx.doi.org/10.1016/0887-6185(94)90009-4

Bar-Haim, Y., Lamy, D., Pergamin, L., Bakermans-Kranenburg, M. J., & van IJzendoorn, M. H. (2007). Threat-related attentional bias in anxious and nonanxious individuals: A meta-analytic study. *Psychological Bulletin, 133,* 1–24. http://dx.doi.org/10.1037/0033-2909.133.1.1

Barsky, A., Thoresen, C. J., Warren, C. R., & Kaplan, S. A. (2004). Modeling negative affectivity and job stress: A contingency-based approach. *Journal of Organizational Behavior, 25,* 915–936. http://dx.doi.org/10.1002/job.285

Bemmels, H. R., Burt, S. A., Legrand, L. N., Iacono, W. G., & McGue, M. (2008). The heritability of life events: An adolescent twin and adoption study. *Twin Research and Human Genetics, 11,* 257–265. http://dx.doi.org/10.1375/twin.11.3.257

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89,* 1996–2009. http://dx.doi.org/10.1111/cdev.13079

Block, J. (1977). Advancing the science of personality: Paradigmatic shift or improving the quality of research? In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 37–63). Hillsdale, NJ: Erlbaum.

Borsboom, D., Cramer, A. O., Schmittmann, V. D., Epskamp, S., & Waldorp, L. J. (2011). The small world of psychopathology. *PLoS ONE, 6,* e27407. http://dx.doi.org/10.1371/journal.pone.0027407

Brackett, M. A., Rivers, S. E., Shiffman, S., Lerner, N., & Salovey, P. (2006). Relating emotional abilities to social functioning: A comparison of self-report and performance measures of emotional intelligence. *Journal of Personality and Social Psychology, 91,* 780–795. http://dx.doi.org/10.1037/0022-3514.91.4.780

Brett, J. F., Brief, A. P., Burke, M. J., George, J. M., & Webster, J. (1990). Negative affectivity and the reporting of stressful life events. *Health Psychology, 9,* 57–68. http://dx.doi.org/10.1037/0278-6133.9.1.57

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365–376. http://dx.doi.org/10.1038/nrn3475

Byrne, D., Golightly, C., & Sheffield, J. (1965). The repression-sensitization scale as a measure of adjustment: Relationship with the CPI. *Journal of Consulting Psychology, 29,* 586–589. http://dx.doi.org/10.1037/h0022751

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitree-multimethod matrix. *Psychological Bulletin, 56,* 81–105. http://dx.doi.org/10.1037/h0046016

Carter, A. R., Tilling, K., & Munafò, M. R. (2017). A systematic review of sample size and power in leading neuroscience journals. Advance online publication. http://dx.doi.org/10.1101/217596

Chapman, A., Devue, C., & Grimshaw, G. M. (2019). Fleeting reliability in the dot-probe task. *Psychological Research, 83,* 308–320. http://dx.doi.org/10.1007/s00426-017-0947-6

Coyne, J. C. (2017). *Stop using the Adverse Childhood Experiences Checklist to make claims about trauma causing physical health and mental problems*. Retrieved from http://www.coyneoftherealm.com/2017/11/15/stop-using-the-adverse-childhood-experiences-checklist-to-make-claims-about-trauma-causing-physical-and-mental-health-problems/

Cronbach, L. J. (1954). Report on a psychometric mission to Clinicia. *Psychometrika, 19,* 263–270. http://dx.doi.org/10.1007/BF02289226

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). Oxford, England: Harper.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302. http://dx.doi.org/10.1037/h0040957

Cyders, M. A., & Coskunpinar, A. (2012). The relationship between self-report and lab task conceptualizations of impulsivity. *Journal of Research in Personality, 46,* 121–124. http://dx.doi.org/10.1016/j.jrp.2011.11.005

Epstein, S. (1979). The stability of behavior: I. On predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37,* 1097–1126. http://dx.doi.org/10.1037/0022-3514.37.7.1097

Epstein, S., & O'Brien, E. J. (1985). The person-situation debate in historical and current perspective. *Psychological Bulletin, 98,* 513–537. http://dx.doi.org/10.1037/0033-2909.98.3.513

Everaert, J., Mogoaşe, C., David, D., & Koster, E. H. (2015). Attention bias modification via single-session dot-probe training: Failures to replicate. *Journal of Behavior Therapy and Experimental Psychiatry, 49,* 5–12. http://dx.doi.org/10.1016/j.jbtep.2014.10.011

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological & Personality Science, 7,* 45–52. http://dx.doi.org/10.1177/1948550615612150

Flake, J. K., & Fried, E. I. (2019). *Measurement schmeasurement: Questionable measurement practices and how to avoid them*. Retrieved from https://psyarxiv.com/hs7wm/

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological & Personality Science, 8,* 370–378. http://dx.doi.org/10.1177/1948550617693063

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., . . . Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy, 22,* 421–435. http://dx.doi.org/10.1111/infa.12182

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders, 208,* 191–197. http://dx.doi.org/10.1016/j.jad.2016.10.019

Garber, J., & Strassberg, Z. (1991). Construct validity: History of application to developmental psychopathology. In W. Grove & D. Cicchetti (Eds.), *Personality and psychopathology* (pp. 219–258). Minneapolis, MN: University of Minnesota Press.

Goodhill, G. J. (2017). *Is neuroscience facing up to statistical power?* Retrieved from https://arxiv.org/abs/1701.01219

Grabitz, C. R., Button, K. S., Munafò, M. R., Newbury, D. F., Pernet, C. R., Thompson, P. A., & Bishop, D. V. M. (2018). Logical and methodological issues affecting genetic studies of humans reported in top neuroscience journals. *Journal of Cognitive Neuroscience, 30,* 25–41. http://dx.doi.org/10.1162/jocn_a_01192

Hare, R. D., Hart, S. D., & Harpur, T. J. (1991). Psychopathy and the *DSM–IV* criteria for antisocial personality disorder. *Journal of Abnormal Psychology, 100,* 391–398. http://dx.doi.org/10.1037/0021-843X.100.3.391

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7,* 238–247. http://dx.doi.org/10.1037/1040-3590.7.3.238

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods, 50,* 1166–1186. http://dx.doi.org/10.3758/s13428-017-0935-1

Hegel, G. F. W. (1823). *Lectures on the philosophy of world history, volume I: Manuscripts of the introduction and the lectures of 1822–1823* (R. F. Brown, P.C. Hodgson, Eds). Oxford, UK: Oxford University Press.

Hughes, B. M. (2018). *Psychology in crisis*. London: U.K. Macmillan International Higher Education.

Hughes, K., Bellis, M. A., Hardcastle, K. A., Sethi, D., Butchart, A., Mikton, C., . . . Dunne, M. P. (2017). The effect of multiple adverse childhood experiences on health: A systematic review and meta-analysis. *Lancet Public Health, 2,* e356–e366. http://dx.doi.org/10.1016/S2468-2667(17)30118-4

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23,* 524–532. http://dx.doi.org/10.1177/0956797611430953

Jones, T. M., Nurius, P., Song, C., & Fleming, C. M. (2018). Modeling life course pathways from adverse childhood experiences to adult mental health. *Child Abuse & Neglect: The International Journal, 80,* 32–40. http://dx.doi.org/10.1016/j.chiabu.2018.03.005

Kappenman, E. S., Farrens, J. L., Luck, S. J., & Proudfit, G. H. (2014). 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, 5,* 1368. http://dx.doi.org/10.3389/fpsyg.2014.01368

Kelley, T. L. (1927). *Interpretation of educational measurement*. Yonkers, NY: World Book.

Kelly-Irving, M., & Delpierre, C. (2019). A critique of the adverse childhood experiences framework in epidemiology and public health: Uses and misuses. *Social Policy and Society, 18,* 445–456. http://dx.doi.org/10.1017/S1474746419000101

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology, 2,* 111–133. http://dx.doi.org/10.1146/annurev.clinpsy.2.022305.095213

Kruijt, A. W., Parsons, S., & Fox, E. (2019). A meta-analysis of bias at baseline in RCTs of attention bias modification: No evidence for dot-probe bias towards threat in clinical anxiety and PTSD. *Journal of Abnormal Psychology, 128,* 563–573. http://dx.doi.org/10.1037/abn0000406

Lilienfeld, S. O. (2004). Taking theoretical risks in a world of directional predictions. *Applied & Preventive Psychology, 11,* 47–51. http://dx.doi.org/10.1016/j.appsy.2004.02.008

Lilienfeld, S. O., Lynn, S. J., & Namy, L. (2018). *Psychology: From inquiry to understanding*. Hoboken, NJ: Pearson.

Lilienfeld, S. O., & Treadway, M. T. (2016). Clashing diagnostic approaches: DSM-ICD versus RDoC. *Annual Review of Clinical Psychology, 12,* 435–463. http://dx.doi.org/10.1146/annurev-clinpsy-021815-093122

Lilienfeld, S. O., & Waldman, I. D. (Eds.). (2017). *Psychological science under scrutiny: Recent challenges and proposed solutions*. New York: Wiley. http://dx.doi.org/10.1002/9781119095910

Lilienfeld, S. O., Waldman, I. D., & Israel, A. C. (1994). A critical examination of the use of the term and concept of comorbidity in psychopathology research. *Clinical Psychology: Science and Practice, 1,* 71–83. http://dx.doi.org/10.1111/j.1468-2850.1994.tb00007.x

Lilienfeld, S. O., Watts, A. L., Francis Smith, S., Berg, J. M., & Latzman, R. D. (2015). Psychopathy deconstructed and reconstructed: Identifying and assembling the personality building blocks of Cleckley's chimera. *Journal of Personality, 83,* 593–610. http://dx.doi.org/10.1111/jopy.12118

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist, 8,* 750–751. http://dx.doi.org/10.1037/h0063675

Manuck, S. B., & McCaffery, J. M. (2010). Genetics of stress: Gene–stress correlation and interaction. In A. Steptoe (Ed.), *Handbook of behavioral medicine* (pp. 455–478). New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-09488-5_31

Meehl, P. E. (1986). Diagnostic taxa as open concepts: Metatheoretical and statistical questions about reliability and construct validity in the grand

strategy of nosological revision. In T. Millon & G. L. Klerman (Eds.), *Contemporary directions in psychopathology* (pp. 215–231). New York, NY: Guilford Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Miller, J. D., Dir, A., Gentile, B., Wilson, L., Pryor, L. R., & Campbell, W. K. (2010). Searching for a vulnerable dark triad: Comparing Factor 2 psychopathy, vulnerable narcissism, and borderline personality disorder. *Journal of Personality, 78,* 1529–1564. http://dx.doi.org/10.1111/j.1467-6494.2010.00660.x

Miyake, A., & Friedman, N. P. (2012). The nature and organization of individual differences in executive functions: Four general conclusions. *Current Directions in Psychological Science, 21,* 8–14. http://dx.doi.org/10.1177/0963721411429458

Monroe, S. M., & Reid, M. W. (2009). Life stress and major depression. *Current Directions in Psychological Science, 18,* 68–72. http://dx.doi.org/10.1111/j.1467-8721.2009.01611.x

Muntaner, C., Walter, D., Nagoshi, C., Fishbein, D., Haertzen, C. A., & Jaffe, J. H. (1990). Self-report vs. laboratory measures of aggression as predictors of substance abuse. *Drug and Alcohol Dependence, 25,* 1–11. http://dx.doi.org/10.1016/0376-8716(90)90133-Y

Murphy, B. A., Costello, T. H., Watts, A. L., Cheong, Y. F., Berg, J. M., & Lilienfeld, S. O. (2020). Strengths and weaknesses of two empathy measures: A comparison of the measurement precision, construct validity, and incremental validity of two multidimensional indices. *Assessment, 27,* 246–260.

Murphy, B. A., & Lilienfeld, S. O. (2019). Are self-report cognitive empathy ratings valid proxies for cognitive empathy ability? Negligible meta-analytic relations with behavioral task performance. *Psychological Assessment, 31,* 1062–1072. http://dx.doi.org/10.1037/pas0000732

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69,* 511–534. http://dx.doi.org/10.1146/annurev-psych-122216-011836

Newson, J. J., Hunter, D., & Thiagarajan, T. C. (2020). The heterogeneity of mental health assessment. *Frontiers in Psychiatry, 11,* 76. http://dx.doi.org/10.3389/fpsyt.2020.00076

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259. http://dx.doi.org/10.1037/0033-295X.84.3.231

Oakes, L. M. (2017). Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy, 22,* 436–469. http://dx.doi.org/10.1111/infa.12186

Park, N. K., Chun, M. Y., & Lee, J. (2016). Revisiting individual creativity assessment: Triangulation in subjective and objective assessment methods. *Creativity Research Journal, 28,* 1–10. http://dx.doi.org/10.1080/10400419.2016.1125259

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science, 2,* 378–395. http://dx.doi.org/10.1177/2515245919879695

Paulhus, D. L. (2017). Socially desirable responding on self-reports. In V. Zeigler-Hill & T. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1–5). New York: Sage. http://dx.doi.org/10.1007/978-3-319-28099-8_1349-1

Poline, J. B., Breeze, J. L., Ghosh, S., Gorgolewski, K., Halchenko, Y. O., Hanke, M., . . . Kennedy, D. N. (2012). Data sharing in neuroimaging research. *Frontiers in Neuroinformatics, 6,* 9. http://dx.doi.org/10.3389/fninf.2012.00009

Saudino, K. J., Pedersen, N. L., Lichtenstein, P., McClearn, G. E., & Plomin, R. (1997). Can personality explain genetic influences on life events? *Journal of Personality and Social Psychology, 72,* 196–206. http://dx.doi.org/10.1037/0022-3514.72.1.196

Schmidt, F. L. (2017). Beyond questionable research methods: The role of omitted relevant research in the credibility of research. *Archives of Scientific Psychology, 5,* 32–41. http://dx.doi.org/10.1037/arc0000033

Schmukle, S. C. (2005). Unreliability of the dot probe task. *European Journal of Personality, 19,* 595–605. http://dx.doi.org/10.1002/per.554

Schwartz, S. J., Lilienfeld, S. O., Meca, A., & Sauvigné, K. C. (2016). The role of neuroscience within psychology: A call for inclusiveness over exclusiveness. *American Psychologist, 71,* 52–70. http://dx.doi.org/10.1037/a0039678

Sharma, L., Markon, K. E., & Clark, L. A. (2014). Toward a theory of distinct types of "impulsive" behaviors: A meta-analysis of self-report and behavioral measures. *Psychological Bulletin, 140,* 374–408. http://dx.doi.org/10.1037/a0034418

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120.

Smith, S. S., & Newman, J. P. (1990). Alcohol and drug abuse-dependence disorders in psychopathic and nonpsychopathic criminal offenders. *Journal of Abnormal Psychology, 99,* 430–439. http://dx.doi.org/10.1037/0021-843X.99.4.430

Staugaard, S. R. (2009). Reliability of two versions of the dot-probe task using photographic faces. *Psychology Science Quarterly, 51,* 339–350.

Streiner, D. L. (2016). Statistics commentary series: Commentary #16—Regression toward the mean. *Journal of Clinical Psychopharmacology, 36,* 416–418. http://dx.doi.org/10.1097/JCP.0000000000000551

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15,* e2000797. http://dx.doi.org/10.1371/journal.pbio.2000797

Tackett, J. L., Lilienfeld, S. O., Patrick, C. J., Johnson, S. L., Krueger, R. F., Miller, J. D., . . . Shrout, P. E. (2017). It's time to broaden the replicability conversation: Thoughts for and from clinical psychological science. *Perspectives on Psychological Science, 12,* 742–756. http://dx.doi.org/10.1177/1745691617690042

Tackett, J. L., & Miller, J. D. (2019). Introduction to the special section on increasing replicability, transparency, and openness in clinical psychology. *Journal of Abnormal Psychology, 128,* 487–492. http://dx.doi.org/10.1037/abn0000455

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. New York, NY: Teachers College, Columbia University. http://dx.doi.org/10.1037/13283-000

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology, 1,* 62. http://dx.doi.org/10.1038/s42003-018-0073-z

Vasey, M. W., Dalgleish, T., & Silverman, W. K. (2003). Research on information-processing factors in child and adolescent psychopathology: A critical commentary. *Journal of Clinical Child and Adolescent Psychology, 32,* 81–93. http://dx.doi.org/10.1207/S15374424JCCP3201_08

Watson, D., & Clark, L. A. (1984). Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin, 96,* 465–490. http://dx.doi.org/10.1037/0033-2909.96.3.465

Wenzel, A., & Holt, C. S. (1999). Dot probe performance in two specific phobias. *British Journal of Clinical Psychology, 38,* 407–410. http://dx.doi.org/10.1348/014466599163006

Received March 16, 2020
Revision received May 11, 2020
Accepted May 12, 2020 ■