Hidden invalidity among 15 commonly used measures in social and personality psychology

Ian Hussey & Sean Hughes

It has recently been demonstrated that metrics of structural validity are severely underreported in social and personality psychology. We comprehensively assessed structural validity in a uniquely large and varied data set (N = 144,496 experimental sessions) to investigate the psychometric properties of some of the most widely used self-report measures (k = 15 questionnaires, 26 scales) in social and personality psychology. When the scales were assessed using the modal practice of considering only internal consistency, 89% of them appeared to possess good validity. Yet when validity was assessed comprehensively (via internal consistency, immediate and delayed test-retest reliability, factor structure, and measurement invariance for age and gender groups), only 4% demonstrated good validity. Furthermore, the less commonly a test was reported in the literature, the more likely the scales were to fail that test (e.g., scales failed measurement invariance much more often than internal consistency). This suggests that the pattern of underreporting in the field may represent widespread hidden invalidity of the measures used and may therefore pose a threat to many research findings. We highlight the degrees of freedom afforded to researchers in the assessment and reporting of structural validity and introduce the concept of validity hacking (v-hacking), similar to the better-known concept of p-hacking. We argue that the practice of v-hacking should be acknowledged and addressed.

Confidence in the replicability and reproducibility of research findings is a foundational pillar upon which theory, application, and progress reside. However, this pillar has recently been shaken. Large-scale efforts to document the replicability of research in psychological science have led many of its core findings to be called into question (Open Science Collaboration, 2015). These discipline-wide efforts have unleashed a tidal wave of new discussion and reflection on those modal practices that have contributed to the so-called replication crisis (LeBel & Peters, 2011; Simmons, Nelson, & Simonsohn, 2011). Numerous research and analytic practices, such as overreliance on and misuse of null-hypothesis significance testing, have been questioned, and the need for increased transparency, data sharing, preregistration, and direct replication has been highlighted and encouraged (Asendorpf et al., 2013; Munafò et al., 2017). Despite these laudable developments, Flake, Pek, and Hehman (2017) noted that the topic of measurement has received far less attention. This is surprising given that measurement plays a key role in replicability and ultimately calibrates the confidence researchers can have in their findings: If a measure is invalid, then theoretical conclusions derived from it are questionable.

Many, if not most, measures in social and personality psychology are designed to assess latent constructs that are unobservable in nature. For instance, a self-report scale may be created to assess belief in a just world or right-wing authoritarianism, or to quantify personality traits.[1] Designing valid measures of latent constructs requires that the measures themselves be subject to an ongoing process known as construct validation (Loevinger, 1957). Although psychological measures most commonly take the form of self-report scales, they can also take a variety of other forms, as in the case of reaction time–based implicit measures (for discussion of the

[1] Following Flake et al. (2017) we define a scale as a self-report measure that relies on items to represent a latent construct.

assessment of the validity of implicit measures specifically, see De Schryver, Hughes, De Houwer, & Rosseel, 2018; for further information on construct validation, see Borsboom, Mellenbergh, & van Heerden, 2004; Cronbach & Meehl, 1955). As Flake et al. (2017) explained, construct validation "is the process of integrating evidence to support the meaning of a number which is assumed to represent a psychological construct" (p. 2; see Cronbach & Meehl, 1955) and consists of three sequential phases (for a more detailed treatment, see Loevinger, 1957). The first, the substantive phase, involves identifying and defining a construct (via literature review and conceptualization of the construct), determining how it will be assessed (via item development and selection), and ensuring that the resulting scale content is both relevant and representative. In the second phase, the structural) phase, a theory about the construct's structure is developed. Quantitative analyses (e.g., item and factor analyses; assessments of consistency, stability, and measurement invariance) are used to determine the psychometric properties of the measure. The third phase, the external phase, involves examining if the measure appropriately represents the construct via checks for convergent and discriminant validity with other measures, predictive or criterion checks using known outcomes, or comparisons of known groups (for a more detailed overview, see American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Cronbach & Meehl, 1955; Loevinger, 1957).

Much of the theoretical work in social and personality psychology centers on the identification and definition of constructs (first phase), and empirical work tends to assess whether these constructs predict, discriminate between, or converge with other measures (third phase). Ascertaining the structure and psychometric properties of the measures used to assess these constructs (second phase) often receives far less attention. For instance, Flake et al. (2017) examined a representative sample of articles from a flagship journal in the field (*Journal of Personality and Social Psychology*) and found that many constructs studied in social and personality research lack appropriate validation. Specifically, they found an overreliance on Cronbach's α as the sole source of evidence for structural validity, and argued that rigorous methodologies for measurement are rarely reported. Indeed, Flake et al. found that the problem with validation was actually more severe than it initially appeared. Specifically, they found that research with well-known measures over relied on Cronbach's α as the sole test of structural validity. In addition, nearly half of the measures sampled were ad hoc and lacked evidence of validity testing at any of the three phases of validation.

Such a situation poses several threats: It (a) increases the potential for questionable theoretical conclusions and (b) decreases the chance that subsequent research will replicate results, given that (c) the three phases of validation are intertwined. Put simply, conclusions about the construct stemming from the external phase may not hold if issues exist at the substantive phase (e.g., the construct lacks a strong theoretical basis) or the structural phase (e.g., the measure lacks acceptable psychometric properties). Thus, substantive and structural validity need to be assessed if researchers wish to engage in theory testing (external validation) or replication. Fortunately, a set of best practices is already available. They involve moving beyond the simple modal practice of assessing internal consistency (Cronbach's α) to investigating the stability of scores across time (test-retest reliability), examining the factor structure of the latent construct (or constructs; confirmatory factor analysis, or CFA), and testing for the equivalence of measurement properties across populations, time points, and contexts (measurement invariance; Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). Although tests such as Cronbach's α and test-retest reliability are widely known and frequently reported, other tests of structural validity, such as measurement invariance, are poorly understood and infrequently conducted, despite their equal importance for theorizing (Flake et al., 2017). Indeed, if evidence for measurement invariance is not obtained—which is typically the case—then it is difficult to determine if a given measure reflects the same construct across samples, contexts, and conditions (see the Results section for a more detailed treatment of different types of structural-validity assessment).

Purpose of the Present Study

In short, measurement validity is central to theory and research in social and personality psychology. Yet rigorous tests of validity are rarely conducted or reported. This widespread tendency to underreport tests of validity leaves the field in a sticky situation: It is currently impossible to know whether the field is facing a mere problem of underreporting (as Flake et al., 2017, highlighted) or the potentially deeper issue of hidden invalidity. It may be that many of the measures used appear to be perfectly adequate on the surface and yet fall apart when subjected to more rigorous tests of validity beyond Cronbach's α.

With this in mind, we used several best practices to examine the structural validity of 15 well-known self-report measures that are often used in social and personality psychology (see Table 1). This unique case study demonstrates what can be achieved when best practices are followed in applying a wide number of

validity metrics to a large number of measures, each tested in a large sample, and reporting the results. For our investigation, we used data from the Attitudes, Identities, and Individual Differences (AIID) study, a large-scale, multivariate, planned-missing-data study that was conducted via the Project Implicit website (implicit.harvard.edu) between 2004 and 2007 (Hussey et al., 2019).

Utilizing this data set provided several advantages and unique opportunities. First, the sheer size of the sample involved (N = 81,986 individuals, N = 144,496 experimental sessions) allowed us to assess the psychometric properties of the 15 measures with numbers that were far greater than those used in many earlier validation studies. Second, the data set's structure allowed us to apply a large range of structural-validity metrics to the same measure in the same study, and to include tests of stability (test-retest reliability) based on multiple delay ranges (immediate vs. up to 1 year). Third, we were able to adopt a comprehensive strategy to structural-validity testing that extended beyond the strategies of previous studies in both nuance and scope. Following best practices, we obtained metrics of consistency (Cronbach's α; McDonald's, 1999, ω_t and ω_h), test-retest reliability (both dependability and stability; Revelle & Condon, 2018), factor structure (CFA), and measurement invariance. Although some of these tests have been applied to some of the scales we examined, this was often done separately, study by study and sample by sample, never comprehensively within and across a range of measures, as in the present study. Fourth, the recent explosion in Internet-based research and renewed reliance on self-report scales within social and personality psychology (Bohannon, 2016; Gosling & Mason, 2015; Sassenberg & Ditrich, 2019) has led to a situation in which many self-report scales are being used in contexts, and with samples, that differ from those in which they were originally validated. If researchers wish to use these measures in online settings, it is imperative that their structural validity be examined in that context to ensure that their psychometric properties are adequate and do not diverge from those observed in traditional (laboratory) settings.

We conducted the tests and report their results in order of the frequency with which the tests are reported in the literature (see Flake et al., 2017). We adopted this strategy in order to demonstrate the inverse relationship between rate of reporting and hidden invalidity. Note we are not suggesting that other researchers should sequence their analyses or reporting in a similar way. Indeed, as argued elsewhere (Flake et al., 2017) the most common test (α) relies on numerous assumptions that can be assessed only by less commonly applied analyses (e.g., within a CFA context).

Before we continue, we want to be clear: Our goal was not to make a final or absolute determination on the validity of any of the scales we assessed, to make a binary determination of their validity or invalidity, or even to present our analytic strategy as a prescriptive set of standards for future work. This is not to say that our results cannot provide input into the ongoing process of validating these scales. Rather, our primary goal was to investigate the issue highlighted by Flake et al. (2017), namely, whether the widespread underreporting of structural-validity information reflects hidden validity or, more worryingly, hidden invalidity.

Disclosures

Preregistration

Our analyses were not preregistered.

Data, materials, and online resources

All code and data to reproduce our analyses are available at the Open Science Framework, at osf.io/23rzk. Additional information on questionnaire items, results not reported here, simplified R scripts for educational purposes, and a change log documenting differences between manuscript versions are also available in supplementary materials at the Open Science Framework, at osf.io/2zx64.

Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study.

Ethical approval

Ethical approval for the underlying AIID study and data set was granted by the University of Virginia's Institutional Review Board for the Social and Behavioral Sciences (Protocol 2003-0173-00). As these data were collected between 2004 and 2007, this study is technically not in accordance with the most recent version (2013) of the Declaration of Helsinki, which requires preregistration prior to data collection. Ethical approval was not required for our analysis of the existing data.

Table 1. *Summary of the Structural-Validity Analyses*

Scale	Internal consistency	Test-retest reliability	Confirmatory factor structure	Measurement invariance	Global structural validity
Balanced Inventory of Desirable Responding					
Impression Management	Good	Good	Mixed	Poor	Questionable
Self-Deception Enhancement	Good	Good	Poor	Poor	Questionable
Bayesian Racism Scale	Good	Good	Good	Poor	Questionable
Belief in a Just World Scale: General Just World scale	Good	Good	Good	Poor	Questionable
Big Five Inventory					
Agreeableness	Good	Good	Mixed	Poor	Questionable
Conscientiousness	Good	Good	Good	Poor	Questionable
Extraversion	Good	Good	Mixed	Poor	Questionable
Neuroticism	Good	Good	Mixed	Poor	Questionable
Openness	Good	Good	Poor	Poor	Questionable
Humanitarianism-Egalitarianism Scale	Good	Good	Good	Poor	Questionable
Intuitions About Controllability and Awareness of Thoughts scales					
Others	Good	Good	Poor	Poor	Questionable
Self	Good	Good	Poor	Poor	Questionable
Need for Cognition Scale	Good	Good	Good	Good	Good
Need for Cognitive Closure Scale					
Ambiguity	Poor	Good	Good	Poor	Questionable
Closed-mindedness	Poor	Good	Mixed	Poor	Questionable
Decisiveness	Good	Good	Good	Poor	Questionable
Order	Good	Good	Good	Poor	Questionable
Predictability	Good	Good	Good	Poor	Questionable
Personal Need for Structure Scale	Good	Good	Mixed	Poor	Questionable
Protestant Ethic Scale	Good	Good	Mixed	Poor	Questionable
Right-Wing Authoritarianism Scale	Good	Good	Mixed	Poor	Questionable
Rosenberg Self-Esteem Scale	Good	Good	Good	Poor	Questionable
Self-Monitoring Scale	Good	Good	Poor	Poor	Questionable
Social Dominance Orientation scale	Good	Good	Poor	Poor	Questionable
Spheres of Control Battery					
Interpersonal Control	Good	Good	Good	Poor	Questionable
Personal Efficacy	Poor	Good	Poor	Poor	Questionable
Summary	88%	100%	73%	4%	4%

Note: The criterion for good internal consistency was McDonald's (1999, chap. 6) criterion of $\omega_t \geq .7$. The criterion for good test-retest reliability was $r \geq .7$ for both dependability ($\leq$ 1-hr delay) and stability (delay between 1 day and 1 year). A scale was labeled as having "good" confirmatory model fit if it met the criteria for all four metrics: comparative-fit index (CFI) ≥ 0.95, Tucker-Lewis fit index ≥ 0.95, root mean square error of approximation (RMSEA) ≤ 0.06, and root mean squared residual (SRMR) ≤ 0.09; confirmatory model fit was labeled "mixed" if the model met the SRMR criterion and any one or two of the other three criteria (see Hu & Bentler, 1999) and was labeled "poor" if the model met none of the other three criteria. Measurement invariance was labeled "good" if the scale met the criteria for configural invariance (the same criteria as for mixed confirmatory model fit) and metric and scalar invariance ($\Delta CFI \geq -0.015$ and $\Delta RMSEA \leq 0.01$; see Chen, 2007) for both age and gender groups. Global structural validity was labeled "good" if internal consistency, test-retest reliability, confirmatory factor structure, and measurement invariance were all labeled "good" or "mixed." The summary row indicates the percentage of scales in each column that were not labeled "poor."

Method

Participants

The data of 144,496 experimental sessions involving 81,986 unique participants (50,141 women and 31,845 men; mean age = 30.84, SD = 11.40) were selected for inclusion from the AIID data set on the basis that the participants met our predefined study criteria (i.e., age 18–65, self-reported fluency in English, and complete data on the individual-differences measures and demographics items). Table 2 lists the sample size for each measure. Repeat participation in the study was possible and allowed for the assessment of test-retest reliability. The modal number of participations was 1 ($M = 1.76$, $SD = 2.22$).

Measures

Fifteen individual-differences questionnaires were selected for inclusion in this study on the basis of their availability in the AIID data set. Five of these questionnaires had a particularly large number of items and were subdivided into two parts that were delivered between participants because of time constraints on the Project Implicit site. This resulted in participants being assigned to 1 of 20 different versions of the study materials. The 15 questionnaires and their subdivisions for purposes of the AIID study were as follows: the Balanced Inventory of Desirable Responding (Version 6; Paulhus, 1988; cited in Robinson, Shaver, & Wrightsman, 1991; Impression Management scale vs. Self-Deception Enhancement scale), Bayesian Racism Scale (Uhlmann, Brescoll, & Machery, 2010), Belief in a Just World Scale (General Just World scale only; Dalbert, Lipkus, Sallay, & Goch, 2001), Big Five Inventory (John & Srivastava, 1999; Extraversion, Conscientiousness, and Neuroticism scales vs. Agreeableness and Openness scales), Humanitarianism-Egalitarianism Scale (Katz & Hass, 1988), Intuitions About Controllability and Awareness of Thoughts scales (Nosek, 2012; Self scale vs. Others scale), Need for Cognition Scale (Cacioppo, Petty, & Kao, 1984), Need for Cognitive Closure Scale (Webster & Kruglanski, 1994; Order and Ambiguity scales vs. Predictability, Decisiveness, and Closed-mindedness scales), Personal Need for Structure Scale (Neuberg & Newsom, 1993), Protestant Ethic Scale (Katz & Hass, 1988), Right-Wing Authoritarianism Scale (Altemeyer, 1981), Rosenberg Self-Esteem Scale (Rosenberg, 1965), Self-Monitoring Scale (Snyder, 1987), Social Dominance Orientation scale (Scale 4; Pratto, Sidanius, Stallworth, & Malle, 1994), and Spheres of Control Battery (Paulhus, 1983; Interpersonal Control scale vs. Personal Efficacy scale).

Fourteen of these questionnaires had previously been employed in a study reported in a published article or book chapter, whereas one (the Intuitions about Controllability and Awareness of Thoughts scales) had not. The psychometric properties of all measures that had been used in previous publications had been examined to at least some extent, with one exception (i.e., the Bayesian Racism Scale, which has been used to make theoretical conclusions without a published validation study). As implemented in the AIID study, the questionnaires employed between 6 and 44 items ($M = 19.5$, $SD = 11.8$) and between 1 and 5 scales ($M = 1.7$, $SD = 1.4$). All scales used the same response format, a Likert scale ranging from 1 (strongly disagree) to 6 (strongly agree). In some cases, the response format differed from the measure's original format, and when significant modifications were made (i.e., change from a dichotomous to a Likert response format), they were carried out in accordance with recommendations in the literature (Dalbert et al., 2001; Stöber, Dette, & Musch, 2002). The wording of a minority of items in several measures was adjusted to make them more appropriate for a general rather than student sample (see the supplementary materials at https://osf.io/2zx64/).

Procedure

In this section, we provide a brief overview of the AIID study (for a more detailed description, see Hussey et al., 2019). Prior to the study, participants voluntarily navigated to the Project Implicit research website, created a unique log-in name and password, and provided demographic information. Those assigned to the AIID study then provided informed consent and completed one Implicit Association Test (Greenwald, McGhee, & Schwartz, 1998) and a subset of self-report measures from an attitudes battery. The IAT and the self-report measures centered on the same attitude domain, selected from a set of 95. Each domain consisted of two concept categories that were related to social groups, political ideologies, preferences, or popular concepts from the wider culture (e.g., African Americans vs. European Americans, Democrats vs. Republicans, coffee vs. tea, or Lord of the Rings vs. Harry Potter). Following the IAT and self-report ratings, participants were randomly assigned to complete 1 of the 20 versions of the individual-differences self-report measures.

In the current study, we made use of data only from the demographics questionnaire (age, gender, and English fluency) and individual-differences measures. Given that people completed only a small subset of the total available measures in any one session, repeat participation in the AIID study was allowed. No restrictions were placed on the time between experimental sessions (i.e., individuals could compete one session immediately after another or up to several years later). In order to maintain a consistent analytic strategy, we analyzed each questionnaire's scales separately. This is consistent with past use of these questionnaires in the almost all cases.

Results

Data preparation

Analyses for a given questionnaire were conducted on data obtained from the first experimental session in which participants completed that questionnaire, with the exception that test-retest reliability analyses were conducted on data obtained from the first two sessions in which participants completed the questionnaire. Reverse scoring of items was conducted according to the recommendations of each scale's original publication.

Analytic strategy

For each scale, we calculated both distributional information and multiple metrics of structural validity (see Tables 2–5), following the recommendations of Flake et al. (2017) and Revelle and Condon (2018). Distributional information (mean, standard deviation, skewness, and kurtosis) was calculated from each scale's sum scores. All analyses were implemented using the R packages lavaan (Rosseel, 2012) and semTools (Jorgensen et al., 2019). Confidence intervals were bootstrapped via the case-removal and quantile method using 1,000 resamples, and were implemented using the R packages *rsample* (Kuhn, Chow, & Wickham, 2019) and *purrr* (Henry & Wickham, 2019).

For all scales, we employed simple measurement models that did not involve method factors (e.g., negatively worded items) or item cross-loadings. We did so for three reasons. First, this uniform analytic strategy allowed us to compare rates of validity across scales, to address our primary research question. Second, with few exceptions (e.g., the Big Five Inventory), the "true" measurement model for most scales either is a matter of long debate (e.g., the Rosenberg Self-Esteem scale; see Mullen, Gothe, & McAuley, 2013; Salerno, Ingoglia, & Lo Coco, 2017; Supple, Su, Plunkett, Peterson, & Bush, 2013; Tomas & Oliver, 1999) or has of yet received no scrutiny (e.g., the Bayesian Racism Scale). Therefore, choices to employ alternative models would be exploratory or weakly informed, and comparisons among these models would detract from answering our primary research question. Third, most researchers who use these scales simply calculate sum scores and rely on these in their subsequent analyses. In doing so, they are tacitly endorsing simple measurement models with no cross-loadings or method factors (Rose, Wagner, Mayer, & Nagengast, 2019). Adopting similar assumptions meant that our findings would reflect how these scales are commonly used and interpreted.

The use of cutoff values for decision making has both potential benefits and potential costs, and results thus obtained should be interpreted with caution (Hu & Bentler, 1999). Following the recommendations of Vandenberg and Lance (2000), we report full results for all tests in order to allow researchers to apply their own decision-making methods if they so wish. Nonetheless, the decision whether or not to employ a scale in a future study is arguably a dichotomous decision, and therefore binary recommendations are useful in many cases. This is particularly true for researchers who do not have a background in psychometrics and want to rely on others' expertise to judge whether a scale is sufficiently valid for use. We therefore apply common and recommended cutoff values to all our test metrics in order to summarize and compare the relative validity of different scales across different aspects of structural validity.

Consistency. Given that researchers have argued that Cronbach's α is frequently misused and of limited utility (Flake et al., 2017; Schmitt, 1996; Sijtsma, 2009), we also used two less frequently reported but arguably superior metrics of internal consistency: McDonald's ω_t (omega total) and ω_h (omega hierarchical; McDonald, 1999, chap. 6); ω_t provides a metric of total measure reliability, or the proportion of variance that is attributable to sources other than measurement error., whereas ω_h provides a metric of factor saturation, or the proportion of variance that is attributable to a measure's primary factor (rather than additional factors or method factors; see Revelle & Condon, 2018). We employed $\omega_t \geq .7$ as the cutoff value for good internal consistency because this cutoff is typically used for α and the two metrics employ the same scale (Nunnally & Bernstein, 1994).

Dependability and stability. Test-retest reliability was estimated for the subset of participants with available data (n = 7,542) using Pearson's r correlations. We calculated two forms of test-retest reliability, according to the recommendations of Revelle and Condon (2018). First, test-retest dependability was calculated using the data from those participants who completed a scale twice within 1 hr. Second, test-retest stability was calculated using the data from those participants who completed a scale twice with a delay between 1 day and 1 year. Our cutoff value for both good test-retest dependability and good test-retest stability was $r \geq .7$, as is commonly recommended in the literature (Nunnally & Bernstein, 1994).

Factor structure. Because of the large number of scales, we specified and assessed the fit of measurement models using a standardized approach based on recommended best practices (Hu & Bentler, 1999; Rose et al., 2019). First, a confirmatory factor-structure model for each scale was defined using the items specified in the scales' original publication. No method factors (e.g., for negatively worded items) or item cross-loadings were included.

Given that an ordinal Likert response format was used for all scales, and that the amount of skew differed

among the scales, we employed the diagonally weighted least squares estimator along with robust standard errors of parameter estimates (i.e., the WLSMV, estimator option within lavaan). Simulation studies have shown that this estimator function is superior to the more common maximum likelihood method (Li, 2016). Performance of all the scales was poorer when the models were refitted using the maximum likelihood method, with or without robust standard errors.

Previous work has repeatedly suggested that multiple indices of a model's goodness of fit should be calculated and reported even if only a subset of these are used for decision-making purposes (Vandenberg & Lance, 2000). We therefore calculated the following indices: Our metrics of absolute fit were chi-square tests (although, given our sample sizes, the p values for these were universally significant and therefore uninformative; nonetheless, chi-square values should be reported), chi-square normalized by number of items, and the root mean squared residual (SRMR). Our measure of relative fit was the Tucker-Lewis fit index (TLI). Our noncentrality indices were the comparative-fit index (CFI) and root mean square error of approximation (RMSEA; its 95% confidence intervals were also calculated). For decision making regarding model fit, we employed the cutoffs suggested by Hu and Bentler (1999: i.e., SRMR $\leq$ 0.09, TLI $\geq$ 0.95, CFI $\geq$ 0.95, RMSEA $\leq$ 0.06). Hu and Bentler argued that basing model-fit decisions two fit indices, rather than one, lowers the combined rate of Type I and Type II errors. Specifically, they recommended that model-fit determinations be based on SRMR combined with one of the following: CFI, TLI, or RMSEA. However, having no strong prior preferences among these multiple fit indices, we labeled individual scales as demonstrating good or poor fit according to the fit observed with all three combinations. That is, if a scale demonstrated good fit with all three metric permutations (i.e., SRMR + CFI, SRMR + TLI, and SRMR + RMSEA), we labeled the fit as "good"; if the fit was good using one or two but not all three permutations, it was labeled as "mixed"; and if it was not good for any of the three permutations, it was labeled as "poor".[2]

Measurement invariance. Assessing a scale's capacity to measure the same construct comparably in different populations or contexts typically involves three component tests: tests of (a) configural invariance (i.e., equivalence of model form: whether the unconstrained model provides adequate fit in each of the groups), (b) metric invariance (or weak factorial invariance; i.e., equivalence of factor loadings), and (c) scalar invariance (or strong factorial invariance; i.e., equivalence of item intercepts, or thresholds; Putnick & Bornstein, 2016). These types of invariance are typically assessed with nested models; the initial measurement model is first fit to each group's data, a second fit constrains factor loadings to be equivalent, and a third fit constrains item intercepts (or thresholds) to be equivalent. Change in fit metrics between these nested models is then typically used to determine whether each test is passed in sequence. When a scale passes all three tests, one can conclude that correlations between scores on the scale and other external variables have equivalent interpretations across the groups. That is, individuals' observed scores on the scale are likely to measure the same latent variable and in a comparable way, regardless of the groups to which the individuals belong. Loosely speaking, one accessible interpretation of meeting measurement invariance is that individuals in the different groups interpret the items in an equivalent manner. Not meeting measurement invariance has important implications for the researcher: It is not possible to meaningfully interpret comparison between groups or associations between scores on the scale and external variables.

Although tests of measurement invariance are typically performed between groups that the researcher wants to compare directly, one should also assess measurement invariance between groups that one tacitly assumes should be invariant. For example, for many studies, researchers recruit adults (e.g., ages 18–65) and both men and women, but do not seek to make comparisons based on either age or gender, or to account for the influence of age or gender within their statistical models. In such cases, the researchers implicitly assume that the scales measure the same construct (or constructs) in different age groups and in both men and women. It is therefore useful to test these two assumptions, specifically, that the employed scales are invariant across gender and, for example, across individuals above versus below the median age in the sample. Similarly, if researchers explicitly wish to make comparisons between such categories (e.g., between men and women), measurement invariance is a requirement for these comparisons to be meaningful. For example, personality differences between men and women are theoretically meaningful only if they represent differences in latent means rather than factor loadings or intercepts. In all cases, measurement

[2] An alternative strategy of employing all four metrics for decision making was considered but ultimately rejected because there was no basis for this analytic strategy in the literature, which prevented us from making informed choices about cutoff values with this approach. Moreover, there was a high probability that employing additional metrics while using cutoff values recommended for a two-index decision-making format would raise the false-negative rate, or at minimum would introduce great uncertainty about the impact of such a strategy on the false-negative rate.

invariance is therefore a necessary prerequisite for subsequent substantive analyses. Therefore, we tested measurement invariance for individuals above versus below the median age in our sample (median age = 27) and for men versus women.

Historically, the most common method used to test measurement invariance was to assess the statistical significance of changes in absolute model fit (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). This was not suitable in the present study because of the sensitivity of chi-square tests to our large sample sizes. In addition, relying exclusively on the significance of chi-square tests, in place of alternative fit indices such as RMSEA, has fallen out of favor over time (Putnick & Bornstein, 2016). Numerous simulation studies have been conducted to explore which indices and cutoffs (if any) should be used. Recommended cutoff values have been described as ranging from liberal (e.g., Cheung & Rensvold, 2002) to conservative (e.g., Meade, Johnson, & Braddy, 2008), and the real-world applicability of these cutoffs is a matter of ongoing debate (Little, 2013). For tests of configural invariance, we elected to employ the same criteria as for mixed CFA fit (Hu & Bentler, 1999), and for tests of metric and scalar invariance, we chose to use Chen's (2007) moderate criteria of both $\Delta CFI \geq -0.015$ and $\Delta RMSEA \leq 0.01$. This two-metric strategy is broadly compatible with the criteria used for CFA and configural invariance fits, as well as being the modal reporting practice according to a recent review (Putnick & Bornstein, 2016). The same estimator was used as in the CFA fits.

Results synthesis

A summary of the results for these metrics of structural validity using recommended cutoff values is presented in Table 1. This table provides a concise summary of the structural-validity evidence for each individual scale, as well as of the evidence across scales. Tables 2 through 5 provide the results for all statistical metrics and the aspects of structural validity to which they speak (i.e., internal consistency, test-retest reliability, factor structure, and measurement invariance for age and gender groups), along with details regarding each scale (number of participants, number of items), and distributional information (mean, standard deviation, skewness, kurtosis). When combined, Tables 1 through 5 provide a wide range of psychometric properties for 15 commonly used self-report individual-differences scales that could inform their future use. Full results of the tests of measurement invariance (i.e., results for each fit index for each test) are available in the supplementary materials. Additionally, recent research has quantified the impact of failure to meet measurement invariance as a continuous variable (e.g., Nye & Drasgow, 2011). Although this is beyond the scope of this article, the supplementary materials provide continuous estimates of the impact of measurement invariance on the magnitude of between-groups comparisons: For each between-groups comparison (i.e., participants above vs. below the median age, male vs. female participants), the between-groups effect size (Cohen's *d*) was calculated separately for the observed sum scores and the latent scores, and then the difference between these two estimates was computed.

The summary labels in Table 1 serve to condense multifaceted metrics of validity into categorical conclusions in order to enable decision making with regard to our core research question (i.e., whether the underreporting in the literature represents hidden validity or invalidity). This trade-off between nuance and heuristic value is analogous to the use of p values, which are natively continuous, but which are often reduced to a significant-versus-nonsignificant dichotomy to facilitate conclusions regarding hypotheses. These categorical labels should not be taken as claims about literal truth for any research question other than our own (e.g., for assessing the adequacy of a scale for future use). Instead, such questions should be informed by the continuous and multifaceted results reported in Tables 2 through 5, which offer a more nuanced perspective on structural validity.

Table 2. *Sample Sizes and Internal-Consistency Results*

Scale	Total *n*	Number of items	Internal consistency		
			α	ω_t	ω_h
Balanced Inventory of Desirable Responding					
Impression Management	6,934	18	.797 [.789, .804]	.798 [.791, .805]	.796 [.789, .803]
Self-Deception Enhancement	6,713	18	.703 [.692, .714]	.708 [.697, .720]	.707 [.697, .719]
Bayesian Racism Scale	6,532	16	.824 [.818, .831]	.828 [.822, .835]	.822 [.815, .829]
Belief in a Just World Scale: General Just World scale	6,758	6	.754 [.744, .765]	.760 [.751, .771]	.761 [.751, .772]
Big Five Inventory					
Agreeableness	6,713	9	.792 [.784, .800]	.793 [.784, .800]	.788 [.779, .796]
Conscientiousness	6,649	9	.820 [.812, .827]	.820 [.812, .826]	.810 [.802, .817]
Extraversion	6,649	8	.869 [.864, .874]	.872 [.867, .877]	.869 [.865, .874]
Neuroticism	6,649	8	.832 [.826, .839]	.834 [.828, .840]	.832 [.826, .838]
Openness	6,713	10	.793 [.785, .801]	.792 [.783, .801]	.784 [.774, .793]
Humanitarianism-Egalitarianism Scale	6,906	10	.840 [.831, .847]	.839 [.831, .847]	.830 [.820, .839]
Intuitions About Controllability and Awareness of Thoughts scales					
Others	6,711	9	.750 [.740, .761]	.753 [.743, .763]	.744 [.733, .754]
Self	6,830	9	.797 [.789, .804]	.800 [.792, .807]	.801 [.793, .808]
Need for Cognition Scale	6,649	18	.889 [.885, .893]	.889 [.885, .893]	.885 [.880, .889]
Need for Cognitive Closure Scale					
Ambiguity	6,585	9	.674 [.661, .686]	.683 [.670, .695]	.680 [.667, .693]
Closed-mindedness	6,559	8	.641 [.627, .655]	.638 [.622, .652]	.631 [.615, .646]
Decisiveness	6,559	7	.816 [.809, .823]	.824 [.817, .831]	.825 [.818, .832]
Order	6,585	10	.819 [.811, .826]	.825 [.818, .832]	.824 [.817, .831]
Predictability	6,559	8	.793 [.784, .801]	.796 [.787, .804]	.795 [.786, .803]
Personal Need for Structure Scale	6,821	12	.861 [.855, .865]	.862 [.857, .866]	.860 [.854, .864]
Protestant Ethic Scale	6,859	11	.791 [.783, .798]	.791 [.783, .798]	.782 [.773, .789]
Ring-Wing Authoritarianism Scale	6,542	20	.922 [.919, .924]	.922 [.919, .925]	.910 [.907, .914]
Rosenberg Self-Esteem Scale	6,971	10	.890 [.886, .895]	.896 [.892, .900]	.887 [.882, .892]
Self-Monitoring Scale	6,623	18	.759 [.750, .768]	.760 [.749, .770]	.740 [.723, .755]
Social Dominance Orientation scale	6,854	12	.831 [.824, .837]	.831 [.824, .837]	.821 [.814, .828]
Spheres of Control Battery					
Interpersonal Control	6,785	10	.808 [.801, .816]	.810 [.803, .818]	.808 [.800, .816]
Personal Efficacy	6,899	10	.641 [.627, .654]	.638 [.623, .651]	.623 [.607, .637]

Note: Total *n* refers to the total number of participants with data available for all the analyses reported in Tables 2 through 5 other than test-retest dependability and stability. Values inside brackets are 95% confidence intervals.

Table 3. *Distributional Statistics and Results for Test-Retest Reliability*

Scale	Distribution				Test-retest dependability		Test-retest stability	
	M	*SD*	Skewness	Kurtosis	*n*	*r*	*n*	*r*
Balanced Inventory of Desirable Responding								
Impression Management	58.61	13.38	0.01	2.99	149	.90 [.85, .94]	158	.77 [.67, .84]
Self-Deception Enhancement	63.38	10.25	0.06	3.22	173	.89 [.83, .93]	132	.76 [.66, .83]
Bayesian Racism Scale	41.09	13.16	0.40	2.91	136	.84 [.74, .90]	154	.88 [.82, .92]
Belief in a Just World Scale: General Just World scale	0.07	5.64	0.07	2.67	170	.88 [.82, .92]	150	.74 [.65, .81]
Big Five Inventory								
Agreeableness	38.90	7.15	–0.28	2.82	158	.95 [.91, .97]	154	.86 [.81, .90]
Conscientiousness	38.84	7.45	–0.29	2.79	144	.94 [.91, .96]	124	.84 [.78, .89]
Extraversion	31.16	8.14	–0.13	2.49	144	.94 [.90, .97]	124	.91 [.85, .95]
Neuroticism	26.70	7.67	0.06	2.65	144	.88 [.81, .93]	124	.87 [.82, .91]
Openness	47.10	7.17	–0.51	3.17	158	.92 [.87, .95]	154	.88 [.84, .92]
Humanitarianism-Egalitarianism Scale	50.69	7.02	–1.20	5.14	131	.91 [.84, .95]	149	.85 [.75, .90]
Intuitions About Controllability and Awareness of Thoughts scales								
Others	23.81	6.74	0.29	3.23	158	.75 [.65, .84]	127	.76 [.67, .82]
Self	30.22	7.86	0.16	2.99	195	.86 [.80, .90]	143	.78 [.69, .85]
Need for Cognition Scale	80.91	13.8	–0.44	2.98	147	.85 [.75, .91]	133	.86 [.80, .90]
Need for Cognitive Closure Scale								
Ambiguity	38.27	6.24	–0.23	3.13	119	.87 [.80, .92]	120	.71 [.61, .79]
Closed-mindedness	22.32	5.44	0.09	2.94	129	.84 [.77, .89]	150	.82 [.75, .88]
Decisiveness	25.96	7.20	–0.13	2.46	129	.93 [.90, .96]	150	.88 [.84, .91]
Order	38.83	8.76	–0.12	2.73	119	.85 [.69, .93]	120	.87 [.81, .91]
Predictability	28.65	7.02	–0.01	2.86	129	.85 [.79, .90]	150	.88 [.83, .92]
Personal Need for Structure Scale	42.10	10.15	–0.06	2.82	149	.88 [.82, .92]	141	.81 [.71, .88]
Protestant Ethic Scale	41.06	8.69	–0.18	3.05	163	.92 [.88, .94]	158	.85 [.80, .89]
Ring-Wing Authoritarianism Scale	51.30	18.84	0.48	2.45	116	.96 [.92, .98]	163	.94 [.91, .96]
Rosenberg Self-Esteem Scale	46.91	9.55	–0.80	3.35	160	.95 [.92, .96]	156	.90 [.86, .93]
Self-Monitoring Scale	63.18	12.20	0.03	2.85	140	.91 [.86, .94]	157	.87 [.82, .91]
Social Dominance Orientation scale	25.58	9.85	0.71	2.98	161	.91 [.86, .94]	149	.84 [.78, .88]
Spheres of Control Battery								
Interpersonal Control	42.57	8.32	–0.43	3.06	127	.90 [.85, .94]	138	.81 [.72, .87]
Personal Efficacy	45.41	6.12	–0.32	3.11	164	.83 [.72, .91]	152	.81 [.74, .86]

Note: Reliability refers to the correlation between scores when the retest occurred within 1 hr of the initial test, and stability refers to the correlation between scores when the retest occurred between 1 day and 1 year after the initial test. Values inside brackets are 95% confidence intervals.

Table 4. *Results of the Confirmatory Factor Analysis*

Scale	χ^2 [a]	χ^2/df	*df*	*CFI*	*TLI*	*RMSEA*	*SRMR*
Balanced Inventory of Desirable Responding							
Impression Management	1,834	13.6	135	0.950	0.944	0.043 [0.041, 0.044]	0.040
Self-Deception Enhancement	3,743	27.7	135	0.819	0.795	0.063 [0.061, 0.065]	0.059
Bayesian Racism Scale	1,667	16.0	104	0.965	0.960	0.048 [0.046, 0.050]	0.046
Belief in a Just World Scale: General Just World scale	153	17.0	9	0.986	0.977	0.049 [0.042, 0.056]	0.031
Big Five Inventory							
Agreeableness	735	27.2	27	0.964	0.952	0.063 [0.059, 0.066]	0.050
Conscientiousness	610	22.6	27	0.976	0.968	0.057 [0.053, 0.061]	0.047
Extraversion	776	38.8	20	0.979	0.970	0.075 [0.071, 0.080]	0.055
Neuroticism	589	29.5	20	0.978	0.969	0.065 [0.061, 0.070]	0.049
Openness	1,152	32.9	35	0.948	0.933	0.069 [0.066, 0.072]	0.060
Humanitarianism-Egalitarianism Scale	485	13.9	35	0.979	0.972	0.043 [0.040, 0.047]	0.048
Intuitions About Controllability and Awareness of Thoughts scales							
Others	1,378	51.0	27	0.903	0.870	0.086 [0.083, 0.090]	0.073
Self	1,509	55.9	27	0.929	0.906	0.090 [0.086, 0.094]	0.072
Need for Cognition Scale	1,959	14.5	135	0.978	0.975	0.045 [0.043, 0.047]	0.044
Need for Cognitive Closure Scale							
Ambiguity	471	13.4	35	0.986	0.982	0.043 [0.040, 0.047]	0.035
Closed-mindedness	440	22.0	20	0.931	0.904	0.057 [0.052, 0.061]	0.042
Decisiveness	301	21.5	14	0.986	0.979	0.056 [0.051, 0.061]	0.039
Order	260	9.6	27	0.973	0.965	0.036 [0.032, 0.040]	0.030
Predictability	374	18.7	20	0.979	0.971	0.052 [0.047, 0.057]	0.040
Personal Need for Structure Scale	1,472	27.3	54	0.969	0.962	0.062 [0.059, 0.065]	0.055
Protestant Ethic Scale	1,244	28.3	44	0.951	0.939	0.063 [0.06, 0.066]	0.056
Ring-Wing Authoritarianism Scale	6,647	39.1	170	0.959	0.954	0.076 [0.075, 0.078]	0.076
Rosenberg Self-Esteem Scale	875	25.0	35	0.982	0.977	0.059 [0.055, 0.062]	0.057
Self-Monitoring Scale	11,631	86.2	135	0.689	0.648	0.113 [0.112, 0.115]	0.102
Social Dominance Orientation scale	1,785	33.1	54	0.946	0.934	0.068 [0.066, 0.071]	0.064
Spheres of Control Battery							
Interpersonal Control	829	23.7	35	0.965	0.955	0.058 [0.054, 0.061]	0.049
Personal Efficacy	1,490	42.6	35	0.836	0.790	0.078 [0.074, 0.081]	0.065

Note: Values inside brackets are 95% confidence intervals. CFI = comparative-fit index; TLI = Tucker-Lewis fit index; RMSEA = root mean square error of approximation; SRMR = root mean squared residual.

[a]For all χ^2 tests, $p < .001$.

Table 5. *Results of the Tests of Measurement Invariance*

Scale	Age groups[a]		Gender groups		Combined results
	Overall result	Test failed	Overall result	Test failed	
Balanced Inventory of Desirable Responding					
Impression Management	Failed	Scalar	Failed	Scalar	Failed
Self-Deception Enhancement	Failed	Configural	Failed	Configural	Failed
Bayesian Racism Scale	Failed	Configural	Failed	Scalar	Failed
Belief in a Just World Scale: General Just World scale	Failed	Scalar	Passed	-	Failed
Big Five Inventory					
Agreeableness	Failed	Configural	Failed	Configural	Failed
Conscientiousness	Failed	Configural	Failed	Configural	Failed
Extraversion	Failed	Configural	Failed	Configural	Failed
Neuroticism	Failed	Configural	Failed	Configural	Failed
Openness	Failed	Configural	Failed	Configural	Failed
Humanitarianism-Egalitarianism Scale	Failed	Configural	Failed	Configural	Failed
Intuitions About Controllability and Awareness of Thoughts scales					
Others	Failed	Configural	Failed	Configural	Failed
Self	Failed	Configural	Failed	Configural	Failed
Need for Cognition Scale	Passed	-	Passed	-	Passed
Need for Cognitive Closure Scale					
Ambiguity	Failed	Scalar	Failed	Scalar	Failed
Closed-mindedness	Failed	Configural	Failed	Configural	Failed
Decisiveness	Failed	Configural	Failed	Configural	Failed
Order	Failed	Configural	Failed	Configural	Failed
Predictability	Failed	Configural	Failed	Configural	Failed
Personal Need for Structure Scale	Failed	Configural	Failed	Configural	Failed
Protestant Ethic Scale	Failed	Configural	Failed	Configural	Failed
Ring-Wing Authoritarianism Scale	Failed	Configural	Failed	Configural	Failed
Rosenberg Self-Esteem Scale	Failed	Configural	Failed	Configural	Failed
Self-Monitoring Scale	Failed	Configural	Failed	Configural	Failed
Social Dominance Orientation scale	Failed	Configural	Failed	Configural	Failed
Spheres of Control Battery					
Interpersonal Control	Failed	Configural	Failed	Configural	Failed
Personal Efficacy	Failed	Configural	Failed	Configural	Failed

Note: A scale passed the test of measurement invariance if it met the criteria for configural invariance, metric invariance, and scalar invariance, as outlined in the footnote in Table 1. Full results of each test of measurement invariance are available in the supplementary materials (osf.io/2zx64).

[a]Age groups were formed using a median spit (median age = 27).

Discussion

The reproducibility and replicability of research findings, as well as confidence in theory and application, require valid measures. Yet as Flake et al. (2017) pointed out, structural validity is rarely reported in the literature, and even when it is, the reported tests are usually restricted to a single and flawed index (Cronbach's α). This raises the question: Is the underreporting of tests of structural validity a mere nuisance, insofar as these measures are in fact valid, or, more troublingly, is there an abundance of invalid measures hiding in plain sight (i.e., hidden invalidity)? To examine this question, we submitted 15 self-report measures from social and personality psychology to a comprehensive battery of structural-validity tests (i.e., we examined their distribution, consistency, test-retest reliability, factor structure, and measurement invariance for gender groups and age groups defined by a median split). Doing so seems timely and necessary given the broader reevaluation of modal practices taking place in psychological science (Munafò et al., 2017) and a growing reliance on self-report data collected from online samples (Sassenberg & Ditrich, 2019).

Before unpacking our findings it seems useful to distinguish between two concepts: the weight of evidence (e.g., presence and quality of evidence, ranging weak to strong) and the nature of conclusions (e.g., given that evidence, what should one conclude about a measure's validity, on a continuum ranging from "good" to "questionable" to "poor"?). We argue that our results have strong evidential weight insofar as they were derived from a large and diverse sample ($n > 6{,}500$ per scale), were obtained across follow-up periods, and were obtained using a wider-than-usual variety of structural-validity metrics applied to many different scales. Indeed, to the best of our knowledge, this is the first study to consider the full range of metrics of structural validity, including multiple metrics of internal consistency, test-retest reliability, confirmatory factor structure, and measurement invariance, and the first to simultaneously apply them to so many scales. We also acknowledge our study's potential evidential weaknesses, in that recruitment was from a single population (i.e., an online sample) and that we considered only the structural phase of validity assessment but not the external phase.

To develop a conclusion, we employed a dichotomization strategy to synthesize the results across the scales. Most of the scales passed certain tests of structural validity: Specifically, 89% demonstrated good internal consistency, and 100% demonstrated good test-retest reliability. Yet many failed other tests of structural validity: Only 73% demonstrated good fit with the expected factor structure, and a surprisingly tiny fraction (4%) demonstrated measurement invariance for both age and gender groups. Only a single scale (Need for Cognition) passed all four metrics and can be said to have good global structural validity. Our results therefore appear to suggest that the widespread underreporting of structural validity highlighted by Flake et al. (2017) may reflect hidden invalidity. Why would this be the case given that most of these scales are widely used throughout psychological science?

One possibility is that invalidity may simply have been hidden until now: The full range of metrics of structural validity has been reported for very few studies. Our findings support this idea, as the metrics the scales tended to pass or fail were not random. The scales were more likely to fail those validity metrics that have been less often reported in the literature (factor structure and measurement invariance). Conversely, the scales were more likely to pass those metrics that have been reported more often in the literature (Cronbach's α and test-retest *r*). Figure 1 illustrates this hierarchical, or Guttman, structure among the validity metrics. The correlation between failure rate and reporting rate highlights the potential for a general pattern of hidden invalidity throughout the discipline.

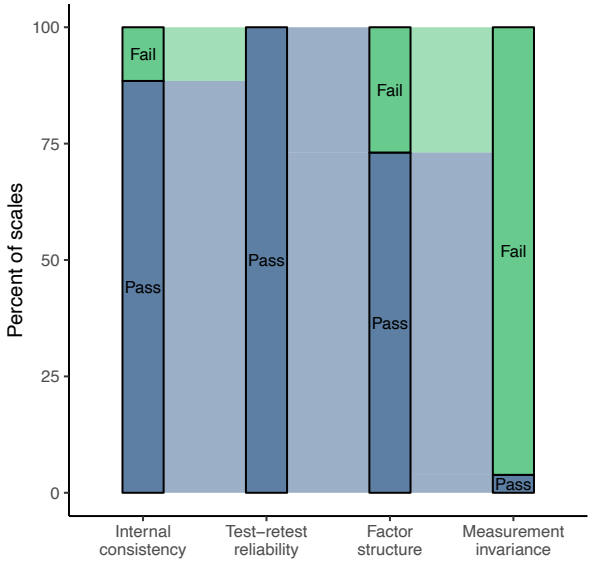

Figure 1. Alluvial plot illustrating the within- and between-scale patterns of passing or failing different tests of structural validity, arranged by frequency of reporting in the literature (from more common on the left to less common on the right).

The question then becomes, why was the structural fit and measurement invariance of these scales mixed or poor when their internal consistency and test-retest reliability were generally so good? One possibility that comes to mind is that tests of confirmatory factor structure and measurement invariance are inherently

stricter. A second is the measures used in the field of psychological science have been overoptimized to demonstrate good consistency, to the detriment of other psychometric properties.

To understand this idea more clearly, imagine that a researcher sets out to develop a new scale assessing negative automatic thoughts among people with depression. After constructing her scale, she attempts to determine how reliable it is, calculates Cronbach's α, and obtains a value of .60. As things currently stand, reviewers and users of the scale might comment that this value is problematically low. The researcher might then spend her limited time and resources attempting to improve α so that it tips over the commonly used and sought-after (yet arbitrary) .70 cutoff, for example, by excluding or rewording items and testing a new version of the scale. As a consequence, she would be less likely to spend her finite resources assessing and attempting to improve other aspects of the scale's structural validity, such as measurement invariance between groups. Yet doing so might have a larger payoff than chasing α: If the scale does not meet criteria of measurement invariance, subsequent longitudinal research using the scale with, for example, depressed individuals before and after therapeutic intervention might lead researchers to incorrectly infer that those individuals changed in terms of the latent variable (e.g., automatic thoughts in depression), when in fact they might simply have interpreted the items differently across the two measurement time points. For example, the therapeutic intervention might not serve to decrease the frequency of automatic thoughts (i.e., produce changes in the underlying latent variable), but instead might increase participants' introspective abilities to more accurately report on the frequency of those thoughts (i.e., only the measurement properties of the scale might have changed). In other words, researchers might incorrectly infer that the intervention is effective in decreasing negative automatic thoughts in depression when in fact it is not.

In short, we are not arguing that internal consistency should be neglected, but rather saying only that it (via Cronbach's α) should not be the sole focus in assessing structural validity, especially given its various flaws (Flake et al., 2017). Instead, researchers should adopt a more considered perspective by probing structural validity from multiple angles, especially those relevant to the context in which the scale in question is likely to be used (e.g., measurement invariance for known groups, test-retest reliability for longitudinal research). Failing to do so risks overoptimizing the measure on a flawed metric and without regard to other important but often overlooked properties.

Of course, the two possible explanations of the tests' differential failure rates (i.e., relative strictness of the tests vs. overoptimization on internal consistency to the neglect of other forms of validity) are not necessarily mutually exclusive. Regardless of the explanation, more rigorous reporting of these metrics is required.

We have also considered a number of other factors—none of which are incompatible with hidden invalidity—that could have contributed to our results. One possibility is that the scales themselves are less than optimal measures of the construct (or constructs) of interest. This could be for several reasons. For example, the items may be more poorly worded than previously appreciated, or the structure among the items may not be as originally assumed. It may also be the case that responding in this sample was influenced by factors that are theoretically relevant; for example, the scales may have unintentionally measured closely related but previously unappreciated constructs. Or responses may have been influenced by theoretically irrelevant factors (e.g., low-quality responding, demand effects, additional latent factors, or item cross-loading among these factors). Indeed, articles considering the confirmatory factor structure of established measures frequently reject the expected model and suggest alternative models with different latent-variable structures or item cross-loadings (e.g., the Rosenberg Self-Esteem Scale: Mullen et al., 2013; Salerno et al., 2017; Supple et al., 2013; Tomas & Oliver, 1999). In many cases, despite subsequent work suggesting that the factor structure is not what the scale's creators originally conceived or evidence that certain items should be dropped or modified, scales are most commonly used with the originally posited items and interpreted according to the originally posited factor structures, so that there is something of a primacy bias in the use of many scales. Indeed, the resistance to incorporating emerging structural-validity evidence into scales' use (e.g., when researchers decide whether to use a given scale, how to score it, how to interpret its scores, and what variations in items or response options to use) is an ongoing issue for the field.

A second possibility is that there was something problematic about the current sample or that participants differed from those used during the original validation processes for these scales. We believe that this is unlikely given that the sample was, if not representative of the general population, far more representative than the samples typically used in laboratory-based research.

Finally, it is possible that scales demonstrated poor structural validity because the constructs they were intended to measure were poorly conceived in the first place (i.e., in the substantive phase of validation; Flake

et al., 2017) or poorly captured by the scale items. Although this may seem unlikely given how well-known many of the scales we tested are, allowing for such a possibility protects against the reification of a construct merely because a scale has been created to assess it. Scales for which such issues do exist could be improved (or even avoided) by following Tay and Jebb's (2018) recent suggestions for continuum specification. For instance, researchers could address issues of polarity ambiguity' within their scales. Do low scores on a scale (e.g., a perfectionism scale) represent the absence of the construct of interest (e.g., low or absent perfectionism) or the presence of its opposite (e.g., high carelessness)? Researchers could also address issues of gradation, that is, the quality, or dimension, separating low from high scores. Take, once again, the example of depression: Multiple scales are available to assess depression, but they differ in their dimension of gradation; one measures the frequency of depressive thoughts, but another measures the degree of belief in the literality of those thoughts, and yet another measures the experienced emotional intensity of those thoughts. The take-home message here is that well-developed frameworks for measurement development already exist for researchers looking to construct or refine their scales. We encourage researchers to make better use of them, including by attending to all three interrelated phases of validation (substantive, structural, external; Flake et al., 2017). Although we have focused on the second phase, all phases of this process must be attended to when making a holistic evaluation of a measure's validity. One phase is neither sufficient nor singularly important relative to the other two, nor should one phase be maximized at the expense of the others.

Implications and future directions

Our findings have implications for individual researchers in particular and for the field more generally. To understand why, imagine that a research team sets out to test a specific hypothesis using one of these scales (e.g., whether belief in a just world predicts some behavior of interest). They run their study and then assess if the scale they used provides a reliable index of the construct of interest. Behaving as most researchers do, they answer this question by examining the consistency of their data, and possibly the data's test-retest reliability. These tests tell the team that the scale demonstrates adequate validity. This necessity taken care of, they then proceed to what is, for them, the real meat of the issue—interpreting their findings relative to their original hypothesis. Yet our findings suggest that if the researchers were to adopt a more comprehensive assessment following best practices, they would discover that the underlying factor structure of their construct and its invariance across samples are problematic, and thus might exert more caution before interpreting their data. In other words, issues at the second phase of validation (structural) moderate researchers' ability to make claims at the third phase (external validation), such as claims about differences between known groups, interrelationships between latent constructs, and the prediction of behavior. Therefore, although questions concerning the structural validity of their measures may not be inherently appealing to all researchers, assessing structural validity is a requirement for making conclusions at other levels.

Another take-home message, one that we have not seen explicated elsewhere, is that a finding can be extremely replicable and yet give rise to invalid conclusions. For example, even if two groups (e.g., depressive and nondepressive individuals) were shown across multiple studies to differ in their observed mean scores on a given scale (e.g., the Rosenberg Self-Esteem Scale), this replicable finding would typically be interesting and useful only if it also reflects differences in a latent variable (e.g., self-esteem), rather than mere differences in how the two groups interpret the items in the questionnaire. In short, replicability does not equal validity. The potential for hidden structural invalidity therefore has implications for the conclusions made using a given scale.

What applies to an individual also applies to the field as a whole. Our findings highlight the possibility that hidden invalidity may be a common feature of many scales in the literature. The overwhelming majority of the scales we examined were found to be structurally invalid in some regard (at least in a categorical sense). As a thought experiment, imagine that the scales examined here are a representative subset of those used in social and personality psychology. If so, there are likely many other instances of hidden invalidity in other scales used in the field. Indeed, even if the true rate of hidden invalidity is only a fraction of that observed here, this would still bring the conclusions of a large number of studies using invalid scales into question. It is currently difficult to assess the true prevalence of hidden invalidity given that researchers often report, and reviewers and editors request, only a single metric of structural validity (Cronbach's α). Therefore, at worst, the literature may be unwittingly advancing a simplistic and overly positive view of how valid many of the most commonly used measures actually are, and reporting invalid conclusions based on these scales. At best, the hidden invalidity we observed may simply reflect underreporting of scales that will ultimately be shown to be valid. Yet until comprehensive reporting of tests of validity is common practice, one cannot know. We therefore encourage a more rigorous, multimetric

approach to structural validity across all areas of psychology, an approach in which researchers identify and report, and reviewers and editors request, multiple sources of validity evidence. Note that although we endorse more widespread assessment of structural validity, we are not prescribing how it should be done, or presenting the methods or any cutoff values we have used here as prescriptive recommendations. For pragmatic advice on improving measurement practices, we encourage readers to consult Flake and Fried (2019). That said, and for educational purposes, we have included in our supplementary materials simplified and commented R code to illustrate how we implemented our validity assessments.

Finally, two barriers limit the field's ability to reach the goals of increasing the frequency with which metrics of structural validity are used and reported and conditioning substantive claims on the basis of evidence: (a) the staggering degrees of freedom available to researchers when they assess the structural validity of their measures and (b) the fact that researchers are heavily motivated to conclude that their measures are valid in order to test their core hypotheses. Imagine, for instance, that a researcher accepts the importance of assessing structural validity and sets out to test the internal consistency, test-retest reliability, factor structure, and measurement invariance of a study's measures. In order to do so, the researcher would have to choose a specific metric for each validity dimension from the many available options, select a cutoff for each metric from among many recommended values, choose an implementation of each test from among multiple options that frequently differ in their results, and make choices among numerous less visible experimenter degrees of freedom. And this is not to mention all the potential interactions between these steps. In the absence of firm guidelines, one's decision-making pathway when choosing how to report structural validity is massively unconstrained, a garden of forking paths (Gelman & Loken, 2013).

This lack of constraint may lead to two practices that are equally detrimental to the reproducibility, replicability, and validity of research findings. Making an analogy with *p*-hacking (Simmons et al., 2011), we refer to the first practice as *v*-hacking: selectively choosing and reporting a combination of metrics, including their implementations and cutoffs, and taking advantage of other degrees of experimenter freedom so as to improve the apparent validity of measures. For example, in 2004, Watson noted that test-retest reliability studies were rarely conducted, but that when they were, authors "almost invariably concluded that their stability correlations were 'adequate' or 'satisfactory' regardless of the size of the coefficient or the length of the retest interval" (p. 326). Researchers may be driven to such conclusions given the current incentive structures for both research and applied work: In research, for example, reporting that a measure demonstrates adequate validity allows one to test one's core hypotheses using that measure, therefore increasing one's chances of being published (theories may be supported or questioned only on the basis of valid measures). Also, a valid measure is more likely to be adopted in applied settings; there are both financial and academic incentives for developing proprietary scales that are deemed valid.

The second practice we refer to as *v*-ignorance: relying on and reporting those metrics that other researchers have used, without considering the issues underlying their use. Indeed, a 2008 review of graduate training in psychology revealed that measurement theory and practice is often ignored in doctoral programs and that only a minority of students know how to apply the methods of reliability correctly (Aiken, West, & Millsap, 2008). Of course, *v*-ignorance can sometimes reflect motivated ignorance. For example, current modal practices do not involve the assessment of measurement invariance. Choosing to test for invariance can greatly decrease one's chances of publication (e.g., measurement issues can undermine theoretical conclusions), and therefore there is little incentive to do so. Both *v*-hacking and v-ignorance can lead to an overinflation of the true structural validity of a measure and thus undermine the validity of research findings.

There are several ways to address and immunize research against these practices. One is for journals, editors, and reviewers to require the psychometric evaluation of all measures used, much as effect sizes, confidence intervals, and precise p values are now commonly required (Parsons, Kruijt, & Fox, 2019). A second is for psychological scientists to come together and discuss issues such as choice of metrics, implementations, and cutoffs, as well as other experimenter degrees of freedom. Let us be clear here: We are not advocating for the introduction of some set of universally applied metrics or cutoff values for those metrics. Such an approach may lead researchers to mindlessly employ such standards and would raise a host of well-known issues (e.g., those associated with using null-hypothesis significance testing and treating $p < .05$ or a Bayes factor ≥ 3 as a sacrosanct threshold; for related arguments see Simmons, Nelson, & Simonsohn, 2018). Rather, we hope that readers will recognize that massive heterogeneity in the choice of cutoffs and metrics serves to inflate research degrees of freedom, and therefore threatens confidence in measurement. If the ongoing debate around p values is any indication (e.g., Benjamin et al., 2018; Lakens et

al., 2018), the effort required to address this issue is unlikely to be trivial, and change may take some time.

However, there is no reason to be pessimistic. Researcher degrees of freedom could be greatly constrained by expanding the practice of preregistration to also include choices concerning the assessment of structural validity (e.g., metrics, cutoffs, measurement models, and decision-making strategies). Preregistration of design and analytic strategy prior to data collection greatly increases confidence in the conclusions of hypothesis-testing research (Nosek, Ebersole, DeHaven, & Mellor, 2018). We expect that preregistration of measurement choices would yield comparable benefits. Finally, providing open access to data also allows future researchers to examine the structural validity of a measure using metrics not originally reported, and enables data to be pooled across studies for reuse and meta-analytic validation. Although ethical considerations are sometimes cited as a barrier to data sharing, innovations such as synthetic data sets (e.g., using the *synthpop* R package; Nowok, Raab, Snoke, & Dibben, 2019) allow researchers to create and share data sets with statistical properties (e.g., covariance matrices, means, and distributions) highly similar to those of original data sets without including any of the original data (see Quintana, 2019, for an accessible primer).

Conclusion

This article provides a psychometrically rich assessment of the structural validity of 15 commonly used questionnaires. These analyses are useful for readers who (a) are interested in a large-scale examination of the structural validity of measures used in social and personality psychology, (b) wish to know more about normative distributions and psychometric properties of several well-known self-report questionnaires (e.g., for deciding whether to employ a measure in a future study or for compare results with those found in large samples elsewhere), (c) want confidence that measures developed offline have good structural validity when used online, or (d) plan to use the AIID data set for other purposes and need information about the structural validity of the scales therein. Perhaps the most important contribution of our findings is that they suggest that the documented underreporting of structural-validity metrics in social and personality psychology presents an even more worrying issue of hidden invalidity among commonly used measures. We have offered recommendations on how this issue might be addressed (e.g., with preregistration of the plan for assessing structural validity and of how validity assessments could impact substantive conclusions). Researchers are currently afforded a large number of degrees of freedom, and validity-related decisions can be hidden or made post hoc. This can lead to situations in which there are few, if any, constraints that prevent researchers from cherry-picking those validity metrics that provide the most favorable impression of their measures (*v*-hacking), to the potential detriment of the validity of their conclusions.

Notes

Author Contributions

Both authors designed the study. I. Hussey wrote the analysis code and analyzed the data. Both authors wrote the manuscript and approved its final submitted version.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was conducted with the support of Ghent University Grant 01P05517 to I. Hussey and FWO Grant BOF16/MET_V/002 to Jan De Houwer.

Open Practices

All data and code have been made publicly available via the Open Science Framework and can be accessed at osf.io/23rzk.

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. *The American Psychologist, 63,* 32–50. doi:10.1037/0003-066X.63.1.32

Altemeyer, B. (1981). *Right-wing authoritarianism.* Winnipeg, Manitoba, Canada: University of Manitoba Press.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

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

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., . . . Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2,* 6–10. doi:10.1038/s41562-017-0189-z

Bohannon, J. (2016). Mechanical Turk upends social sciences. *Science, 352,* 1263–1264. doi:10.1126/science.352.6291.1263

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071. doi:10.1037/0033-295X.111.4.1061

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48,* 306–307. doi:10.1207/s15327752jpa4803_13

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14,* 464–504. doi:10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 9,* 233–255. doi:10.1207/S15328007SEM0902_5

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

Dalbert, C., Lipkus, I. M., Sallay, H., & Goch, I. (2001). A just and an unjust world: Structure and validity of different world beliefs. *Personality and Individual Differences, 30,* 561–577. doi:10.1016/S0191-8869(00)00055-6

De Schryver, M., Hughes, S., De Houwer, J., & Rosseel, Y. (2018). On the reliability of implicit measures: Current practices and novel perspectives. PsyArXiv. doi:10.31234/osf.io/w7j86

Flake, J. K., & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. PsyArXiv. doi:10.31234/osf.io/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. doi:10.1177/1948550617693063

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Retrieved from osf.io/n3axs/

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology, 66,* 877–902. doi:10.1146/annurev-psych-010814-015321

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74,* 1464–1480. doi:10.1037/0022-3514.74.6.1464

Henry, L., & Wickham, H. (2019). purrr: Functional programming tools (R package Version 0.3.3) [Computer software]. Retrieved from https://CRAN.R-project.org/package=purrr

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 1–55. doi:10.1080/10705519909540118

Hussey, I., Hughes, S., Lai, C. K., Ebersole, C. R., Axt, J., & Nosek, B. A. (2019). The Attitudes, Identities, and Individual Differences (AIID) Study and Dataset. Retrieved from https://osf.io/pcjwf/

John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. Pervin & O. John (Eds.), Handbook of personality: Theory and research (2nd ed., Vol. 2, pp. 102–138). New York, NY: Guilford Press.

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., Rosseel, Y., Miller, P., Quick, C., . . . Mansolf, M. (2019). semTools: Useful tools for structural equation modeling (R package Version 0.5-2) [Computer software]. Retrieved from https://CRAN.R-project.org/package=semTools

Katz, I., & Hass, R. G. (1988). Racial ambivalence and American value conflict: Correlational and priming studies of dual cognitive structures. *Journal of Personality and Social Psychology, 55,* 893–905. doi:10.1037/0022-3514.55.6.893

Kuhn, M., Chow, F., & Wickham, H. (2019). rsample: General resampling infrastructure (R package Version 0.0.5) [Computer software]. Retrieved from https://CRAN.R-project.org/package=rsample

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., . . . Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour, 2,* 168–171. doi:10.1038/s41562-018-0311-x

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15,* 371–379. doi:10.1037/a0025172

Li, C. H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychological Methods, 21,* 369–387. doi:10.1037/met0000093

Little, T. (2013). Longitudinal structural equation modeling. New York, NY: Guilford Press.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3,* 635–694. doi:10.2466/pr0.1957.3.3.635

McDonald, R. P. (1999). Test theory: A unified treatment. Mahwah, NJ: Erlbaum.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93,* 568–592. doi:10.1037/0021-9010.93.3.568

Mullen, S. P., Gothe, N. P., & McAuley, E. (2013). Evaluation of the factor structure of the Rosenberg Self-Esteem Scale in older adults. *Personality and Individual Differences, 54,* 153–157. doi:10.1016/j.paid.2012.08.009

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1,* Article 0021. doi:10.1038/s41562-016-0021

Neuberg, S. L., & Newsom, J. T. (1993). Personal need for structure: Individual differences in the desire for simpler structure. *Journal of Personality and Social Psychology, 65,* 113–131. doi:10.1037/0022-3514.65.1.113

Nosek, B. A. (2012). Intuitions about controllability of feelings, thoughts, and behaviors. Retrieved from osf.io/puwyq

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, USA, 115,* 2600–2606. doi:10.1073/pnas.1708274114

Nowok, B., Raab, G. M., Snoke, J., & Dibben, C. (2019). synthpop: Generating synthetic versions of sensitive microdata for statistical disclosure control (R package Version 1.5-1) [Computer software]. Retrieved from https://CRAN.R-project.org/package=synthpop

Nunnally, J., & Bernstein, I. (1994). Psychometric theory (3rd ed.). New York, NY: McGraw-Hill.

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology, 96,* 966–980. doi:10.1037/a0022955

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* Article aac4716. doi:10.1126/science.aac4716

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. doi:10.1177/2515245919879695

Paulhus, D. (1983). Sphere-specific measures of perceived control. *Journal of Personality and Social Psychology, 44,* 1253–1265. doi:10.1037/0022-3514.44.6.1253

Paulhus, D. (1988). Assessing self-deception and impression management in self-reports: The Balanced Inventory of Desirable Responding. Unpublished manuscript, Department of Psychology, University of British Columbia, Vancouver, Canada.

Pratto, F., Sidanius, J., Stallworth, L. M., & Malle, B. F. (1994). Social dominance orientation: A personality variable predicting social and political attitudes. *Journal of Personality and Social Psychology, 67,* 741–763. doi:10.1037/0022-3514.67.4.741

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41,* 71–90. doi:10.1016/j.dr.2016.06.004

Quintana, D. (2019). Synthetic datasets: A non-technical primer for the behavioural sciences to promote reproducibility and hypothesis-generation. PsyArXiv. doi:10.31234/osf.io/dmfb3

Revelle, W., & Condon, D. (2018). Reliability from α to ω: A tutorial. PsyArXiv. doi:10.31234/osf.io/2y3w9

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.). (1991). Measures of personality and social psychological attitudes. San Diego, CA: Academic Press.

Rose, N., Wagner, W., Mayer, A., & Nagengast, B. (2019). Model-based manifest and latent composite scores in structural equation models. *Collabra: Psychology, 5,* Article 9. doi:10.1525/collabra.143

Rosenberg, M. (1965). Society and the adolescent self-image. Princeton, NJ: Princeton University Press.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2). doi:10.18637/jss.v048.i02

Salerno, L., Ingoglia, S., & Lo Coco, G. (2017). Competing factor structures of the Rosenberg Self-Esteem Scale (RSES) and its measurement invariance across clinical and non-clinical samples. *Personality and Individual Differences, 113,* 13–19. doi:10.1016/j.paid.2017.02.063

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science, 2,* 107–114. doi:10.1177/2515245919838781

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8,* 350–353. doi:10.1037/1040-3590.8.4.350

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,* 107–120. doi:10.1007/s11336-008-9101-0

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. doi:10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-positive citations. *Perspectives on Psychological Science, 13,* 255–259. doi:10.1177/1745691617698146

Snyder, M. (1987). Public appearances, private realities: The psychology of self-monitoring. New York, NY: W. H. Freeman/Times Books/Henry Holt.

Stöber, J., Dette, D. E., & Musch, J. (2002). Comparing continuous and dichotomous scoring of the Balanced Inventory of Desirable Responding. *Journal of Personality Assessment, 78,* 370–389. doi:10.1207/S15327752JPA7802_10

Supple, A. J., Su, J., Plunkett, S. W., Peterson, G. W., & Bush, K. R. (2013). Factor structure of the Rosenberg Self-Esteem Scale. *Journal of Cross-Cultural Psychology, 44,* 748–764. doi:10.1177/0022022112468942

Tay, L., & Jebb, A. T. (2018). Establishing construct continua in construct validation: The process of continuum specification. *Advances in Methods and Practices in Psychological Science, 1,* 375–388. doi:10.1177/2515245918775707

Tomas, J. M., & Oliver, A. (1999). Rosenberg's self-esteem scale: Two factors or method effects. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 84–98. doi:10.1080/10705519909540120

Uhlmann, E. L., Brescoll, V., & Machery, E. (2010). The motives underlying stereotype-based discrimination against members of stigmatized groups. *Social Justice Research, 23,* 1–16. doi:10.1007/s11211-010-0110-7

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4–70. doi:10.1177/109442810031002

Watson, D. (2004). Stability versus change, dependability versus error: Issues in the assessment of personality over time. *Journal of Research in Personality, 38,* 319–350. doi:10.1016/j.jrp.2004.03.001

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67,* 1049–1062. doi:10.1037/0022-3514.67.6.1049