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

Ian Hussey & Sean Hughes
Ghent University

Flake, Pek, and Hehman (2017) recently demonstrated that metrics of structural validity are severely underreported in social and personality psychology. We apply their recommendations for the comprehensive assessment of structural validity to a uniquely large and varied dataset (N = 144496 experimental sessions) to investigate the psychometric properties of some of the most widely used self-report measures (k = 15 questionnaires, 26 subscales) in social and personality psychology. When assessed using the modal practice of considering only their internal consistency, 89% of scales 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 median age and gender) only 4% demonstrated good validity. Furthermore, the less commonly a test is reported in the literature, the more likely it was to be failed (e.g., measurement invariance). This suggests that the pattern of under-reporting in the field may represent widespread hidden invalidity of the measures we use, and 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. Similar to the better-known concept of p-hacking, we introduce the concept of validity hacking (v-hacking) and argue that it should be acknowledged and addressed.

Our 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 has 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 which have contributed to the so-called 'replication crisis' (LeBel & Peters, 2011; Simmons, Nelson, & Simonsohn, 2011). Numerous research and analytic practices have now been subject to questioning, from an over-reliance on null hypothesis significance testing, to the need for increased transparency and sharing of data, pre-registrations, and replications (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 we can have in our 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 one's 'belief in a just world', right-wing authoritarianism, or to quantify personality traits.[1] Designing valid measures of latent constructs requires that the measures themselves are subject to an ongoing process known as construct validation (where measures could be self-report scales, implicit measures, or otherwise: see De Schryver, Hughes, De Houwer, Hussey, & Rosseel, 2019; De Schryver et al., 2018; see Borsboom, Mellenbergh, & van Heerden, 2004; Cronbach & Meehl, 1955 regarding construct validation). As Flake et al. (2017) point out, 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 (*substantive*) involves identifying and defining a construct (via literature review and construct conceptualization), determining how it will be assessed (via item development and selection), and ensuring that the resulting scale content is both relevant and representative. The second (*structural*) phase develops a theory about the construct's structure. Quantitative analyses are used to determine the psychometric properties of the measure (e.g., by engaging in item and factor analyses, assessing the measure's consistency or stability, and checking for measurement invariance). The third (*external*) phase examines 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 known groups comparisons (for a more detailed overview see Cronbach & Meehl, 1955; Loevinger, 1957; and the Standards of Educational and Psychological Testing: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Much of the theoretical work in social and personality psychology centers on the identification and definition of constructs (first phase) while empirical work tends to assess whether these constructs predict, discriminate between, or converge with other measures (third phase). Yet 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 papers 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 that there is an over-reliance on Cronbach's α as the

[1]In-line with Flake et al. (2017) we define a scale as a measure which relies on items to represent a latent construct.

sole source of structural validity evidence, and argue that rigorous methodologies for measurement are rarely reported. Indeed, Flake et al. (2017) found that the problem with validation was actually more severe that it initially appeared. Specifically, they not only found that research with well-known measures over-relied on Cronbach's α as the sole test of structural validity, but that 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, given that (c) the three phases of validation are intertwined. Put simply, conclusions about the construct stemming from the third (*external*) phase may not hold if issues exist at the first (*substantive*) phase (e.g., the construct lacks a strong theoretical basis) or the second (*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. This involves moving beyond the simple modal practice of assessing for internal consistency (Cronbach's α) to investigating the stability of scores across time (test-retest reliability), examining the factor structure of the latent construct(s) (Confirmatory Factor Analysis), and testing for the equivalence of measurement properties across populations, time points, and contexts (measurement invariance: Putnick & Bornstein, 2016; Vandenberg & Lance, 2000). Although analyses 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 the same measure reflects the same construct across samples, contexts, and conditions (see the Method section for a more detailed treatment of different types of structural validity assessment).

Purpose of the 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 under-report tests of validity leaves the field in a sticky situation: it is currently impossible to know whether we are facing a mere problem of under-reporting (as Flake et al. highlighted) or the potentially deeper issue of hidden invalidity. It may be that many of the measures we use appear 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 examined the structural validity of fifteen well-known self-report measures that are often used in social and personality psychology using several best practices (see Table 1). This provided a unique case study in what can be achieved when a wide number of validity metrics are applied using best practices to a large number of measures, each tested in a large sample. To achieve this, we used the 'Attitudes, Identities, and Individual Differences' (AIID) dataset, a large-scale, multivariate, planned-missing data study that was collected via the Project Implicit website (implicit.harvard.edu) between 2004 and 2007 (Hussey, Hughes, & Nosek, 2019; see osf.io/pcjwf).

Utilizing this dataset provides 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 these measures with numbers that were far greater than those used in many earlier validation studies. Second, the dataset's structure allowed us to apply a large range of structural validity metrics to the same measure in the same study, include tests of stability (test-retest reliability) based on multiple delay ranges (immediate vs. up to 1 year later). Third, we adopted a comprehensive strategy to structural validity testing that extends beyond previous studies in both its nuance and scope. In line with best practices, we obtained measures of consistency (Cronbach's α, McDonalds ω_t and ω_h), test-retest reliability (both dependability & stability: Revelle & Condon, 2018), factor structure (Confirmatory Factor Analysis), and measurement invariance. Although some of these tests have been applied to some of these scales, this was often done separately across papers and between samples, never comprehensively within and across a range of measures as we do here. 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 where many self-report scales are being used in contexts, and with samples, that differ to those in which they were originally validated. If we wish to use these measures in online settings, it is imperative that we examine their structural validity in this context to ensure that their psychometric properties are adequate and do not diverge from those observed in traditional (laboratory) settings.

It is worth noting that the sequential ordering of the tests we carried out, as reported in text and in Tables 1 and 2, was according to the frequency with which they 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) makes numerous assumptions that can only be assessed by less commonly applied analyses (e.g., within a CFA context).

Finally, and before we continue, let us be clear: our goal was not to make a final or absolute determination on the (in)validity of any of the scales we assessed; to make a binary determination of their (in)validity; 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 test the issue highlighted by Flake and colleagues (2017) - namely - whether the widespread under-reporting of structural validity information reflects hidden validity or, more worryingly, hidden invalidity.

Table 1. *Summary of 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	Good	Good	Poor	Poor	Questionable
Bayesian Racism	Good	Good	Good	Poor	Questionable
Belief in a Just World: General Just World Scale	Good	Good	Good	Poor	Questionable
Big 5 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	Good	Good	Good	Poor	Questionable
Intuitions about Controllability and Awareness of Thoughts					
Others	Good	Good	Poor	Poor	Questionable
Self	Good	Good	Poor	Poor	Questionable
Need for Cognition	Good	Good	Good	Good	Good
Need for Cognitive Closure					
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	Good	Good	Mixed	Poor	Questionable
Protestant Ethic	Good	Good	Mixed	Poor	Questionable
Ring-Wing Authoritarianism	Good	Good	Mixed	Poor	Questionable
Rosenberg Self-Esteem	Good	Good	Good	Poor	Questionable
Self-Monitoring	Good	Good	Poor	Poor	Questionable
Social Dominance Orientation	Good	Good	Poor	Poor	Questionable
Spheres of Control					
Interpersonal Control	Good	Good	Good	Poor	Questionable
Personal Efficacy	Poor	Good	Poor	Poor	Questionable
Summary	89%	100%	73%	4%	4%

Notes: Good internal consistency refers to McDonald's (1999) $\omega_t \geq 0.7$; good dependability refers to 1-hour test-retest $r \geq 0.7$; good stability refers to test-retest with follow up between 1 day and 1 year $r \geq 0.7$; good confirmatory model fit refers to meeting all of CFI $\geq .95$, TLI $\geq .95$, RMSEA $\leq .06$, and SRMR $\leq .09$, mixed confirmatory model fit refers to meeting SRMR $\leq .09$ and any one of CFI $\geq .95$, TLI $\geq .95$, or RMSEA $\leq .06$; see Hu & Bentler, 1999), and poor fit refers to meeting neither of these; good measurement invariance refers to meeting configural invariance (using same criteria as mixed CFA fit), metric invariance and scalar invariance (for each, meeting both ΔCFI $\geq$ -.15 and ΔRMSEA $\leq .01$; see Chen, 2007) for both median age and gender; good global structural validity refers to having no poor fits on any of these metrics.

Method

Disclosures

Preregistration. Our analyses were not preregistered.

Data, materials, and online resources. All code and data to reproduce our analyses is available at osf.io/23rzk. Supplementary materials, including additional results not reported in the manuscript, simplified R scripts for educational purposes, and a changelog documenting differences between manuscript versions, are available 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 dataset was approved by the University of Virginia's Institutional Review Board for the Social and Behavioral Sciences (protocol 2003-0173-00). As this data was collected between 2005 and 2007, this study is technically not in accordance with the most recent version of the Declaration of Helsinki (2013), which requires pre-registration prior to data collection. Ethical approval was not required for our analysis of this existing data.

Participants

The data of 144,496 experimental sessions representing 81,986 unique participants (50,141 women and 31,845 men, $M_{age} = 30.84$, $SD = 11.40$) were selected for inclusion from the AIID dataset on the basis that they met our predefined study criteria (i.e., age 18-65, self-reported fluent English, and complete data on the individual differences measure and demographics items). Sample size for each measure can be found in Table 2. Repeat participation in the study was possible and allowed for the assessment of test-retest reliability. The modal number of participations was one ($M = 1.76$, $SD = 2.22$).

Measures

Fifteen questionnaires were selected for inclusion in this study based on availability in the AIID dataset. Five of these with a larger number of items were subdivided into two parts and delivered between participants due to time constraints on the Project Implicit site (as noted below). This resulted in participants being assigned to one of 20 different questionnaires: the Balanced Inventory of Desirable Responding (Version 6: Paulhus, 1988; cited in Robinson, Shaver, & Wrightsman, 1991, split into Impression Management and Self Deception subscales), Bayesian Racism Scale (Uhlmann, 2002; Uhlmann, Brescoll, & Machery, 2010), Belief in a Just World (General Just World subscale: Dalbert, Lipkus, Sallay, & Goch, 2001), Big Five Inventory (John & Srivastava, 1999; split into extraversion, conscientiousness & neuroticism vs. agreeableness & openness subscales), Humanitarianism-Egalitarianism Scale (Katz & Hass, 1988), Intuitions About Controllability and Awareness of Thoughts for Others

(Nosek, 2002; split into self and others subscales), Need for Cognition (Cacioppo, Petty, & Kao, 1984). Need for Cognitive Closure (Webster & Kruglanski, 1994; split into order & ambiguity vs. predictability, decisiveness, & closed-mindedness subscales), 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 number 4: Pratto, Sidanius, Stallworth, & Malle, 1994), and Spheres of Control Battery (Paulhus, 1983; split into interpersonal control vs. personal efficacy subscales).

Fourteen of these questionnaires have previously been employed in a published article or book chapter, whereas one had not (i.e., it was author created: the Intuitions about Controllability and Accessibility of Thoughts scales). In cases where a measure was previously published, its psychometric properties 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: Uhlmann et al., 2010). Overall, the questionnaires employed between 6 and 44 items ($M = 19.5$, $SD = 11.8$), and between 1 and 5 subscales ($M = 1.7$, $SD = 1.4$). All scales employed the same response format, a Likert scale ranging from 1 (*strongly disagree*) to 6 (*strongly agree*), which, in some cases, differed from the measure's original response format. Note that in cases where more significant modifications were made (i.e., from a dichotomous to Likert response format), this was carried out based on the recommendations of research elsewhere in the literature (Dalbert et al., 2001; Stöber, Dette, & Musch, 2002). A minority of items in several measures was also subject to wording adjustments to make them more appropriate for a general rather than student sample (see Supplementary Materials).

Procedure

In what follows we provide a brief overview of the AIID study (for a more detailed description see Hussey et al., 2019). Prior to the study participants navigated to the Project Implicit research website on their own accord, created a unique login 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 a battery which asked about the same attitude domain as probed in the IAT. Both IAT and self-report measures centered on one of 95 attitude domains. Each domain consisted of two concept categories that were related to either 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-reported ratings, participants were

randomly assigned to complete one of the twenty individual difference self-report measures.

In the current study we only made use of data from the demographics questionnaire (age, gender, and English fluency) and individual difference measures. Given that people only completed 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 apart). In order to maintain a consistent analytic strategy, we analyzed each questionnaire's subscales separately. This is consistent with past use of these questionnaires' in the almost all cases. In what follows, we refer to these subscales as 'scales' for convenience.

Results

Data Preparation

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

Analytic Strategy

For each scale, we calculated both distribution information and multiple metrics of structural validity following the recommendations of Flake et al. (2017) and Revelle and Condon (2018; see Table 2). Distribution information (mean, standard deviation, skewness, & 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., 2018). Confidence intervals were bootstrapped via the case removal and quantile method using 1000 resamples, and were implemented using the R packages rsample (Kuhn & Wickham, 2019) and purrr (Henry & Wickham, 2019).

For all scales, simple measurement models were employed which 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 (in)validity across scales, in line with our primary research question. Second, with few exceptions (e.g., the BFI), most scale's 'true' measurement model is either 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). As such, the choice to employ alternative models would represent exploratory or weakly informed model choices, comparisons among which would detract from 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). By adopting similar assumptions here, our findings better reflect how these scales are commonly used and interpreted.

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

Consistency. Given that Cronbach's α is frequently argued to be misused and of limited utility (Flake et al., 2017; Schmitt, 1996; Sijtsma, 2009), we also provide two less frequently reported but arguably superior metrics of internal consistency: McDonald's ω_t (Omega total) and ω_h (Omega hierarchical; McDonald, 1999). ω_t provides a metric of total measure reliability, or the proportion of variance that is attributable to sources other than measurement error. 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 a cutoff value of $\omega_t \geq 0.7$ on the basis that 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 that subset of participants with available data ($n = 7542$) using Pearson's r correlations. We calculated two forms of test-retest reliability based on the recommendations of Revelle and Condon (2018). First, test-retest 'dependability' was calculated using those participants who completed a scale twice within one hour. Second, test-retest 'stability' was calculated using those participants who completed a scale twice with a period of between one day and one year between the two sessions. We employed a cutoff value of $r \geq 0.7$ for both test-retest dependability and stability based on common recommendations in the literature (Nunnally & Bernstein, 1994).

Factor structure. Due to the large number of scales, we employed a standardized approach to specifying and assessing the fit of measurement models based on recommended best practices (Hu & Bentler, 1999; Rose et al., 2019). First, confirmatory factor structure models for each scale were defined using the items specified in a scale's original publication. For example, if a questionnaire was

constructed to contain two (sub)scales (e.g., the Spheres of Control scale), separate CFA models were specified for each scale with the appropriate items loading onto a latent variable. No methods factors (e.g., for negatively worded items) were included.

Based on the use of ordinal Likert response formats across all scales, and differential skew between the sum scores of different scales, we employed the diagonally weighted least squares (DWLS) estimator along with a robust standard errors of parameter estimates (i.e., the WLSMV estimator option within lavaan). This estimator function has been shown in simulation studies to be superior to the more common Maximum Likelihood method (Li, 2016). Refitting the models using the Maximum Likelihood method, with or without robust standard errors, produced poorer performance across all scales.

Previous work has repeatedly suggested that multiple model goodness-of-fit indices 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: measures of absolute fit: Chi square tests (although, given our sample sizes the p values for these are universally significant and therefore uninformative; nonetheless Chi square values should be reported), Chi square normalized by number of items, the Root Mean Square of the Residual (RMSR); measures of relative fit: the Tucker Lewis Fit Index (TLI); and noncentrality indices: Comparative Fit Index (CFI), and Root Mean Square Error of Approximation (RMSEA and its 95% confidence intervals). For decision-making purposes regarding model fit, we employed the cutoffs suggested by Hu and Bentler (1999: i.e., CFI $\geq$ 0.95, TLI $\geq$ 0.95, RMSEA $\leq$ 0.06, SRMR $\leq$ 0.09). Hu and Bentler argue that model fit decisions on the basis of two fit indices lower the combined rate of Type I and Type II errors relative to methods based on a single index. Specifically, they recommend 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 observed that individual scales could be said to demonstrate good or poor fit based on which of these three indices was chosen in combination with SRMR. As such, if a scale demonstrated good fit when considering all three metric permutations (i.e., SRMR+CFI, SRMR+TLI, and SRMR+RMSEA) we labeled it as being 'good'; if it demonstrated good fit using one or two but not all three permutations it was labeled as 'mixed'; and if it demonstrated good fit using none of the three permutations it was labeled as 'poor'.[2]

[2]An alternative strategy of employing all four metrics for decision-making was considered but ultimately rejected due to the fact that: (a) there was no basis for this analytic strategy in the literature, thus preventing us from making informed choices about cutoff values when using four indices, and (b) the 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 its impact on false negative rates.

Measurement invariance. A scale's capacity to measure the same construct in a comparable way between populations or contexts typically involves three component tests: (1) configural invariance (i.e., equivalence of model form: does the unconstrained model provide adequate fit in each of the groups), (2) metric invariance (or weak factorial invariance; i.e., equivalence of factor loadings), and (3) scalar invariance (or strong factorial invariance; i.e., equivalence of item intercepts or thresholds; Putnick & Bornstein, 2016). These are typically assessed as nested models, whereby 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 between the groups. That is, individuals' observed scores on the scale are likely to measure the same latent variable and in a comparable way between the groups. Loosely speaking, one accessible interpretation of meeting measurement invariance is that individuals in both subgroups 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 the subgroups, nor associations between scores on the scale and external variables.

Although tests of measurement invariance are typically performed between groups that the researcher wants to directly compare, one should also assess measurement invariance between groups that one tacitly assumes should be invariant. For example, many studies recruit adults (e.g., age 18 to 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 researcher implicitly assumes that the scales measure the same construct(s) across both groups. It is therefore useful to test these two assumptions, specifically that the employed scales are invariant across gender (female vs. male) and median age (age $\geq$ 27 in our data). Equally, if a study explicitly wishes to make comparisons based on these categories (e.g., between men and women), measurement invariance would still be a requirement for these comparisons to be meaningful. For example, differences between men and women in personality are theoretically meaningful only if they represent differences in latent means rather than factor loadings or intercepts. In all cases, measurement invariance is therefore necessary to subsequent substantive analyses.

Historically, the most common method 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 here due to the sensitivity of

Chi square tests to our large sample sizes. In addition, relying exclusively on the significance of Chi square tests has fallen out of favor over time in place of alternative fit indices such as RMSEA (Putnick & Bornstein, 2016). Numerous simulation studies have been conducted on which indices to employ and cutoffs to use (if any). Recommendations range from liberal (Cheung & Rensvold, 2002) to conservative (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 for 'mixed' CFA fit employed above (Hu & Bentler, 1999) combined with Chen's (2007) moderate criteria of meeting both ΔCFI $>$ -.01 & ΔRMSEA $<$ .015 for each of metric and scalar invariance. 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.

A summary of the results of these metrics of structural validity using recommended cutoff values can be found in Table 1. This table provides a concise summary of the structural validity evidence for each individual scale as well as general conclusions about structural validity across scales using best-practice recommendations. Table 2 provides the results of all tests and metrics (i.e., internal consistency, test-retest reliability, factor structure, and measurement invariance for median age and gender), along with details of each scale (n participants, k items, number of assumed factors), and distributional information (M, SD, skewness, kurtosis). When combined, Tables 1-2 provide a wide range of psychometric properties for fifteen commonly used self-report individual differences scales that could inform their future use. Full results of the tests of measurement invariance (i.e., each fit index for each test) can be found in the Supplementary Materials. Additionally, recent research has also sought to quantify the impact of failure to meet measurement invariance as a continuous variable (e.g., Nye & Drasgow, 2011). While this is beyond the scope of this article, estimates of differences in between groups effect sizes between the latent and observed sum scores can be found in the Supplementary Materials.

Results synthesis

In Table 1, we use categorical terms such as 'Good', 'Questionable', or 'Poor' to summarize conclusions about the structural validity of scales based on the cutoff values discussed above for each dimension of validity. These labels serve to condense multifaceted metrics of validity to categorical conclusions in order to enable decision-making with regard to our core research question (i.e., whether under-reporting represents hidden validity or invalidity). This tradeoff 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 non-significant dichotomy to facilitate conclusions

regarding hypotheses. These categorical labels should not be taken as literally ‘true’ for any other research question than our own (e.g., when assessing the adequacy of a scale for future use). Instead, such questions should be informed by the continuous and multifaceted results reported in Table 2, which offer a more nuanced perspective on structural validity.

Table 2. *Results of structural validity analyses.*

			Internal consistency								
				95% CI			95% CI			95% CI	
Scale	Total n	Items	α	Lower	Upper	ω_t	Lower	Upper	ω_h	Lower	Upper
Balanced Inventory of Desirable Responding											
Impression Management	6934	18	.797	.789	.804	.798	.791	.805	.796	.789	.803
Self Deception	6713	18	.703	.692	.714	.708	.697	.720	.707	.697	.719
Bayesian Racism	6532	16	.824	.818	.831	.828	.822	.835	.822	.815	.829
Belief in a Just World: General Just World Scale	6758	6	.754	.744	.765	.760	.751	.771	.761	.751	.772
Big 5 Inventory											
Agreeableness	6713	9	.792	.784	.800	.793	.784	.800	.788	.779	.796
Conscientiousness	6649	9	.820	.812	.827	.820	.812	.826	.810	.802	.817
Extraversion	6649	8	.869	.864	.874	.872	.867	.877	.869	.865	.874
Neuroticism	6649	8	.832	.826	.839	.834	.828	.840	.832	.826	.838
Openness	6713	10	.793	.785	.801	.792	.783	.801	.784	.774	.793
Humanitarianism-Egalitarianism	6906	10	.840	.831	.847	.839	.831	.847	.830	.820	.839
Intuitions about Controllability and Awareness of Thoughts											
Others	6711	9	.750	.740	.761	.753	.743	.763	.744	.733	.754
Self	6830	9	.797	.789	.804	.800	.792	.807	.801	.793	.808
Need for Cognition	6649	18	.889	.885	.893	.889	.885	.893	.885	.880	.889
Need for Cognitive Closure											
Ambiguity	6585	9	.674	.661	.686	.683	.670	.695	.680	.667	.693
Closed mindedness	6559	8	.641	.627	.655	.638	.622	.652	.631	.615	.646
Decisiveness	6559	7	.816	.809	.823	.824	.817	.831	.825	.818	.832
Order	6585	10	.819	.811	.826	.825	.818	.832	.824	.817	.831
Predictability	6559	8	.793	.784	.801	.796	.787	.804	.795	.786	.803
Personal Need for Structure	6821	12	.861	.855	.865	.862	.857	.866	.860	.854	.864
Protestant Ethic	6859	11	.791	.783	.798	.791	.783	.798	.782	.773	.789
Ring-Wing Authoritarianism	6542	20	.922	.919	.924	.922	.919	.925	.910	.907	.914
Rosenberg Self-Esteem	6971	10	.890	.886	.895	.896	.892	.900	.887	.882	.892
Self-Monitoring	6623	18	.759	.750	.768	.760	.749	.770	.740	.723	.755
Social Dominance Orientation	6854	12	.831	.824	.837	.831	.824	.837	.821	.814	.828
Spheres of Control											
Interpersonal Control	6785	10	.808	.801	.816	.810	.803	.818	.808	.800	.816
Personal Efficacy	6899	10	.641	.627	.654	.638	.623	.651	.623	.607	.637

Notes: Total n refers to the total number of participants with data available for internal consistency, distribution, confirmatory factor structure and measurement invariance analyses.

Table 2 (continued)

	Distribution				Test-retest dependability				Test-retest stability			
							95% CI				95% CI	
Scale	*M*	*SD*	Skewness	Kurtosis	*n*	*r*	Lower	Upper	*n*	*r*	Lower	Upper
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	63.38	10.25	0.06	3.22	173	.89	.83	.93	132	.76	.66	.83
Bayesian Racism	41.09	13.16	0.40	2.91	136	.84	.74	.90	154	.88	.82	.92
Belief in a Just World: General Just World Scale	0.07	5.64	0.07	2.67	170	.88	.82	.92	150	.74	.65	.81
Big 5 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	50.69	7.02	-1.20	5.14	131	.91	.84	.95	149	.85	.75	.90
Intuitions about Controllability and Awareness of Thoughts												
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	80.91	13.8	-0.44	2.98	147	.85	.75	.91	133	.86	.80	.90
Need for Cognitive Closure												
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	42.10	10.15	-0.06	2.82	149	.88	.82	.92	141	.81	.71	.88
Protestant Ethic	41.06	8.69	-0.18	3.05	163	.92	.88	.94	158	.85	.80	.89
Ring-Wing Authoritarianism	51.30	18.84	0.48	2.45	116	.96	.92	.98	163	.94	.91	.96
Rosenberg Self-Esteem	46.91	9.55	-0.80	3.35	160	.95	.92	.96	156	.90	.86	.93
Self-Monitoring	63.18	12.20	0.03	2.85	140	.91	.86	.94	157	.87	.82	.91
Social Dominance Orientation	25.58	9.85	0.71	2.98	161	.91	.86	.94	149	.84	.78	.88
Spheres of Control												
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

Notes: *r* refers to Pearson's *r* correlations between time points, dependability refers to test-retest reliability within 1 hour, stability refers to test-retest within between one day and one year.

Table 2 (continued)

	Confirmatory Factor Analysis								
							95% CI		
Scale	χ^2	χ^2/df	df	CFI	TLI	RMSEA	Lower	Upper	SRMR
Balanced Inventory of Desirable Responding									
Impression Management	1834	13.6	135	0.950	0.944	0.043	0.041	0.044	0.040
Self Deception	3743	27.7	135	0.819	0.795	0.063	0.061	0.065	0.059
Bayesian Racism	1667	16.0	104	0.965	0.960	0.048	0.046	0.050	0.046
Belief in a Just World: General Just World Scale	153	17.0	9	0.986	0.977	0.049	0.042	0.056	0.031
Big 5 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	1152	32.9	35	0.948	0.933	0.069	0.066	0.072	0.060
Humanitarianism-Egalitarianism	485	13.9	35	0.979	0.972	0.043	0.040	0.047	0.048
Intuitions about Controllability and Awareness of Thoughts									
Others	1378	51.0	27	0.903	0.870	0.086	0.083	0.090	0.073
Self	1509	55.9	27	0.929	0.906	0.090	0.086	0.094	0.072
Need for Cognition	1959	14.5	135	0.978	0.975	0.045	0.043	0.047	0.044
Need for Cognitive Closure									
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	1472	27.3	54	0.969	0.962	0.062	0.059	0.065	0.055
Protestant Ethic	1244	28.3	44	0.951	0.939	0.063	0.06	0.066	0.056
Ring-Wing Authoritarianism	6647	39.1	170	0.959	0.954	0.076	0.075	0.078	0.076
Rosenberg Self-Esteem	875	25.0	35	0.982	0.977	0.059	0.055	0.062	0.057
Self-Monitoring	11631	86.2	135	0.689	0.648	0.113	0.112	0.115	0.102
Social Dominance Orientation	1785	33.1	54	0.946	0.934	0.068	0.066	0.071	0.064
Spheres of Control									
Interpersonal Control	829	23.7	35	0.965	0.955	0.058	0.054	0.061	0.049
Personal Efficacy	1490	42.6	35	0.836	0.790	0.078	0.074	0.081	0.065

Notes: All χ^2 tests demonstrated p values $< .001$.

Table 2 (continued)

	Measurement invariance				
Scale	Median age	Test failed	Gender	Test failed	Combined
Balanced Inventory of Desirable Responding					
Impression Management	Failed	Scalar	Failed	Scalar	Failed
Self Deception	Failed	Configural	Failed	Configural	Failed
Bayesian Racism	Failed	Configural	Failed	Scalar	Failed
Belief in a Just World: General Just World Scale	Failed	Scalar	Passed	-	Failed
Big 5 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	Failed	Configural	Failed	Configural	Failed
Intuitions about Controllability and Awareness of Thoughts					
Others	Failed	Configural	Failed	Configural	Failed
Self	Failed	Configural	Failed	Configural	Failed
Need for Cognition	Passed	-	Passed	-	Passed
Need for Cognitive Closure					
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	Failed	Configural	Failed	Configural	Failed
Protestant Ethic	Failed	Configural	Failed	Configural	Failed
Ring-Wing Authoritarianism	Failed	Configural	Failed	Configural	Failed
Rosenberg Self-Esteem	Failed	Configural	Failed	Configural	Failed
Self-Monitoring	Failed	Configural	Failed	Configural	Failed
Social Dominance Orientation	Failed	Configural	Failed	Configural	Failed
Spheres of Control					
Interpersonal Control	Failed	Configural	Failed	Configural	Failed
Personal Efficacy	Failed	Configural	Failed	Configural	Failed

Notes: Passing measurement invariance requires meeting configural invariance (using same criteria as mixed CFA fit), metric invariance and scalar invariance (for each, meeting both ΔCFI $\geq$ -.15 and ΔRMSEA $\leq$ .01; see Chen, 2007). For full results of each test of measurement invariance see Supplementary Materials.

Discussion

The reproducibility and replicability of research findings, as well as our confidence in theory and application, requires valid measures. Yet as Flake et al. (2017) point out, structural validity is rarely reported in the literature, and even when it is, these tests are usually restricted to a single and flawed index (Cronbach's α). This raises the question: does the under-reporting of tests of structural validity represent a mere nuisance, insofar as these measures are ultimately valid, or the more troubling issue that there is an abundance of invalid measures hiding in plain sight (i.e., hidden invalidity). To examine this question, we submitted fifteen 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 median age and gender). Doing so seems timely and necessary given the broader re-evaluation 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 *nature of conclusions* (e.g., based on that evidence do we conclude that a measure is relatively valid or invalid, ranging 'good' to 'questionable' to 'poor'). We would argue that our results have strong *evidential weight* insofar as they are derived from a large and diverse sample (n per scale ≈ 6700), obtained across follow up periods, that speak to a wider than usual variety of structural validity metrics, and which consider many different measures. Indeed, to the best of our knowledge, this is the first paper to consider the full range of measures of structural validity, including multiple measures of internal consistency, test-retest reliability, confirmatory factor structure, and measurement invariance, and the first to simultaneously apply them to so many measures. We also acknowledge its 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 (we return to this latter point below).

In terms of the nature of conclusion, we employed a dichotomization strategy to synthesize results across scales. Most scales passed certain tests of structural validity: 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 surprisingly a tiny fraction (4%) demonstrated measurement invariance for both median age and gender. When considering all metrics together, only a single scale (Need for Cognition) passed all four metrics and can be said to have 'good' global structural validity. Results therefore

appear to suggest that the widespread under-reporting of structural validity highlighted by Flake and colleagues (2017) may reflect hidden invalidity. This begs the question: why was this is 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: very few studies have reported the full range of measures of structural validity. Our findings support this idea: the metrics that scales tended to pass or fail were not random. The less often a validity metric is reported in the literature (factor structure and measurement invariance) the more likely scales are to fail it. Conversely the more likely a metric is to be reported in the literature (Cronbach's α and test-retest *r*) the more likely scales are to pass it. This hierarchical or 'Guttman' structure among validity metrics can be seen in Figure 1. The correlation between failure rates and reporting rates highlights the potential for a general pattern of hidden invalidity throughout the discipline.

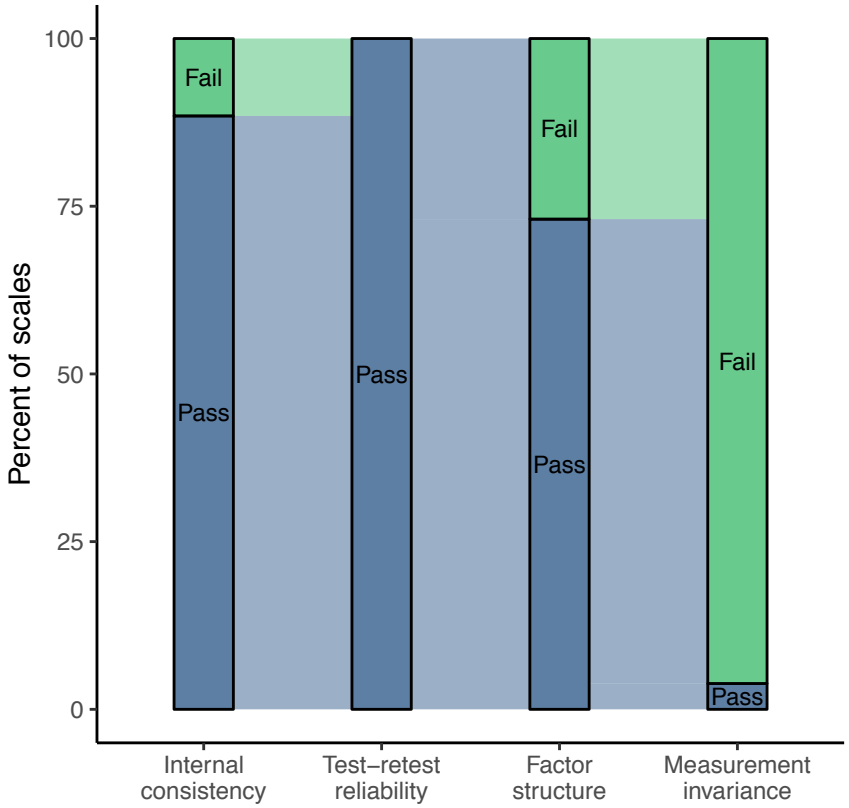

Figure 1. Alluvial plot illustrating the within and between scale patterns of passing or failing different tests of structural validity, arranged by frequency of reporting in the literature (from more common on the left to less common on the right)

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 were generally so good? One immediate possibility is that tests of confirmatory factor structure and measurement invariance are inherently stricter. A second is that we, as a field, have over-optimized our measures to demonstrate good consistency to the detriment of other psychometric properties.

To illustrate this idea more clearly, imagine that a researcher sets out to develop a new scale assessing 'negative automatic thoughts' within depression. After constructing her scale she attempts to determine how 'reliable' it is, calculates Cronbach's α, and obtains a value of $\alpha = .60$. As things currently stand, reviewers and users of the scale may comment that this value is problematically low. She may 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 (e.g., by excluding or rewording items and testing a new version of the scale). As a consequence, she is therefore less likely to spend her finite resources assessing and attempting to improve other aspects of structural validity, such as measurement invariance between groups. Yet doing so may have a larger pay-off than chasing α: without meeting measurement invariance, subsequent research using the scale with a certain group (e.g., depressed individuals pre- and post-therapeutic intervention) may incorrectly infer that those groups differ in terms of the latent variable (e.g., automatic thoughts in depression), when in fact they may simply interpret the items differently across the two measurement time points. For example, the therapeutic intervention may not serve to decrease the frequency of automatic thoughts (i.e., produce changes in the underlying latent variable), but instead increase participants' introspective abilities to more accurately report on the frequency of those thoughts (i.e., changes only in the measurement properties of the scale). This may lead to problematic or incorrect inferences 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, only that it (via Cronbach's α) is should not be the sole focus in structural validity assessment, 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 is likely to be used (e.g., measurement invariance for known groups, test-retest for longitudinal research, etc.). Failing to do so risks 'over optimizing' the measure on a flawed metric and without regard to other important but often overlooked properties.

Of course, the two possibilities discussed above (i.e., relative strictness of the tests vs. over optimization on internal consistency to the neglect of other forms of

validity) are not necessarily mutually exclusive in explaining the differential failure rates between the tests. In both cases, more rigorous reporting of these metrics is required.

We also considered a number of additional possibilities that could have contributed to our results, none of which are incompatible with the concept of hidden invalidity. A first possibility is that the scales themselves are less than optimal measures of the construct(s) 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 is not as originally assumed. It may also be the case that responding was influenced by additional factors that were either theoretically relevant (e.g., unintentional measurement of closely related but previously unappreciated constructs), or irrelevant (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 and/or item cross loadings (e.g., the Rosenberg Self-Esteem Scale: Mullen, Gothe, & McAuley, 2013; Salerno, Ingoglia, & Lo Coco, 2017; Supple, Su, Plunkett, Peterson, & Bush, 2013; Tomas & Oliver, 1999). In many cases, despite subsequent work suggesting that the factor structure is not what the scale's creators originally conceived, the originally-posited factor structures often represent the most common interpretation of scores on the scale, representing somewhat of a primacy bias in the use of many scales. Indeed, the resistance to incorporating emerging structural validity evidence for a given scale represents 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 scale validation process. We believe that this is unlikely given the sample was, if not representative of the general population, far more representative than that typically used in laboratory-based research.

Finally, it is possible that a given measure demonstrates poor structural validity because the construct it seeks to measure is poorly conceived of in the first place (i.e., in the substantive phase of validation: Flake et al., 2017), or poorly captured by the scale. Although this may seem unlikely given how well known many of these scales are, allowing for such a possibility protects against the reification of a construct merely because a scale has been created to assess it. In cases where such issues do exist, they 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. That is, whether 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). They 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 seek to assess depression but differ in their dimension of gradation: where one measures the *frequency* of depressive thoughts, another may measure the *degree of belief* in the literality of those thoughts, and yet another the experienced *emotional intensity* of those thoughts. The take home message here is that well-developed frameworks for measurement development already exist for those looking to construct or refine their scales. We encourage researchers to make better use of them. This includes attending to all three interrelated phases of validation (substantive, structural, external; Flake et al., 2017). Although we focused on the second phase, all phases of this process must be attended to when making a holistic evaluation about a measure's validity. One phase (e.g., structural) is neither sufficient nor singularly important relative to the other two (e.g., substantive and external), nor should one strive to maximize it at the expense of the others.

Implications and Future Directions

Our findings have implications for individual researchers in particular and the field more generally. To illustrate why, imagine that a researcher 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). She runs her study and then assesses if the scale she used provides a reliable index of the construct of interest. If she were to behave as most researchers do, she would answer this question by examining the consistency of her data, and in some cases, its test-retest reliability. These tests would tell her that the scale demonstrates adequate 'validity'. This necessity taken care of, she then proceeds to what is, for her, the real meat of the issue - interpreting her findings relative to her original hypothesis. Yet our findings suggest that if she were to adopt a more comprehensive assessment following best practices, then she would discover that the underlying factor structure of her construct and its invariance across samples would be problematic, thus leading her to exert more caution before interpreting her data. In other words, issues at the second phase of validation (structural) moderate our ability to make claims at the third phase (external validation), such as differences between known groups, interrelationships between constructions, and the prediction of behavior. As such, while questions of the structural validity of their measures may not be inherently appealing to all researchers, it is a requirement for making conclusions at other levels.

Another take-home message, that we have not seen explicated elsewhere, is that a finding can be both extremely replicable and yet give rise to invalid conclusions. For example, even if two groups (e.g., between depressive and non-depressive individuals) were shown across multiple studies to differ in their mean scores on a given scale (i.e., differences in the observed variable, such as the Rosenberg Self-Esteem Scale), this

replicable finding would only be interesting and useful 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 validity therefore has implications for the conclusions made on the basis of these measures.

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 we use. Indeed, even if the true rate of hidden invalidity were only a fraction of that observed here, this would still bring the conclusions of a large number of papers 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*, we may be unwittingly advancing a simplistic and overly positive view of how valid many of our most commonly used measures actually are, and drawing invalid conclusions on the basis of these scales. *At best*, this may be simply an issue of under-reporting scales that are ultimately valid. Yet until comprehensive reporting of tests of validity is common practice, we cannot know. We therefore encourage a more rigorous, multi-measure approach to structural validity across all areas of psychology where researchers identify and report, and reviewers and editors request, multiple sources of validity evidence. Note that although we endorse the idea that more widespread structural validity assessment should be done, we are not prescribing how it should be done, or presenting the methods or any cutoff values we use here as prescriptive recommendations. For pragmatic advice on improving measurement practices, readers are encouraged to read Flake and Fried (2019). That said, and for educational purposes, we have included simplified and commented R code to illustrate how we implemented our validity assessments in the Supplementary Materials.

Finally, two barriers exist that limit our ability to reach the aforementioned goal: (a) the staggering degrees of freedom available to researchers when assessing 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 their measures. In order to do so they

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 which frequently differ in their results, as well as making choices among numerous less visible experimenter degrees of freedom - 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, representing 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. Based on an analogy with *p*-hacking (Simmons et al., 2011), the first practice is what we will refer to as *v*-hacking, and refers to researchers selectively choosing and reporting a combination of metrics, their implementations, cutoffs, and other degrees of experimenter freedom in the assessment of structural validity that improve the apparent validity of their measures. For example, Watson (2004) noted that test-retest reliability studies are rarely conducted, but that when they are, 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). They may be driven to do so given the incentive structure present in research (e.g., reporting that a measure demonstrates adequate validity allows tests of one's core hypotheses using that measure, therefore increasing one's chances of being published; theories may only be supported or questioned on the basis of valid measures) and application (a valid measure is more likely to be adopted in applied settings; proprietary scales that are concluded to be valid are financially as well as academically incentivized). The second practice we refer to as *v*-ignorance, and refers to researchers simply relying on and reporting those metrics that others have used, without considering the issues underlying their use. Indeed, a recent 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, even *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 over inflation of the true structural validity of a measure and thus undermine the validity we have in our findings.

There are several ways to address and immunize research against these issues. One is for journals, editors, and reviewers to require the psychometric evaluation of all

measures used in a similar fashion to how effect sizes, confidence intervals, and precise p values are now commonly required (Parsons, Kruijt, & Fox, 2018). A second is for the field to come together and discuss issues such as choice of metrics, implementations, cutoffs, and other experimenter degrees of freedom. Let us be clear here: we are not advocating for the introduction of some set of universally applied cutoffs and metrics. Such an approach may lead researchers to mindlessly employ such values and raises a host of well-known issues (e.g., those associated with treating $p < .05$ or $BF_{10} \geq 3$ 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 our confidence in measurement. If the ongoing debate elsewhere around p values is any indication (e.g., Benjamin et al., 2018; Lakens et al., 2018), addressing this issue may take time and is unlikely to be trivial.

However, there is no reason to be pessimistic. Researcher degrees of freedom could be greatly constrained by expanding the use of pre-registration to also include measurement choices (e.g., metrics, cutoffs, measurement models, and decision-making strategies). Pre-registration 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 pre-registration 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 reported in a given article, and enables data to be pooled across studies for reuse and meta-analytic validation.

Conclusion

The current paper provides a psychometrically rich assessment of the structural validity of fifteen commonly used questionnaires. These analyses are useful for those who (a) are interested in a large scale examination of the structural validity and current state of measurement 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., regarding the decision to employ a measure in a future study, or compare their results with that found in large samples elsewhere); (c) want confidence that measures developed offline have good structural validity when used online; and (d) plan to use the AIID dataset for other purposes and need information about the structural validity of the scales therein. Perhaps most importantly, our findings suggest that the documented under-reporting of structural validity metrics in social and personality psychology represents an even more worrying issue of hidden invalidity among commonly used measures. We offer recommendations on how this might be addressed, with particular emphasis on pre-registration. The degrees of freedom afforded

to researchers are currently high, and validity-related decisions can be hidden or made post-hoc. This can lead to situations where there are little, if any, constraints that prevent researchers from cherry picking those validity metrics that presents the most favorable impression of their measures (v-hacking), to the potential detriment of the validity of their conclusions.

Author contributions

Both authors designed the study. IH wrote the analysis code and analyzed the data. Both authors wrote the article. Both authors approved the final submitted version of the manuscript.

Acknowledgments

Thanks to Jan De Houwer for his continued support of the Irish diaspora in Belgium.

Declaration of Conflicting Interests

IH and SH declare we have no conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was conducted with the support of Ghent University grant 01P05517 to IH and BOF16/MET_V/002 to Jan De Houwer.

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*(1), 32–50. https://doi.org/10.1037/0003-066X.63.1.32

Altemeyer, B. (1981). *Right-wing authoritarianism*. University of Manitoba press.

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

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: Recommendations for increasing replicability. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/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*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Bohannon, J. (2016). Mechanical Turk upends social sciences. *Science*, *352*(6291), 1263–1264. https://doi.org/10.1126/science.352.6291.1263

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/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*(3), 306–307. https://doi.org/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*(3), 464–504. https://doi.org/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*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281. https://doi.org/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*(4), 561–577. http://dx.doi.org/10.1016/S0191-8869(00)00055-6

De Schryver, M., Hughes, S., De Houwer, J., & Rosseel, Y. (2019). On the Reliability of Implicit Measures: Current Practices and Novel Perspectives. Preprint: https://psyarxiv.com/w7j86

Flake, J. K., & Fried, E. I. (2019, January 17). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. https://doi.org/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 and Personality Science*, *8*(4), 370–378. https://doi.org/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. Preprint: https://osf.io/n3axs/

Gosling, S. D., & Mason, W. (2015). Internet Research in Psychology. *Annual Review of Psychology*, *66*(1), 877–902. https://doi.org/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*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

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), 1–55. https://doi.org/10.1080/10705519909540118

Henry, L., & Wickham, H. (2019). purrr: Functional Programming Tools. https://CRAN.R-project.org/package=purrr

Hussey, I., Hughes, S., & Nosek, B. A. (2019). Attitudes, Identities and Individual Differences: A large dataset for investigating relations among implicit and explicit attitudes and identity. *Unpublished Manuscript*. 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* (Vol. 2, pp. 102–138). Elsevier.

Jorgensen, D. T., Pornprasertmanit, S., Schoemann, M. A., Rosseel, Y., Miller, P., Quick, C., & Garnier-Villarreal, M. (2018). *semTools: Useful tools for structural equation modeling*. 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*(6), 893. http://dx.doi.org/10.1037/0022-3514.55.6.893

Kuhn, M., & Wickham, H. (2019). rsample: General Resampling Infrastructure. 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*(3), 168–171. https://doi.org/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*(4), 371–379. https://doi.org/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*(3), 369–387. https://doi.org/10.1037/met0000093

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

Loevinger, J. (1957). Objective Tests as Instruments of Psychological Theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

McDonald, R. P. (1999). Test homogeneity, reliability, and generalizability. In *Test theory: A unified approach* (pp. 76–120). Mahwah, NJ: Lawrence Erlbaum Associates.

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*(3), 568. https://doi.org/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*(2), 153–157. https://doi.org/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*(1), 0021. https://doi.org/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*(1), 113. http://dx.doi.org/10.1037/0022-3514.65.1.113

Nosek, B. A. (2002). *Intuitions About Controllability and Awareness of Thoughts*. Unpublished data.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 201708274. https://doi.org/10.1073/pnas.1708274114

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory* (3rd edition). 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*(5), 966–980. https://doi.org/10.1037/a0022955

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Parsons, S., Kruijt, A.-W., & Fox, E. (2018). Psychological Science needs a standard practice of reporting the reliability of cognitive behavioural measurements, 25. https://doi.org/10.31234/osf.io/6ka9z

Paulhus, D. (1983). Sphere-specific measures of perceived control. *Journal of Personality and Social Psychology*, *44*(6), 1253–1265. https://doi.org/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, University of British Columbia, Vancouver, B.C., 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*(4), 741. http://dx.doi.org/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 : DR*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Revelle, W., & Condon, D. (2018). Reliability from α to ω: A Tutorial. Preprint: https://doi.org/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*(1), 9. https://doi.org/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), 1–36. https://doi.org/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. https://doi.org/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,* Advance online publication. https://doi.org/10.1177/2515245919838781

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*(4), 350–353. https://doi.org/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*(1), 107–120. https://doi.org/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*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018). False-Positive Citations. *Perspectives on Psychological Science, 13*(2), 255–259. https://doi.org/10.1177/1745691617698146

Snyder, M. (1987). *Public appearances, Private realities: The psychology of self-monitoring.* WH Freeman/Times Books/Henry Holt & Co.

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*(2), 370–389. https://doi.org/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*(5), 748–764. https://doi.org/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*(3), 375–388. https://doi.org/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*(1), 84–98. https://doi.org/10.1080/10705519909540120

Uhlmann, E. L. (2002). *The Bayesian Racism Scale.* Unpublished data.

Uhlmann, E. L., Brescoll, V., & Machery, E. (2010). The Motives Underlying Stereotype-Based Discrimination Against Members of Stigmatized Groups. *Social Justice Research*, *23*(1), 1–16. https://doi.org/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*(1), 4–70. https://doi.org/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*(4), 319–350. https://doi.org/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*(6), 1049. http://dx.doi.org/10.1037/0022-3514.67.6.1049