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 = 151,698 experimental sessions) to investigate the psychometric properties of some of the most wide used self-report measures (k = 15) in social and personality psychology. When assessed using the modal practice of considering only their internal consistency, all measures 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 60% demonstrated good validity. Furthermore, the less commonly a test is reported in the literature, the more likely it was to be failed in our analyses. This suggests that the pattern of underreporting 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, and suggest that the issue of validity hacking (v-hacking) should be acknowledged and addressed as well as the better-known concept of p-hacking.

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 often 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 on-going process known as construct validation (where measures could be self-report scales, implicit measures, or otherwise: see De Schryver, Hughes, De Houwer, & Rosseel, 2018; 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) and consists of three sequential phases. 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 reliability, 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 Social and Personality 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 sole source of structural validity evidence, and argue that rigorous methodologies for measurement are rarely reported.

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

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 construct 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, or 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.

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 robust 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 scales 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 with a large sample. To achieve this, we used the Attitudes 2.0 dataset, a large-scale, multivariate, planned-missing data study that was collected via the Project Implicit website (https://implicit.harvard.edu) between 2004 and 2007 (Hussey et al., 2018; 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 = 151,698 experimental sessions) allowed us to assess the psychometric properties of these measures with numbers that were far greater than those used in many earlier validations. Second, the dataset's structure allows for the application of a large range of metrics, including test-retest reliability across multiple timescales (immediate vs. up to 1 year later), something that has yet to be reported for many of these measures. 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 and stability: Revelle & Condon, 2018), factor structure (Confirmatory Factor Analysis), and measurement invariance. Although some of these analyses have been applied to individual measures, this is 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 within social and personality psychology (Bohannon, 2016; Gosling & Mason, 2015) 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 do not diverge from those observed in traditional (laboratory) settings.

Table 1. *Summary of structural validity analyses.*

Full scale	Split scales	Internal consistency	Test-retest reliability	Confirmatory factor structure	Measurement invariance	Overall measure evaluation
Balanced Inventory of Desirable Responding		Good	-	Mixed	Poor	Questionable
	Impression Management	Good	Good	Mixed	Poor	Questionable
	Self Deception	Good	Good	Poor	Poor	Questionable
Bayesian Racism		Good	Good	Good	Good	Good
Belief in a Just World (General Just World)		Good	Good	Good	Poor	Questionable
Big 5 Inventory		Good	-	Mixed	Good	Good
	Agreeableness & Openness	Good	Good	Mixed	Good	Good
	Extroversion, Conscientiousness & Neuroticism	Good	Good	Poor	Poor	Questionable
Humanitarianism-Egalitarianism		Good	Good	Good	Good	Good
Intuitions about Controllability & Awareness of Thoughts		Good	-	Poor	Good	Questionable
	Others	Good	Good	Poor	Poor	Questionable
	Self	Good	Good	Poor	Poor	Questionable
Need for Cognition		Good	Good	Good	Good	Good
Need for Cognitive Closure		Good	-	Mixed	Good	Good
	Order & Ambiguity	Good	Good	Mixed	Good	Good
	Predictability, Decisiveness & Close-Mindedness	Good	Good	Poor	Poor	Questionable
Personal Need for Structure		Good	Good	Mixed	Good	Good
Protestant Ethic		Good	Good	Mixed	Good	Good
Ring-Wing Authoritarianism		Good	Good	Mixed	Good	Good
Rosenberg Self-Esteem		Good	Good	Good	Good	Good
Self-Monitoring		Good	Good	Poor	Poor	Questionable
Social Dominance Orientation		Good	Good	Poor	Poor	Questionable
Spheres of Control		Good	-	Mixed	Poor	Questionable
	Interpersonal Control	Good	Good	Good	Good	Good
	Personal Efficacy	Poor	Good	Poor	Poor	Questionable
Summary	*Full scales*	100%	100%	80%	67%	60%
	Split scales	90%	100%	40%	30%	30%

Notes: Good internal consistency refers to McDonald's (1999) $\omega_t \geqslant 0.7$; good dependability refers to 1-hour test-retest $r \geqslant 0.7$; good stability refers to test-retest with follow up between 1 day and 1 year $r \geqslant 0.7$; good confirmatory model fit refers to meeting all of CFI $\geqslant .95$, TLI $\geqslant .95$, RMSEA $\leqslant .06$, and SRMR $\leqslant .09$, mixed confirmatory model fit refers to meeting SRMR $\leqslant .09$ and any one of CFI $\geqslant .95$, TLI $\geqslant .95$, or RMSEA $\leqslant .06$; 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 $\geqslant -.15$ and ΔRMSEA $\leqslant .01$; see Chen, 2007) for both median age and gender; good overall measure evaluation refers to having no poor fits on any of these metrics.

Method

Participants

The data of 151,698 experimental sessions representing 81,986 unique participants (50,141 women, 31,845 men, $M_{\text{age}} = 30.84$, $SD = 11.40$) were selected for inclusion 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. The modal number of participations was 1 (M = 1.77, SD = 2.22).

Measures

Fifteen scales were initially selected for inclusion in this study based on availability in the Attitudes 2.0 dataset: these are referred to as 'full scales'. Five of the full scales with a larger number of items were subdivided into two parts and delivered between participants due to time constraints on the Project Implicit site. These 10 total scales are referred to as 'split scales'. This resulted in a total of 20 scales being delivered to participants: the Balanced Inventory of Desirable Responding (Version 6: Paulhus, 1988; cited in J. P. 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 the full scales have previously been employed in a published article or book chapter, whereas one was 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 has been used to make theoretical conclusions without a published validation study: Uhlmann et al., 2010). Overall, the full scales 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 on OSF).

Procedure

In what follows we provide a brief overview of the Attitudes 2.0 study (for a more detailed description see Hussey et al., 2018). 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 Attitudes 2.0 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, and 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 Attitudes 2.0 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). Two final points are worth noting. First, participants had an approximately 5% chance of completing the same scale across a subsequent session and we used the data from this subset to assess test-retest reliability. Second, recall that five scales were subdivided into two for the purpose of the study (e.g., the Big Five Inventory). For scales that were split into two parts, participants also had an approximately 5% chance of completing the second half of that scale across a subsequent session. Note that this meant that no test-retest analyses were possible for the full scales. Data from this subset was pooled to assess the structural validity of the entire scale, using only participants who completed both of the two experimental sessions within one day. In what follows, we will refer to both the split and (recombined) full scales as ‘scales’. Where an original scale was split into two halves, we report analyses for both of the separate halves as well as the full scale.

Results

Data Preparation

Analyses were conducted on data obtained from the first experimental session in which a participant completed a given scale, 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 (Revelle & 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 package broom (D. Robinson, 2015). All code to reproduce our analyses is available on the Open Science Framework (osf.io/23rzk).

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 full in order to allow researchers to apply their own decision making methods if they so wish (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: a situation that applies to many researchers in social and personality psychology. We therefore apply common and recommended cutoff values to all metrics of structural validity in order to summarize and compare the relative validity of different scales and across different dimensions of structural validity.

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 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 = 6422, 4.8%) 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 analytic strategy which was informed based on recommended best practices (Hu & Bentler, 1999). First, confirmatory factor structure models for each scale were defined using the number of subscales (factors) stated in a scale's original publication. For example, if a scale was constructed to contain 2 subscales (e.g., the Spheres of Control scale), the CFA model was specified as two correlated latent variables with the appropriate items loading onto a latent variable for each subscale, and correlations between these latent variables to form a higher order latent variable. No factor cross loadings or methods factors (e.g., for reverse scored items) and were included. Refitting the models with orthogonal rather than correlated latent variables produced poorer performance across all scales.

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 was shown 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), 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 (CLI), and Root Mean Square Error of Approximation (RMSEA and 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 the combination of determining model fit based on scores within the cutoff values on both SRMR and one of CFI, TLI, or RMSEA. However, having no strong prior preferences among these multiple fit indices that could be applied across all scales, 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 cutoff 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]

Measurement invariance. A scale’s capacity to measure the same construct between populations or contexts typically involves three tests: configural invariance (i.e., does the same model fit to the whole sample also fit to both subgroups), metric invariance (refit the model to the whole sample but constrain the intercepts across indicators), and scalar invariance (refit the model but constrain both intercepts and factor loadings across indicators). When a scale passes all three steps, one can conclude that individuals in both groups interpret the items in an equivalent manner, and that the scale measures the same construct along the same latent continuum in both groups. While tests of measurement invariance are often performed between groups where the researcher wishes to later run inferential tests to compare these groups’ scores, it is also reasonable to assess measurement invariance between groups that one is not analyzing but which 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 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 wished 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 in personality between men and women are theoretically meaningful only if they represent differences in latent means rather than

[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.

factor loadings or intercepts. In all cases, measurement invariance is therefore necessary to subsequent substantive analyses.

Historically, the most common method of testing measurement invariance has been by assessing the statistical significance of changes in absolute model fit (Putnick & Bornstein, 2016; Vandenberg & Lance, 2000), but this was not suitable here due to the sensitivity of Chi square tests to our large sample sizes. Assessing measurement invariance via alternative fit indices such as RMSEA is also possible and increasingly popular (Putnick & Bornstein, 2016). Numerous simulation studies have been conducted on which index or indices to employ, along with which cutoffs to use. 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*, *k* items, number of assumed factors), and distributional information (*M*, *SD*, skewness, kurtosis). When combined, these tables detail a wide range of psychometric properties for fifteen commonly used self-report individual differences scales that can 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 (osf.io/23rzk).

Table 2. *Results of structural validity analyses.*

				Internal consistency								
					95% CI			95% CI			95% CI	
Scale	Total *n*	Factors	Items	α	Lower	Upper	ω_t	Lower	Upper	ω_h	Lower	Upper
Balanced Inventory of Desirable Responding	456	2	36	0.84	0.82	0.86	0.85	0.83	0.87	0.86	0.84	0.88
Impression Management	6934	1	18	0.80	0.79	0.80	0.80	0.79	0.81	0.80	0.79	0.80
Self Deception	6713	1	18	0.70	0.69	0.72	0.71	0.70	0.72	0.71	0.70	0.72
Bayesian Racism	6532	1	16	0.82	0.82	0.83	0.83	0.82	0.83	0.82	0.81	0.83
Belief in a Just World (General Just World)	6758	1	6	0.75	0.74	0.76	0.76	0.75	0.87	0.76	0.75	0.77
Big 5 Inventory	397	5	44	0.76	0.73	0.79	0.84	0.82	0.86	0.86	0.83	0.88
Agreeableness & Openness	6713	2	19	0.77	0.76	0.78	0.81	0.80	0.81	0.81	0.80	0.81
Extroversion, Conscientiousness & Neuroticism	6649	3	25	0.68	0.67	0.69	0.79	0.79	0.80	0.80	0.79	0.81
Humanitarianism-Egalitarianism	6906	1	10	0.84	0.83	0.85	0.84	0.83	0.85	0.83	0.82	0.84
Intuitions about Controllability & Awareness of Thoughts	446	2	18	0.89	0.87	0.90	0.89	0.87	0.91	0.89	0.87	0.90
Others	6711	1	9	0.75	0.74	0.76	0.75	0.74	0.76	0.74	0.73	0.75
Self	6830	1	9	0.80	0.79	0.80	0.80	0.79	0.81	0.80	0.79	0.81
Need for Cognition	6649	1	18	0.89	0.88	0.89	0.89	0.88	0.89	0.89	0.88	0.89
Need for Cognitive Closure	410	5	42	0.84	0.81	0.86	0.87	0.84	0.89	0.85	0.80	0.88
Order & Ambiguity	6585	2	19	0.78	0.77	0.79	0.80	0.79	0.81	0.80	0.79	0.81
Predictability, Decisiveness & Close-Mindedness	6559	3	23	0.75	0.73	0.75	0.80	0.79	0.81	0.78	0.76	0.79
Personal Need for Structure	6821	1	12	0.86	0.86	0.87	0.86	0.86	0.87	0.86	0.85	0.86
Protestant Ethic	6859	1	11	0.79	0.78	0.80	0.79	0.78	0.80	0.78	0.77	0.79
Ring-Wing Authoritarianism	6542	1	20	0.92	0.92	0.92	0.92	0.92	0.92	0.91	0.91	0.91
Rosenberg Self-Esteem	6971	1	10	0.89	0.89	0.89	0.90	0.89	0.90	0.89	0.88	0.89
Self-Monitoring	6623	1	18	0.76	0.75	0.77	0.76	0.75	0.77	0.74	0.72	0.75
Social Dominance Orientation	6854	1	12	0.83	0.82	0.84	0.83	0.82	0.85	0.82	0.81	0.83
Spheres of Control	402	2	20	0.78	0.75	0.81	0.80	0.77	0.83	0.80	0.75	0.82
Interpersonal Control	6785	1	10	0.78	0.78	0.79	0.79	0.78	0.79	0.78	0.77	0.79
Personal Efficacy	6899	1	10	0.64	0.63	0.65	0.64	0.62	0.65	0.62	0.61	0.64

Notes: Total *n* refers to the total number of participants with data available for internal consistency, distribution, confirmatory factor structure and measurement invariance analyses; Factors refers to the expected number of factors based on the original publication.

Table 2 (continued)

	Distribution				Test-Retest Dependability				Test-Retest Stability			
							95% CI				95% CI	
Scale	*M*	*SD*	Skew	Kurtosis	*n*	*r*	Lower	Upper	*n*	*r*	Lower	Upper
Balanced Inventory of Desirable Responding	120.36	20.90	-0.03	2.76	-	-	-	-	-	-	-	-
Impression Management	58.61	13.38	0.01	2.99	149	0.90	0.85	0.93	158	0.76	0.67	0.83
Self Deception	63.38	10.25	0.06	3.22	173	0.89	0.83	0.93	132	0.76	0.66	0.83
Bayesian Racism	41.09	13.16	0.40	2.91	136	0.83	0.73	0.90	154	0.88	0.83	0.92
Belief in a Just World (General Just World)	19.05	5.64	0.07	2.67	170	0.87	0.81	0.92	150	0.74	0.65	0.81
Big 5 Inventory	182.83	17.10	0.09	2.68	-	-	-	-	-	-	-	-
Agreeableness & Openness	85.99	10.49	-0.23	3.09	158	0.93	0.89	0.96	154	0.86	0.81	0.9
Extroversion, Conscientiousness & Neuroticism	96.70	11.64	-0.10	3.15	144	0.94	0.91	0.96	124	0.87	0.75	0.94
Humanitarianism-Egalitarianism	50.69	7.02	-1.20	5.14	131	0.91	0.84	0.95	149	0.84	0.76	0.9
Intuitions about Controllability & Awareness of Thoughts	51.45	13.88	0.02	2.97	-	-	-	-	-	-	-	-
Others	23.81	6.74	0.29	3.23	158	0.75	0.64	0.84	127	0.76	0.67	0.83
Self	30.22	7.86	0.16	2.99	195	0.86	0.79	0.9	143	0.78	0.69	0.85
Need for Cognition	80.91	13.80	-0.44	2.98	147	0.85	0.75	0.91	133	0.86	0.8	0.9
Need for Cognitive Closure	135.20	19.32	-0.12	3.41	-	-	-	-	-	-	-	-
Order & Ambiguity	76.93	12.33	-0.02	3.35	119	0.89	0.80	0.95	120	0.83	0.75	0.88
Predictability, Decisiveness & Close-Mindedness	60.93	10.36	-0.16	2.94	129	0.91	0.85	0.94	150	0.86	0.82	0.9
Personal Need for Structure	42.10	10.15	-0.06	2.82	149	0.88	0.83	0.92	141	0.81	0.71	0.88
Protestant Ethic	41.06	8.69	-0.18	3.05	163	0.92	0.88	0.94	158	0.85	0.8	0.89
Ring-Wing Authoritarianism	51.30	18.84	0.48	2.45	116	0.96	0.92	0.98	163	0.94	0.92	0.96
Rosenberg Self-Esteem	46.91	9.55	-0.80	3.35	160	0.95	0.92	0.96	156	0.9	0.86	0.93
Self-Monitoring	63.18	12.20	0.03	2.85	140	0.91	0.86	0.95	157	0.87	0.82	0.91
Social Dominance Orientation	25.58	9.85	0.71	2.98	161	0.91	0.86	0.94	149	0.84	0.77	0.88
Spheres of Control	83.74	11.06	-0.28	3.04	-	-	-	-	-	-	-	-
Interpersonal Control	38.22	7.49	-0.41	3.03	127	0.90	0.84	0.94	138	0.81	0.73	0.87
Personal Efficacy	45.41	6.12	-0.32	3.11	164	0.83	0.71	0.91	152	0.81	0.74	0.86

Notes: *r* refers to Pearson's *r* correlations between time points.

Table 2 (continued)

	Confirmatory Factor Structure										Measurement Invariance			
							95% CI							
Scale	χ^2	χ^2/df	df	CFI	TLI	RMSEA	Lower	Uppe r	SRMR	Fit	Median Age	Faile d	Gende r	Failed
Balanced Inventory of Desirable Responding	1048	1.8	593	0.93	0.92	0.04	0.04	0.05	0.06	Mixed	Poor	M	Poor	M
Impression Management	1834	13.6	135	0.95	0.94	0.04	0.04	0.04	0.04	Mixed	Good		Poor	S
Self Deception	3743	27.7	135	0.82	0.80	0.06	0.06	0.07	0.06	Poor	Poor	C	Poor	C
Bayesian Racism	1667	16.0	104	0.97	0.96	0.05	0.05	0.05	0.05	Good	Good		Good	
Belief in a Just World (General Just World)	153	17.0	9	0.99	0.98	0.05	0.04	0.06	0.03	Good	Poor	S	Good	
Big 5 Inventory	2037	2.3	892	0.91	0.90	0.06	0.05	0.06	0.07	Mixed	Good		Good	
Agreeableness & Openness	2971	19.7	151	0.93	0.93	0.05	0.05	0.05	0.05	Mixed	Good		Good	
Extroversion, Conscientiousness & Neuroticism	9917	36.5	272	0.91	0.90	0.07	0.07	0.07	0.07	Poor	Poor	C	Poor	C
Humanitarianism-Egalitarianism	485	13.9	35	0.98	0.97	0.04	0.04	0.05	0.05	Good	Good		Good	
Intuitions about Controllability & Awareness of Thoughts	451	3.4	134	0.94	0.94	0.07	0.07	0.08	0.08	Poor	Good		Good	
Others	1378	51.0	27	0.90	0.87	0.09	0.08	0.09	0.07	Poor	Poor	C	Poor	C
Self	1509	55.9	27	0.93	0.91	0.09	0.09	0.09	0.07	Poor	Poor	C	Poor	C
Need for Cognition	1959	14.5	135	0.98	0.98	0.05	0.04	0.05	0.04	Good	Good		Good	
Need for Cognitive Closure	1584	2.4	655	0.90	0.89	0.06	0.06	0.06	0.08	Mixed	Good		Good	
Order & Ambiguity	1556	17.5	89	0.95	0.95	0.05	0.05	0.05	0.05	Mixed	Good		Good	
Predictability, Decisiveness & Close-Mindedness	7027	31.0	227	0.88	0.87	0.07	0.07	0.07	0.07	Poor	Poor	C	Poor	C
Personal Need for Structure	1472	27.3	54	0.97	0.96	0.06	0.06	0.07	0.06	Mixed	Good		Good	
Protestant Ethic	1244	28.3	44	0.95	0.94	0.06	0.06	0.07	0.06	Mixed	Good		Good	
Ring-Wing Authoritarianism	6647	39.1	170	0.96	0.95	0.08	0.08	0.08	0.08	Mixed	Good		Good	
Rosenberg Self-Esteem	875	25.0	35	0.98	0.98	0.06	0.06	0.06	0.06	Good	Good		Good	
Self-Monitoring	11631	86.2	135	0.69	0.65	0.11	0.11	0.12	0.10	Poor	Poor	C	Poor	C
Social Dominance Orientation	1785	33.1	54	0.95	0.93	0.07	0.07	0.07	0.06	Poor	Poor	C	Poor	C
Spheres of Control	289	1.9	151	0.93	0.92	0.05	0.04	0.06	0.07	Mixed	Poor	M	Poor	M
Interpersonal Control	669	24.8	27	0.96	0.95	0.06	0.06	0.06	0.05	Good	Good		Good	
Personal Efficacy	1490	42.6	35	0.84	0.79	0.08	0.07	0.08	0.07	Poor	Poor	C	Poor	C

Notes: All χ^2 tests demonstrated p values $<$ .001; good confirmatory model fit refers to meeting all of CFI $\geqslant$ .95, TLI $\geqslant$ .95, RMSEA $\leqslant$ .06, and SRMR $\leqslant$ .09, mixed fit refers to meeting SRMR $\leqslant$ .09 and any one of CFI $\geqslant$ .95, TLI $\geqslant$ .95, or RMSEA $\leqslant$ .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 $\geqslant$ -.15 and ΔRMSEA $\leqslant$ .01; see Chen, 2007); Failed refers to the first test of measurement invariance failed when conducted in the order configural (C), metric (M), and then scalar (S).

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 index (Cronbach's α). This raises the question: does the field suffer from merely *under-reporting* tests of structural validity for measures that are ultimately valid, or from the more troubling issue of whether we have an abundance of invalid measures hiding in plain sight (*hidden invalidity*). To examine this question, we submitted fifteen self-report scales 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).

Our analyses benefitted from both a large sample size (n per delivered scale ≈ 6700) and, in order to accommodate a large number of scales, a consistent analytic strategy that followed best practices. 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 simultaneously applied them to so many measures used in social and personality psychology.

On the one hand, all of the full scales we assessed passed certain tests of structural validity: 100% demonstrated both good internal consistency and test-retest reliability. On the other hand, many scales also failed other tests of structural validity: only 80% demonstrated good fit with the expected factor structure, and only 67% demonstrated measurement invariance for both median age and gender. When considering all metrics together, only 60% of full scales passed all four metrics and can be said to have good overall structural validity. Importantly, the metrics that scales tended to pass or fail were not random: the less often a validity metric is used 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 used in the literature (Cronbach's α and test-retest r) the more likely scales are to pass it. This Guttman structure among validity metrics can be seen in Figure 1. This correlation between failure rates and reporting rates highlights the potential for a general pattern of hidden invalidity across the field.

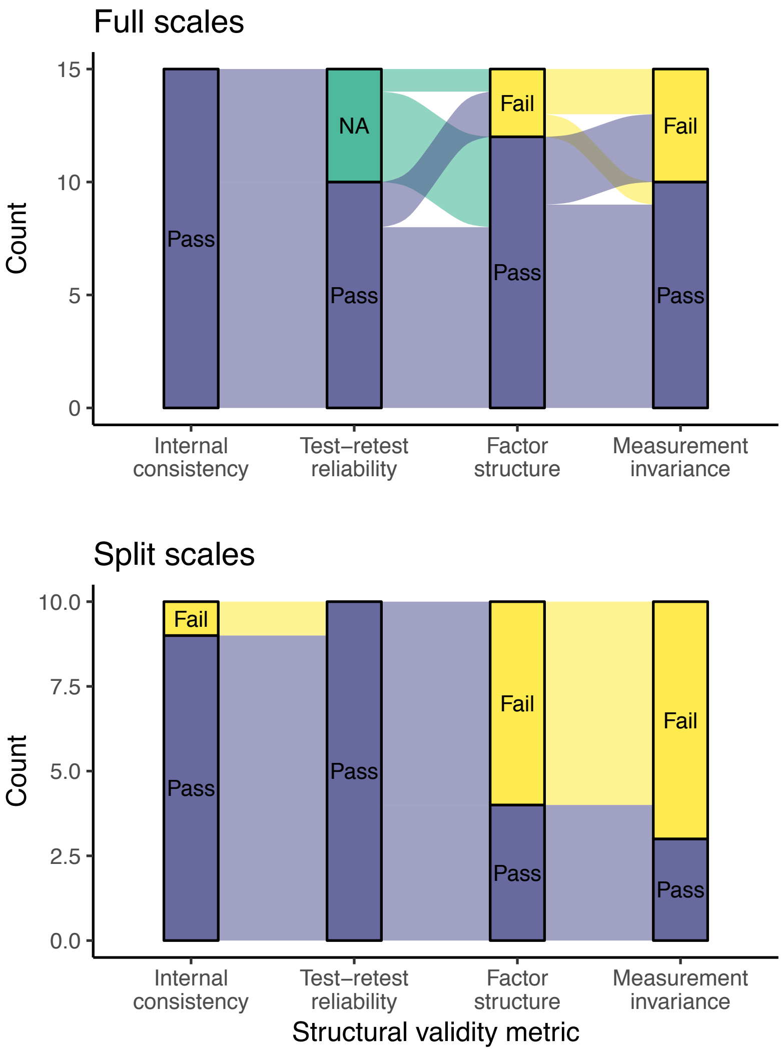

Figure 1. Failure rates between structural validity metrics

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 good? One possibility is that tests of confirmatory factor structure and measurement invariance are inherently stricter. An alternative is possibility is that we, as a field, have over-optimized our measures to demonstrate good consistency to the detriment of other psychometric properties. In either case, more rigorous reporting of these metrics is required. This also highlights the broader question of why only 60% of full scales demonstrated overall good structural validity when many of these scales are commonly used in the literature. One possibility highlighted by our analyses is that this invalidity may simply have been hidden until now. Few studies report the full range of measures of structural validity. This risk of hidden invalidity poses a threat to the validity of findings in social and personality psychology.

We also considered a number of additional possibilities that could contribute to our results, none of which are mutually exclusive 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 (e.g., because the items are more poorly worded that previously appreciated, or the structure among the items is not as originally suggested). 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 Rosenburg 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 unique 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 external phase of validiation: Flake et al., 2017). Although this may seem unlikely in many cases given how well-known many of these scales are, allowing for this possibility protects against the reification of a construct merely because a scale has been created to assess it.

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.

A key take-home message is that a finding can be both extremely replicable and yet invalid. For example, even if depressive and non-depressive individuals were reliably shown across multiple studies to differ in their mean scores on a self-esteem scale (i.e., on the observed or manifest variable), this is only interesting and useful if this reflects differences in mean Self-Esteem (i.e., scores on the latent variable) rather than merely differences in how the two groups interpret the items, producing spurious differences between the groups. 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. One third of the full scales we examined were found to be structurally invalid in some regard. As a thought experiment, imagine that the fifteen 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 reliable many of our most commonly used measures actually are, and drawing invalid conclusions on the basis of these scales. At best, we may not, and this may be an issue of under-reporting scales that are ultimately valid. However, until comprehensive tests of validity are common practice, we cannot know. We therefore echo Flake et al.'s (2017) call for 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.

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 numerous less visible experimenter degrees of freedom, as well as all the potential interactions these steps. In the absence of firm-guidelines, one's decision-making pathway when choosing how to report structural validity is massively unconstrained.

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 may involve financial as well as academic rewards). The second practice we refer to as *v*-ignorance, and refers to researchers simply relying on and reporting those cutoffs 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 notionally be motivated ignorance. For example, current modal practices do not involve the assessment of measurement invariance, and choosing to test for invariance can greatly decrease one's chances of publication (e.g., by revealing that a scale merely measures different latent variables between groups, and apparent differences between them are in fact invalid) while providing little added incentive to run such tests which are not currently rewarded by editors, reviewers, and readers. 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 immunize research against these biases. 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 statisticians to provide firmer recommendations on choice of metrics, implementations, cutoffs, and other experimenter degrees of freedom on the basis of simulation studies (for related arguments see De Schryver et al., 2018; Putnick & Bornstein, 2016). These efforts are ongoing within researchers who focus directly on psychometrics and statistics, but the speed at which statisticians conduct these results, communicate their findings to researchers who employ the techniques, and the adoption of techniques and best practices by researchers remains slow. A third is to explicate and pre-register one's decision-making pathway for tests of structural validity (e.g., the metrics, cutoffs, and other decisions made). Just as researchers should clearly outline and pre-register their design and analytic strategy prior to data-collection (Nosek, Ebersole, DeHaven, & Mellor, 2018), we recommend that they also start doing so for tests of structural validity. 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 meta-analytic validation.

Conclusion

The current paper provides a psychometrically rich assessment of the structural validity of fifteen commonly used measures. 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 measures (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 Attitudes 2.0 dataset for other purposes and need information about the structural validity of the scales therein. Our findings speak to the potential for hidden invalidity among many measures employed in social and personality psychology, highlight that this may be a prevalent issue, and offer recommendations on how it might be addressed. In the absence of firmer guidelines and standards for estimating validity, the degrees of freedom afforded to researchers are high and the validity-related decisions they make 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, to the potential detriment of the validity of their conclusions.

Authors' Note

Both authors designed the study. IH compiled and analyzed the data. Both authors wrote the article. Correspondence should be addressed to ian.hussey@ugent.be or sean.hughes@ugent.be.

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

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.

De Schryver, M., Hughes, S., De Houwer, J., & Rosseel, Y. (2018). On the Reliability of Implicit Measures: Current Practices and Novel Perspectives. *Unpublished Manuscript.* Retrieved from psyarxiv.com/w7j86

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

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

Hussey, I., Hughes, S., Lai, C. K., Ebersole, C. R., Axt, J. R., & Nosek, B. A. (2018). Attitudes 2.0: A large dataset for investigating relations among implicit and explicit attitudes and identity. *Unpublished Manuscript*. 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* (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.

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.

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.

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.

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. *Under Review.* https://doi.org/10.31234/osf.io/2y3w9

Robinson, D. (2015). *broom: Convert statistical analysis objects from R into tidy format.* Retrieved from https://github.com/tidymodels/broom

Robinson, J. P., Shaver, P. R., & Wrightsman, L. S. (Eds.). (1991). *Measures of Personality and Social Psychological Attitudes.* San Diego, CA: Academic Press.

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

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

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

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.