Special Issue Article

Construct Validation in Social and Personality Research: Current Practice and Recommendations

Jessica K. Flake[1], Jolynn Pek[1], and Eric Hehman[2]

Abstract

The verity of results about a psychological construct hinges on the validity of its measurement, making construct validation a fundamental methodology to the scientific process. We reviewed a representative sample of articles published in the *Journal of Personality and Social Psychology* for construct validity evidence. We report that latent variable measurement, in which responses to items are used to represent a construct, is pervasive in social and personality research. However, the field does not appear to be engaged in best practices for ongoing construct validation. We found that validity evidence of existing and author-developed scales was lacking, with coefficient α often being the only psychometric evidence reported. We provide a discussion of why the construct validation framework is important for social and personality researchers and recommendations for improving practice.

> Whatever exists at all exists in some amount. To know it thoroughly involves knowing its quantity as well as its quality.
>
> —Edward Thorndike

The dependability of research findings from diverse fields has recently come under scrutiny, including psychology (Pashler & Wagenmakers, 2012; Sijtsma, 2016; Simmons, Nelson, & Simonsohn, 2011). Such concerns have sparked discussion and facilitated the reexamination of many core statistical and methodological practices which might have contributed to a "replication crisis." These include the role of null hypothesis significance testing, the reporting of effect sizes and their confidence intervals (Cumming, 2014; Wilkinson & Task Force on Statisical Inference, 1999), data sharing and conducting replications (Nosek et al., 2015; Open Science Collaboration, 2015), and the preregistration of studies (Moore, 2016). To this point in discussions, measurement has gone relatively unexamined.[1] However, measurement is at the foundation of the scientific process: If a study's measures are not valid, then the conclusions have questionable meaning. The present research sampled from published papers to assess the quality of current practices in measurement.

Psychological phenomena investigated in psychology are often latent, in that the constructs of interest are typically unobservable (e.g., attitudes). Measures are developed and employed in the pursuit of studying these phenomena. For instance, the construct of life satisfaction is often measured by the 5-item Satisfaction with Life Scale (Diener, Emmons, Larsen, & Griffin, 1985). After responses to these items are collected and scored, these scores are taken to represent the construct of life satisfaction in data analysis and in the interpretations from analysis. Studying latent constructs of this nature, as opposed to observable variables such as height or weight, require the process of construct validation. This process begins with identifying a construct, defining it, developing a theory about the structure of the construct (e.g., how many factors are present, how they are related), selecting a means of measuring the construct (e.g., Likert-type scales), and establishing that the measure appropriately represents the construct. This process of construct validation is the means by which evidence is generated to support that scores reflect the target construct (i.e., have construct validity).

The verity of results about a psychological construct hinges on the validity of its measurement, making construct validation a fundamental methodology in the scientific process, particularly in psychology. If the construct of interest is studied with poor measurement, the ability to make any claims about the phenomenon is severely curtailed because what exactly is being measured is unknown and that uncertainty trickles down into the primary results.

[1] Department of Psychology, York University, Toronto, Ontario, Canada
[2] Department of Psychology, Ryerson University, Toronto, Ontario, Canada

Corresponding Author:
Jessica K. Flake, Department of Psychology, York University, 101 BSB, 4700 Keele St., Toronto, Ontario, Canada M3J 1P3.
Email: kayflake@gmail.com

Table 1. Examples of Validity Evidence and Resources for Each Phase of Construct Validation.

Phase	Validity Evidence	Description
Substantive	Literature review and construct conceptualization	Identifying depth and breadth of construct (Gehlbach & Brinkworth, 2011)
	Item development and scaling selection	Expert review (Gehlbach & Brinkworth, 2011)
	Content relevance and representativeness	Item mapping (Dawis, 1987), focus groups, and cognitive interviewing (i.e., think aloud; Willis, 2004), investigate construct under representation or irrelevancy (i.e., content validity; Sireci, 1998)
Structural	Item analysis	Response distributions, item–total correlations, and difficulty
	Factor analysis	Exploratory and confirmatory analyses including structural equation models and item response theory
	Reliability	Coefficients: α and ω (Mcdonald, 1999); interitem correlations, test–retest (McCrae, Kurtz, Yamagata, & Terracciano, 2011), dependability (Chmielewski & Watson, 2009)
	Measurement invariance (i.e., differential item functioning) testing	Multiple group factor analysis, item response theory, and differential item functioning tests (Millsap, 2011)
External	Convergent and discriminant	Correlations between other scales meant to capture similar and different constructs, multitrait-multimethod matrix analyses (Campbell & Fiske, 1959)
	Predictive/criterion	Regressions on criterion variables of import
	Known groups	Detecting differences between groups known to differ on construct

Note. Table draws from a collection of seminal works and texts on validation and measurement more broadly including Benson (1998), Clark and Watson (1995), Crocker and Algina (2006), Loevinger (1957), Strauss and Smith (2009), and Raykov and Marcoulides (2011).

Purpose of Study

To assess current practice, we conducted a systemic review of the use of psychological measures using a random sample of 30% of the empirical articles published in the *Journal of Personality and Social Psychology* (*JPSP*) in 2014. Many consider *JPSP* the flagship journal of social and personality psychology; accordingly, we assumed that all aspects of research within would be exemplary. Thus, we set out to determine to what extent researchers are utilizing rigorous methodology for construct validation. Prior to reporting results from our review, we briefly review the established standards for generating validity evidence of measures, reiterating the fundamental role of construct validity in strengthening the conclusions drawn from psychological research. Subsequent to reporting our results, we offer recommendations for improving the use of psychological measures so as to strengthen research findings in the areas of personality and social psychology.

Construct Validation

Construct validation is the process of integrating evidence to support the meaning of a number which is assumed to represent a psychological construct. Cronbach and Meehl (1955) described construct validation as necessary whenever "an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings." Further, construct validity pertains to a specific use of a scale (e.g., diagnosis or research) and can often be context or population dependent (Kane, 2013; Messick, 1995). Stated differently, a particular scale may only measure the intended construct within a specific context. Van Bavel, Mende-Siedlecki, Brady, and Reinero (2016) discuss this same issue, termed "contextual sensitivity," in relation to scientific reproducibility broadly.

They found that studies were less likely to replicate when the psychological processes under study were contextually sensitive. Just as some psychological processes may be influenced by context, so too can their measurement. Thus, the process of construct validation is best viewed as ongoing in which validity evidence is continually gathered in defense of findings.

The Standards of Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014) serve as an official reference, outlining best practice and methodology for conducting construct validation. These recommended practices have been categorized into three phases: substantive, structural, and external (Loevinger, 1957). The substantive phase comprises the theoretical underpinnings of a measure where previous literature is used to define the construct and outline its scope, describing the necessary content required for reasonably measuring the construct (i.e., items which tap certain dimensions). In the structural phase, quantitative analyses are used to examine the psychometric properties of the measure such as the factor structure or internal consistency. In the final, external phase, researchers gather evidence for how the construct relates to other constructs or predicts criteria, placing it in a larger nomological network (Cronbach & Meehl, 1955). These phases encompass a host of potential studies and methodologies which cannot be comprehensively reviewed here. Table 1 provides a nonexhaustive summary of validity evidence from each phase.

The majority of research conducted in the social and personality areas can be couched in the external phase. Although researchers may not explicitly be aware they are engaged in construct validation, they are implicitly conducting external validation when they gather information about a construct. The three phases of construct validation progress sequentially such

that conclusions made in the external (third) phase may not be valid if the construct does not have a strong theoretical foundation (first phase), and the scale which measures it does not have acceptable psychometric properties (second phase). Further, even though acceptable psychometric properties were previously determined by the researchers developing the scale, it does not mean that the scale will have these same properties in a different study (i.e., the measurement may not replicate; for a discussion, see Fabrigar & Wegener, 2016). And critically, if a scale does not have acceptable properties in current research, it is questionable whether the scale is measuring the same construct as determined previously. Thus, substantive and structural evidence of validity are prerequisites for considering findings that relate to the external phase or a replication study. If the measurement properties of a scale do not replicate, then the replicability of the results from analyses using those measures is suspect.

Given the recent interest in the replication of psychological findings, we investigated what methodologies social and personality researchers use to provide ongoing structural validity evidence for measures they employ. We first obtained a snapshot of the most pervasive applications of measurement and then focused specifically on *scales*, defined as measures that use items to represent a latent construct. We coded the structural validity evidence provided in support of the use of these scales. Using this review as a basis, we develop recommendations for improving measurement and ultimately, psychological findings. Specifically, we aimed to answer the following questions:

1. What types of measures are social and personality researchers using?
2. How often do authors report a previous validation study?
3. How often do authors report psychometric information?

Method

Sampling and Data Sources

The total number of articles published in the *JPSP* in 2014 ($N = 122$) served as the finite population. Among these articles, seven were editorials, errata, commentaries, or meta-analyses. A random sample of 39 (34%) empirical articles, stratified by substantive area (i.e., Attitudes and Social Cognition [ASC], Interpersonal Relations and Group Processes [IRGP], and Personality Processes and Individual Differences [PPID]), was drawn from the remainder for coding. Of the $N = 39$ sampled articles, 26% ($n = 10$) were from the section on ASC, 33% ($n = 13$) were from IRGP, and 41% ($n = 16$) were from personality PPID. These percentages corroborated with the finite population percentages observed (26%, 35%, and 39% for ASC, IRGP, and PPID, respectively) indicating fidelity of the sampling procedure.

For this review, we focus on empirical studies in which effects based on latent constructs are of interest; as such, we removed four articles from the sample (leaving $n = 35$) because they were a research synthesis, a theoretical paper, or a scale development paper. Although scale development papers focus on measurement, we considered them a different population from those utilizing scales because they focus on the full-scale development process and all phases of construct validation.

Coding of Articles

In this review, we focus on how researchers engaged in ongoing construct validation, specifically the validity evidence from the structural phase reported in "Method" section. Common approaches to this phase of construct validation are listed in Table 1. We focused on Method section because that is where the primary variables of interest and their psychometric properties (e.g., factor analysis or reliability) are typically described. Accordingly, our results exclude substantive or external construct validity evidence (e.g., theoretical breadth or predictive validity) possibly present in other sections. Additionally, we did not code for validity evidence of manipulation checks or measures that were not used in the final analysis. The current work is only a snapshot of one part of the construct validation process and should not be construed as reviewing all evidence researchers should report, such as other phases of validation and the validity of manipulation checks. We stress that manipulation checks serve the essential role of quantifying the internal validity (and construct validity) of experimental designs.

All articles were coded independently by a senior and junior coder for the frequency of evidence reported. The senior coders are authors of this article with formal training in measurement and statistics, whereas junior coders were student research assistants trained specifically for this project. We coded the frequency of reported evidence for each measure, which were objective observations (e.g., number of items on a scale or presence of a reliability coefficient), as opposed to subjective judgments. To ensure there were no data entry errors, we used double entry for all articles, whereby coders met to cross check any disagreement with the original work. Errors of entry were corrected, which resulted in one accurate data set for analysis, a common approach in reviews of measurement properties (e.g., Hulleman, Schrager, Bodmann, & Harackiewicz, 2010; Weidman, Steckler, & Tracy, 2016).

Results

Types of Measures Used

On average, we coded 4.02 (standard deviation [SD] = 2.16) experiments per article and a total of 700 instances of measures with an average of 20.00 (SD = 9.71) measures per article. Some of these measures were only used once within an article, but most were used repeatedly across experiments within a paper. When taking into account measures which were used repeatedly across experiments, we coded $N = 500$ unique measures, with an average of 14.29 (SD = 6.54) unique measures per article. Eighty-seven percent ($n = 433$) of these unique

measures were item-based scales in which questions or statements were combined in some way to form a composite score, meant to represent the construct of interest. These scales included 1-item measures, surveys, questionnaires, and tests. Thirteen percent of measures were not scales and varied in their approach to measurement; these measures included demographic variables, tasks, qualitative data, and observations.

In this study, we focus on *scales*, defined as measures that use items to capture a latent construct for which the process of construct validation is applicable. Among the unique scales, 30% ($n = 132$) were 1-item scales and 70% ($n = 301$) included more than 1 item. However, the specific number of items was not reported for 19% of unique scales ($n = 79$). For those with the specific number of items reported, the mean scale length was 4.69 ($SD = 6.35$, range $= 1$–58, $n = 354$). Excluding 1-item scales, the average scale length was 6.87 ($SD = 7.18$, range $= 2$–58, $n = 222$). Finally, 81% (351) used a Likert-type response scale. The second most common response scale was binary (e.g., yes/no or right/wrong), representing 4% of scales. Nine percent did not report the response scale.

Validity Evidence Reported

Validity evidence for a scale can take on two major forms: using evidence from a previous study (which assumes that evidence extends to the current study) and conducting sample specific analyses to provide ongoing evidence. As such, we coded and report how often authors used existing scales and how often structural validity evidence was reported for those scales. We intended to code for other information, but it was not routinely reported in Method section. However, it may have been presented in other areas. For example, some authors reported correlations between variables in "Results" section, but it was not presented as validity evidence for the scale in Method section. Such results are not reflected in our study.

Use of existing scales. Roughly half the unique scales, 53% ($n = 230$), were accompanied by a citation, suggesting that the scales had previously established validity evidence. Forty percent of the scales had no stated source and are assumed to be author created, whereas 7% of the scales were explicitly stated to have been developed by the author. Notably, of the scales ($n = 230$) which were cited from existing literature, 19% were modified or adapted in some way such that the psychometric information provided by the citation may not extend to the adapted version. Scales accompanied by a citation were longer on average ($M = 6.18$, $SD = 7.20$) than scales with no citation ($M = 3.43$, $SD = 5.25$).

Psychometric information. For this analysis, we focused on scales with 2 or more items, as 1-item scales require different validation methodologies (discussed later). Two types of psychometric evidence were presented in the reviewed articles: reliability coefficients and factor analyses. Table 2 presents frequencies and percentages of the type of structural validity evidence reported, split by whether or not a citation was provided.

Authors reporting the use of a previously developed scale, which accompany a citation, were more likely to report a factor analysis. Scales without a citation were likely to be shorter in length, with scales of 2–3 items not being appropriate for a factor analysis, which partly explains why so few researchers reported a factor analysis.

Some of these scales include instances where the author combined multiple scales to form a new scale or index. These *combination scales* included scales which were used separately in a previous experiment, a combination of previously published scales or a combination of items with multiple modes of responses such as a qualitative response with a Likert-type response. For example, one author noted two separate scales had low αs, reported combining the scales resulted in a higher α, and then created an average score from items across both scales. We coded 22 combination scales and 18 of those reported coefficient α as sole justification for combining measures.

Reliability coefficients. Given the frequent reporting of reliability coefficients, we further examined their characteristics. Of the scales that included 2 or more items ($n = 301$), coefficient α (Cronbach, 1951) was by far the most common reliability coefficient provided, comprising 73% ($n = 222$) of reported reliability information, with a correlation between 2 items representing 4%, the remaining scales did not report reliability information. One article utilized numerous scales and reported test–retest reliability in addition to α.

Of the scales for which α was reported, 15% were not specific estimates. Instead, a range across repeated measures designs or groups (e.g., $\alpha = .80$–.86) or the lowest estimate (e.g., $\alpha > .80$) were reported. Scales without their specific reliability coefficients ($n = 45$) were not included in the analyses.

Many scales were used multiple times within an article and some authors reported sample-specific α coefficients. Two hundred and forty-five estimates of α were reported for 166 unique scales. The average coefficient α estimate was .79, $SD = .13$, range $= .17$–.87. Figure 1 shows the distribution of α by whether a citation was provided for the scale. This plot shows that the variance in reliability is smaller for cited scales. We also ran a multilevel model to take into account the nested structure of these αs, as multiple αs were reported for unique scales within an article. This model included three parameter estimates: the expected grand mean (the intercept, γ_{00}), the variance within unique measures (σ^2), and the variance between unique measures (τ_{00}). The estimated grand mean of α across all unique measures was $\hat{\gamma}_{00} = .786$ and the variability of those αs across scales was $\hat{\tau}_{00} = .014$ ($SD = .12$).

Discussion

Latent variable measurement is at the foundation of social and personality psychological research. The importance of establishing construct validity for these measures is reflected in the many resources which outline best practices (AERA et al., 2014; Borsboom & Mellenbergh, 2004). These resources are

Table 2. Structural Validity Evidence Reported by Presence of a Citation for the Scale.

Evidence	Citation Provided (n = 177)		Author Developed or No Citation Provided (n = 124)	
	Count	%	Count	%
Reliability	138	78.0	100	80.6
Factor analysis	37	20.9	3	2.4
Reliability only	108	61.1	97	78.2
No information	31	17.5	24	19.3

Note. These percentages do not sum to 100% because scales sometimes included reliability coefficients and factor analyses.
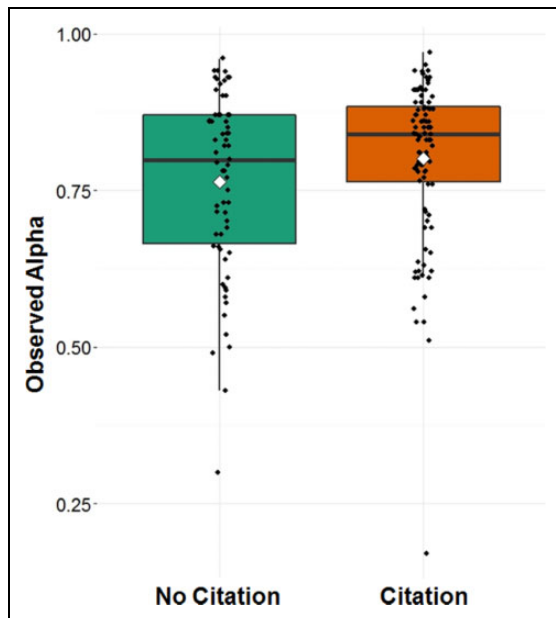
Figure 1. Boxplots of the α distributions for both novel and previously developed scales.

keys to strengthening the verity and dependability of findings. Generally, our results indicate that researchers who report structural evidence of ongoing construct validation in Method section of their paper are in the minority. This suggests that many constructs studied in social and personality research lack appropriate validation, which will contribute to questionable conclusions and difficulty of subsequent research to replicate. There is a vast field of measurement research that has together created best measurement practices (e.g., see Table 1), and we highlight key findings of our review and provide recommendations for improving current practice.

On the Fly Measurement

There is an abundance of latent variable measurement in social and personality psychology research. An average article in *JPSP* used 20 measures and latent variable measurement accounted for 87% of these measures. Roughly, half of these scales (46%) included no reference to previous validation, appearing to have been developed on the fly. α was the only psychometric information reported for half of these scales which had no previously published validity evidence, and 19% had no accompanying psychometric information.

These scales are intended to represent latent constructs, and the lack of validity evidence suggests that rigorous methodology for measurement has been overlooked by authors and reviewers. Valid measurement is a necessary prerequisite to the interpretation of results and cannot be ensured if no evidence is reported. For instance, researchers studying temperature need to ensure that their thermometer provides accurate readings of temperature before interpreting their results. In psychology, ensuring accurate scores from measures is more complicated, requiring an entire process of construct validation. When newly developed scales are reported, evidence is required to indicate that scores from these scales reflect the purported construct of interest, because these scores have a direct and dramatic impact on the theory researchers are developing. Until that evidence is available, any conclusions are questionable.

We recommend researchers consider their studies as part of a broader literature which encompasses substantive theory *including* what is known about how to best measure the constructs central to that theory. If a new scale is needed, then the full process of construct validation is necessary. We recognize that construct validation is a lengthy process, but it is theoretically and methodologically rich, providing the potential for numerous contributions to one's field. As such we recommend researchers and reviewers gather and look for multiple sources of validity evidence, especially when a scale has no cited source.

The Importance of Ongoing Validation

Nighteen percent of scales accompanied by a citation were explicitly said to have been adapted or modified, but new validity evidence for these scales was often not provided. Further, when citations were reported for existing scales, there was little discussion of why the scale would be valid for the current research context. For example, some research utilizes a small number of publically available items from the Graduate Record Exam (GRE), which was originally designed for graduate admissions. The GRE was intended have hundreds of items, and using a small subset of this total makes it unlikely that scores based on these items continue to reflect the intended construct. Although the items are not modified, the validity evidence supporting the original purpose of graduate admissions is unlikely to extend to this new purpose.

Constructs are in a constant state of validation, where researchers attempt to hone and expand existing theory using the evidence they garner in their studies. The measurement of these constructs similarly requires continual evaluation and refinement, which is why construct validation is discussed as an ongoing process in *The Standards*. Just as primary research findings can be context dependent (Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016), so too can measurement properties. If the validity evidence for a scale does not hold

in an adapted version or in a new context, then the scores do not represent the same construct and results based on these scores will not be comparable to the previous research. If researchers are using an adapted scale, or a scale in a new way, evidence is needed to show that the scale scores are valid representation of the construct. Examples of such psychometric evidence are described in Table 1 and include factor analyses which indicate the same factor structure as previous research. Another approach is studies of measurement invariance which test for the same measurement properties across different populations (see Millsap, 2011). We observed numerous studies which tested hypotheses across numerous populations (e.g., age-groups, cultures), but only one tested measurement invariance.

Big Theories, Small Scales

Thirty percent of scales we reviewed had 1 item, and the majority of scales without a citation had less than 3 items. Construct validation is built on the notion that when researchers develop items for a scale, they are sampling from a population of possible items. As such, short scales have historically been discouraged by the measurement community (e.g., Nunnally, 1978) because they would not adequately represent the construct and would lack in predictive power compared to multi-item scales (Diamantopoulos, Sarstedt, Fuchs, Wilczynski, & Kaiser, 2012). The case of using a 1-item scale requires careful consideration and validation (e.g., see, Robins, Hendin, & Trzesniewski, 2001).

We recommend researchers consider the construct they wish to measure and the adequacy of a couple items to fully capture the breadth of that construct (see construct representativeness in Table 1). For example, the construct of status includes multiple dimensions, such as wealth, social affiliation, and prestige (Cheng & Tracy, 2014), which would be difficult to capture with a short scale. If 2–3 items were used to represent status, they would provide an extremely narrow conceptualization of the construct and may not generalize to the larger theoretical domain or existing literature. Measurement of broader or multi-dimensional constructs requires longer scales. For narrow conceptualizations of a construct, strong validity evidence can justify the use of a shorter scale. Such analyses include comparing the predictive power of single item or short scales to a longer scale and using correction formulas to estimate scale reliability (see Eisinga, Grotenhuis, & Pelzer, 2013).

Limitations of α

Coefficient α was by far the most common type of evidence reported with regard to the psychometric properties of a scale. The average α was .78, and ranged from .17 to .97, with lower estimates of .60 and below being somewhat common (10%). We reiterate that these low αs were associated with scales used in primary analyses, suggesting that a substantial number of primary variables are measured with poor reliability. Further, there was a heavy reliance on α as the sole source of structural validity evidence. Over half of the scales which did not

accompany a citation (i.e., were explicitly said to have been author developed or a source was not stated) reported α as the only psychometric property. Although α is a useful tool for summarizing the internal consistency of items on a scale as a measure of reliability, reliability is necessary but not sufficient evidence of validity. Further, α has a long history of misuse and abuse in the social sciences (Schmitt, 1996), which our results corroborate.

A comprehensive review of the assumptions and uses of α is beyond the scope of this article. We highlight key information relevant to our findings and refer readers to comprehensive references in Table 1. Given certain assumptions, the α derived from a sample provides an estimate of internal consistency of items within a scale. These assumptions are expressed as an essentially tau-equivalent measurement model, which is a factor model where each item indicates only one factor, items have equal loadings, but item intercepts and error variances can differ. α was the most common and sole source of psychometric evidence reported for scales with no previously published source (78%), making it unclear whether such assumptions were met. To the extent that these assumptions are not met, α can be biased. We referred readers to Graham (2006), Sijtsma, (2009), and Yang and Green (2011) for details on how to test these assumptions in classical test theory and structural equation modeling framework. Additionally, we saw no reporting of McDonald's (1999) ω, which can be used under circumstances in which items measure the same factor but have unequal loadings.

It is incorrect to use α as a measure of unidimensionality, when unidimensionality is a prerequisite for its computation (Cronbach, 1951). In our review, α was used to justify combining multiple scales to form a single variable 18 times, implying that the misinterpretation of α as a measure of unidimensionality (Schmitt, 1996) continues today. There are numerous demonstrations showing that α can be high even if the scale has multiple and completely orthogonal factors (Cortina, 1993; Schmitt, 1996). When authors combine items or scales to form a combination scale, they are assuming that their scores represent a single construct. If the score used is a blend of numerous constructs, the results cannot capture the theoretical insights which would be gained by representing the factors separately, conflating several distinct psychological processes.

The heavy reliance on α also suggests that researchers are using it as a criterion for scale use and even item selection. Indeed, we noted numerous instances in which α was reported to justify item removal. Reliability is important to consider in construct validation, but it should not be maximized at the expense of other evidence. Drawing from the example of the broad construct of status in the previous section, we would expect a scale with numerous and similarly worded items, which captures a narrow conceptualization of status, to have a high reliability coefficient. However, this scale would not capture the breadth of the construct and lack in content validity. The construct validity of a scale cannot be boiled down into a single number, as evidenced by the list of potential validity

studies one could conduct in Table 1. Even with high reliability as measured by α, researchers should offer evidence from the substantive and structural phase of construct validation before moving on to interpreting results from primary analyses.

Conclusions and Recommendations

Our review indicates that the use of scales is pervasive in social and personality psychology research and highlights the crucial role of construct validation in the conclusions derived from the use of scale scores. It also indicates that the practice of conducting and reporting evidence of ongoing construct validation could be increased, which would be the benefit of the field. We recommend the many resources and practices regarding construct validity for researchers and the reviewers who will be evaluating their work in Table 1. In summary, the key points to take away are:

1. Consider valid measurement a prerequisite for interpreting the results of a study or a replication. If adequate measurement properties are not replicated, the rest of the results are necessarily not replicated.
2. Incorporate ongoing validation from all phases into your program of research and report on it, particularly if you have created a new scale, adapted an existing scale, or are using an existing scale in a new context or population.
3. Consider the construct representation and relevance when choosing items. Broad constructs will generally require longer scales.
4. Halt the sole and incorrect use of coefficient α.

In closing, we want to stress that a fundamental step toward supporting a research community in utilizing more rigorous methodology is formal training. In the most recent review of graduate training in psychology (Aiken, West, & Millsap, 2008), few departments offered a full course on measurement such as test construction or classical test theory (20–24% depending on the specific topic), with 20–42% offering no curriculum on any measurement topics. Given this lack of graduate training, it is likely that many social and personality researchers are unaware of the vast methodologies associated with construct validation. Psychometrics often utilizes advanced statistical modeling such as item response theory and structural equation modeling. However, full courses devoted to such topics are rare. So even for the researcher who is mindful of measurement, they may have had little experience using the methodologies needed to evaluate scales. We hope the present assessment of the field and recommendations may serve as a starting point for strengthening the research methodology of social and personality psychology.

Authors' Note

All authors designed the study. Eric Hehman and Jessica K. Flake compiled the data. Jessica K. Flake and Jolynn Pek analyzed the data. All authors wrote the article.

Note

1. Note that Andrew Gelman has written about this repeatedly on his blog http://andrewgelman.com/2016/03/03/more-on-replication-crisis/ and a few recent publications have focused on measurement issues within specific areas (Chmielewski, Sala, Tang, & Baldwin, 2016; Weidman et al., 2016).

References

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology. *63*, 32–50. doi:10.1037/0003-066X.63.1.32

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

Benson, J. (1998). Developing a strong program of construct validation: A test anxiety example. *Educational Measurement: Issues and Practice*, *17*, 10–17.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81–105.

Cheng, J. T., & Tracy, J. L. (2014). Toward a unified science of hierarchy: Dominance and prestige are two fundamental pathways to human social rank. In C. Anderson (Ed.), *The psychology of social status* (pp. 3–28). New York, NY: Springer. Retrieved from http://doi.org/10.1007/978-1-4939-0867-7

Chmielewski, M., Sala, M., Tang, R., & Baldwin, A. (2016). Examining the construct validity of affective judgments of physical activity measures. *Psychological Assessment*, *28*, 1128–1141.

Chmielewski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of transient error on trait research. *Journal of Personality and Social Psychology*, *97*, 186–202. doi:10.1037/a0015618

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*, 309–309.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.

Crocker, L. M., & Algina, J. (2006). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth Publishing Company.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological science. *Psychological Bulletin*, *52*, 281–302.

Cumming, G. (2014). *The new statistics: Why and how*. doi:10.1177/0956797613504966

Dawis, R. V. (1987). Scale construction. *Journal of Counseling Psychology*, *34*, 481–489. doi:10.1037//0022-0167.34.4.481

Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science*, *40*, 434–449. doi:10.1007/s11747-011-0300-3

Diener, E. D., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The satisfaction with life scale. *Journal of Personality Assessment*, *49*, 71–75.

Eisinga, R., Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health*, *58*, 637–642. doi:10.1007/s00038-012-0416-3

Fabrigar, L. R., & Wegener, D. T. (2016). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*, 68–80. doi:10.1016/j.jesp.2015.07.009

Gehlbach, H., & Brinkworth, M. E. (2011). Measure twice, cut down error: A process for enhancing the validity of survey scales. *Review of General Psychology*, *15*, 380–387. doi:10.1037/a0025704

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement*, *66*, 930–944. doi:10.1177/0013164406288165

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., & Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: Different labels for the same constructs or different constructs with similar labels? *Psychological Bulletin*, *136*, 422–449. doi:10.1037/a0018947

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1–73. doi:10.1111/jedm.12000

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*, 635–694.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal consistency, retest reliability, and their implications for personality scale validity. *Personality and Social Psychology Review*, *15*, 28–50. doi:10.1177/1088868310366253

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. Florence, KY: Routledge.

Moore, D. A. (2016). Pre-register if you want to. *American P Manuscript Manuscript*, *71*, 238–239. doi:10.1037/a0040195

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*, 1422–1425. doi:10.1126/science.aab2374

Nunnally, J. (1978). *Psychometric methods*. New York, NY: McGraw-Hill.

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

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. doi:10.1177/1745691612465253

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York, NY: Routledge.

Robins, R. W., Hendin, H. M., & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin*, *27*, 151–161. doi:10.1177/0146167201272002

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353.

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

Sijtsma, K. (2016). Playing with data: Or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*, *81*, 1–15. doi:10.1007/s11336-015-9446-0

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

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, *45*, 83–117.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 25. doi:10.1146/annurev.clinpsy.032408.153639

Van Bavel, J. J., Mende-siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, *I*, 6454–6459. doi:10.1073/pnas.1521897113

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2016). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*. Advance online publication. doi:10.1037/emo0000226

Wilkinson, L., & Task Force on Statisical Inference. (1999). Statistical methods in psychology journals. *American Psychologist*, *54*, 594–604. doi:10.1037/0003-066X.54.8.594

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks, CA: Sage.

Yang, Y., & Green, S. B. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment*, *29*, 377–392. doi:10.1177/0734282911406668

Author Biographies

Jessica K. Flake's research focuses on applications and evaluations of latent variable and random effects models for educational and social-psychological research. She works to develop and apply models for measurement, measurement invariance, and instrument design in ways that are accessible to substantive research communities.

Jolynn Pek's research focuses on quantifying uncertainties in statistical results of popular statistical models, and bridging the gap between methodological developments and their application.

Eric Hehman's research examines how individuals perceive and evaluate one another across group boundaries (e.g., race, gender, sexual-orientation, occupation, etc). To address these questions, he takes a multi-method approach, incorporating a broad range of behavioral (e.g., computer-mouse tracking, digital face modeling, group interactions), neural (e.g., fMRI, EEG), and statistical techniques (e.g., multilevel modeling, structural equation modeling).

Handling Editor: Gregory Webster