Advances in Methods and Practices in Psychological Science
1–4
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245920957618
www.psychologicalscience.org/AMPPS

Commentary on Hussey and Hughes (2020): Hidden Invalidity Among 15 Commonly Used Measures in Social and Personality Psychology

Eunike Wetzel[1] and Brent W. Roberts[2]
[1]Department of Psychology, Otto-von-Guericke University Magdeburg, and [2]Department of Psychology, University of Illinois at Urbana-Champaign

Received 8/26/19; Revision accepted 8/18/20

Hussey and Hughes (2020) analyzed four aspects relevant to the structural validity of a psychological scale (internal consistency, test-retest reliability, factor structure, and measurement invariance) in 15 self-report questionnaires, some of which, such as the Big Five Inventory (John & Srivastava, 1999) and the Rosenberg Self-Esteem Scale (Rosenberg, 1965), are very popular. In this Commentary, we argue that (a) their claim that measurement issues like these are ignored is incorrect, (b) the models they used to test structural validity do not match the construct space for many of the measures, and (c) their analyses and conclusions regarding measurement invariance were needlessly limited to a dichotomous decision rule.

First, we believe it is important to note that we are in agreement with the sentiment behind Hussey and Hughes's study and the previous work that appeared to inspire it (Flake, Pek, & Hehman, 2017). Measurement issues are seldom the focus of the articles published in the top journals in personality and social psychology, and the quality of the measures used by researchers is not a top priority in evaluating the value of the research. Furthermore, the use of ad hoc measures is common in some fields. Nonetheless, we disagree with the authors' analyses, interpretations, and conclusions concerning the validity of these 15 specific measures for the three reasons we discuss here.

Measurement Issues Are Not Ignored

First, the authors argued that structural validity is rarely reported in the literature. Readers may conclude that the field is cavalier about the quality of its measures. However, even though the top journals may not publish work on structural validity, hundreds of studies on exactly that topic are published annually in more specifically focused journals. Thus, the situation is actually worse than that portrayed by recent criticisms. The field does not undervalue good measurement practices because the research has not been done. Rather, the field so undervalues good measurement practices that it ignores the measurement research that *has* been done.

As a case in point, we conducted a cursory Google Scholar search for works focusing on the measurement invariance[1] or structural validity[2] of each of the scales Hussey and Hughes (2020) assessed. The results are reported in Table 1 at https://osf.io/msv2f/. As the table shows, we found, in most cases, multiple studies testing the structural validity or specifically the measurement invariance of each measure. Moreover, in several cases there was a robust, decades-long lineage of research showing not only that scholars do care about measurement issues, but also that these issues have been and continue to be the obsession of many researchers and the focus of many research reports. This is particularly the case in applied fields, such as industrial-organizational psychology. Leaving readers with the impression that researchers in the field of social and personality psychology do not care about or publish research on measurement issues would be a disservice to the many researchers who do wrestle with these issues on a continuing basis.

Simple-Structure Confirmatory Factor Analysis Models Do Not Match the Construct Space for Many of the Measures

The second issue that we would like to raise is the danger of unexamined assumptions and the application of

Corresponding Author:
Eunike Wetzel, Department of Psychology, University of Koblenz-Landau, Fortstrasse 7, 76829 Landau, Germany
E-mail: eunike.wetzel@uni-landau.de

potentially inappropriate standards when they are not warranted. Hussey and Hughes used confirmatory factor analysis (CFA) models with several assumptions built in, presumably because they considered such models to be the best and most appropriate for evaluating the measures in their study. There are at least two reasons not to make this assumption.

First, the standards used in CFA models, although accepted by the methodologists who employ them, were not established to develop and evaluate new measures. As we discuss, these standards are arbitrary in their own right (also see Hopwood & Donnellan, 2010).

Second, Hussey and Hughes chose to assume that constructs should have items with no cross-loadings within the CFA measurement models. This assumption fails to consider the theoretical construct space being tested in the case of several measures, especially measures of personality traits. Specifically, the assumption that factors should have no cross-loadings contravenes what is known about the items that go into personality-trait inventories. If one examines the history of measurement research on the Big Five, one finds that almost all Big Five personality items are multifactorial (Hofstee, de Raad, & Goldberg, 1992). For example, "dependable" and "reliable" load on both agreeableness and conscientiousness. On average, most terms psychologists use to describe people load on two factors, not one. Thus, the seemingly reasonable practice that calls for no cross-loadings for any factors in CFA models is inconsistent with the construct space of personality-inventory items. The choice to value and prioritize no cross-loadings may be something to aspire to in some measurement spaces. Nonetheless, it does not match the factorial complexity represented in the content of most personality items, and most likely most of the items used in all of the measures Hussey and Hughes assessed. Given the history of attempts to use CFA with the Big Five (Vassend & Skrondal, 1997), and given the factorial complexity of the Big Five, this seemingly innocuous decision was problematic from the start. If the goal is to test measures' structural validity, ideal modeling practices should be driven by theoretical and conceptual understanding of the construct space, and not by what may be common, but less optimal, default methodological practices.

A Dichotomous Decision Rule for Measurement Invariance Can Be Misleading

Finally, Hussey and Hughes accurately described the evaluative systems for determining measurement invariance across groups as unsettled, but then portrayed the field as if there is one accepted threshold and ignored several solutions that for years have helped applied researchers evaluate the importance of putative differences in measurement-invariance models. They used a "two-metric strategy" (p. 176) to evaluate measurement invariance across groups, such that a change of 0.015 in the comparative-fit index (CFI) and a change of 0.01 in the root mean square error of approximation (RMSEA) were the thresholds used to judge whether a scale demonstrated configural, metric, or scalar invariance. These thresholds were the only standards provided and used for qualitative distinctions such as "failing" to establish measurement invariance. Hussey and Hughes concluded that measurement invariance was poor for 14 out of the 15 questionnaires (and 25 out of the 26 subscales) and that the measures' global structural validity is therefore questionable, implying that researchers should stop administering these scales.

There are a number of problems with their analyses and in particular their conclusion: (a) Measurement invariance is not a pass/fail dichotomy, but rather is a matter of degree; (b) partial measurement invariance allows drawing comparisons between groups; and (c) there are many methods of testing measurement invariance, some of which also take the effect size of the noninvariance into account.

Hussey and Hughes implied that when the fit criteria are not met, measures fail the test of structural validity and therefore should not be used to make comparisons between groups. This is too simplistic. The procedure they applied for testing measurement invariance is a global test of whether full measurement invariance exists across all items in a scale. However, it is possible that the majority of items are invariant and only one or a few items are noninvariant. In this case, partial measurement invariance might still be achieved. All that is necessary for partial measurement invariance is for some items to be invariant, not all of them. The invariant items establish a common metric across groups, allowing comparisons to be drawn in the final partial-measurement-invariance model (Byrne, Shavelson, & Muthén, 1989; Reise, Widaman, & Pugh, 1993; Schmitt & Kuljanin, 2008; Steenkamp & Baumgartner, 1998). When there are few noninvariant items relative to the number of invariant items, estimates of the mean differences across groups will be unbiased (Guenole & Brown, 2014). Thus, with partial measurement invariance, it is still possible to compare groups, although this requires the estimation of latent mean differences because that is the only way to control for the noninvariance. Mean differences computed from observed scores, on the other hand, will be biased.

Hussey and Hughes used one particular method of testing measurement invariance, a method that is based on comparing model fit with cutoff criteria. It should be noted that there are many other methods that might lead to different results. These include nonparametric methods and methods in the framework of item response theory (IRT), in which measurement invariance is referred to as differential item functioning or differential test functioning (for an overview, see Penfield & Camilli, 2007, and for a comparison between CFA and IRT approaches, see Tay, Meade, & Cao, 2015). In contrast to CFA methods, IRT-based methods usually rely on considering the effect size of the noninvariance (for an overview, see DeMars, 2011), as in the classification system developed by Educational Testing Service (Zieky, 1993), although effect-size methods have also been developed for CFA (Nye & Drasgow, 2011). The advantage of these effect-size-based methods is that only substantial (e.g., moderate or large) noninvariance is flagged, whereas with other methods, especially those that are sample-size dependent (significance testing, ΔCFI), even noninvariance of negligible size might lead to the rejection of measurement invariance. Thus, the conclusion that 14 out of 15 measures have questionable structural validity because they had poor measurement invariance in one study using one criterion is unwarranted. Moving forward, more rigorous and nuanced investigations of the measurement invariance of popular personality measures are needed. Ideally, checking items for invariance across commonly formed groups (e.g., gender or age groups) would be part of the test-construction process, as in educational testing.

Conclusion

Taken at face value, the arguments of Hussey and Hughes, as well as other researchers, leave one with the impression that the field of social and personality psychology is experiencing a crisis caused by either poor measurement or a lack of good measurement research. In our opinion, the real problem is that psychological measurement is difficult, is complex, and requires more effort and energy than most researchers are willing to invest. Consumers of psychological measurement, arguably, want nothing more than to pull measures off the metaphorical shelf. This is unwise. We believe that a more productive way forward is for all researchers to more actively engage with prior measurement research, know the limits of existing measures, and invest in a deeper examination of the psychometric properties of their own measures in each of their studies. This can involve using measurement models that conform to the theoretical construct space by allowing cross-loadings, using observed scores only when it is justified, and considering effect sizes and partial invariance in measurement-invariance analyses. We believe that a more measured approach to measurement will improve all research efforts.

Transparency

Action Editor: Frederick L. Oswald
Editor: Daniel J. Simons
Author Contributions
Both authors wrote and revised this Commentary.
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Open Practices
Open Data: not applicable
Open Materials: not applicable
Preregistration: not applicable

ORCID iD

Eunike Wetzel https://orcid.org/0000-0002-4224-0366

Notes

[1] We combined "measurement invariance" with the name of each specific scale Hussey and Hughes reported on and limited the search to the first page of hits.

[2] When the specific scale had not been the focus of prior work, we looked at the literature on one or more closely related scales.

References

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456–466. doi:10.1037/0033-2909.105.3.456

DeMars, C. E. (2011). An analytic comparison of effect sizes for differential item functioning. *Applied Measurement in Education*, *24*, 189–209.

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*, 370–378. doi:10.1177/1948550617693063

Guenole, N., & Brown, A. (2014). The consequences of ignoring measurement invariance for path coefficients in structural equation models. *Frontiers in Psychology*, *5*, Article 980. doi:10.3389/fpsyg.2014.00980

Hofstee, W. K., de Raad, B., & Goldberg, L. R. (1992). Integration of the Big Five and circumplex approaches to trait structure. *Journal of Personality and Social Psychology*, *63*, 146–163.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review*, *14*, 332–346. doi:10.1177/1088868310361240

Hussey, I., & Hughes, S. (2020). Hidden invalidity among 15 commonly used measures in social and personality psychology. *Advances in Methods and Practices in Psychological Science*, *3*, 166–184. doi:10.1177/2515245919882903

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

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

Penfield, R. D., & Camilli, G. (2007). Differential item functioning and item bias. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 125–167). Amsterdam, The Netherlands: North-Holland.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566.

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

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review*, *18*, 210–222. doi:10.1016/j.hrmr.2008.03.003

Steenkamp, J. B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90. doi:10.1086/209528

Tay, L., Meade, A. W., & Cao, M. Y. (2015). An overview and practical guide to IRT measurement equivalence analysis. *Organizational Research Methods*, *18*, 3–46. doi:10.1177/1094428114553062

Vassend, O., & Skrondal, A. (1997). Validation of the NEO Personality Inventory and the five-factor model. Can findings from exploratory and confirmatory factor analysis be reconciled? *European Journal of Personality*, *11*, 147–166.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Erlbaum.