Dear Prof. Tackett,

We thank you and Reviewers 1-3 for your positive feedback on our original manuscript. We have revised the manuscript in order to address all the points that were raised and feel that the manuscript is much stronger thanks to this review process.

Below we provide point-by-point details of the changes we have made, including quotes and page numbers for these additions.

We look forward to hearing from you.

Kind regards,

Ian

Dear Dr. Hussey:  
  
Thank you for submitting your Empirical Article (AMPPS-18-0154) entitled ‘Hidden invalidity among fifteen commonly used measures in social and personality psychology’ to Advances in Methods and Practices in Psychological Science (AMPPS). I have now received evaluations of your submission from three reviewers with expertise in this area of research and scientific inquiry. As you will see, all reviewers were quite strongly positive about your submission, as was I on my own reading. All reviewers offered suggestions for consideration in a potential revision, as well, all of which I thought were quite feasible to address in a revised manuscript. Thus, it would be my pleasure to evaluate a revised manuscript for further consideration of publication.  
  
The reviewers offered such thoughtful and comprehensive reviews (of which I am very appreciative), that I have very little to add by way of substantive comment. I will note that I generally agreed that all of the reviewer comments would improve an already excellent manuscript, if attended to, and I did not find them to be in conflict with one another. My own interest in this topic could easily lead me to add on a number of suggestions of my own, but I think there are plenty of things to attend to already, in the reviewer suggestions.  I will point the authors to a relevant recent publication that should likely be incorporated in the discussion here, published also in this outlet (Tay & Jebb, AMPPS, 2018). Otherwise, I’d like to see careful attention to the comments raised by the reviewers.

If you choose to submit a revision, please include a letter detailing your point-by-point responses to each reviewer comment and indicating how you changed the manuscript to address them. The revised manuscript may undergo further peer review. We ask that you submit your revision within three months. Please let us know if you will not be able to meet this deadline.  
  
To submit your revision, log into <http://mc.manuscriptcentral.com/ampps> and enter your Author Center, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript number has been appended to denote a revision.  
  
IMPORTANT: Your original files are available to you when you upload your revised manuscript. Please delete any redundant or outdated files before completing the submission.  
  
Once again, thank you for submitting your manuscript to AMPPS and I look forward to receiving your revision.  
  
Sincerely,  
  
Jennifer Tackett  
Associate Editor, Advances in Methods and Practices in Psychological Science  
[jenniferltackett@gmail.com](mailto:jenniferltackett@gmail.com)

**General Comments and Replies**

**Editor:** I will point the authors to a relevant recent publication that should likely be incorporated in the discussion here, published also in this outlet (Tay & Jebb, AMPPS, 2018). Otherwise, I’d like to see careful attention to the comments raised by the reviewers.

**Authors**: We thank the Editor and Reviewers 1-3 for their positive feedback on our original manuscript. In our revision we address each of Reviewer 1-3’s comments in detail (*see below*). We have also incorporated and unpacked the implications of the above suggested article (Tay & Jebb, 2018) in our General Discussion (see new material on p.30).

Along with our resubmission, we have also included documentation of analytic and code changes between submission 1 and 2 (“changes between versions.docx”, found on the OSF supplementary materials). This document acknowledges that (a) certain results in our manuscript have changed during the revision process (namely that an increased proportion of measures fail our structural validity assessments) and (b) highlights the changes made to the data and analyses that are likely to have contributed to these changes in results (most of which came from changes made on the basis of reviewers comments). We decided to include this document in order to highlight and be transparent about the source and nature of these changes across versions of our manuscript.

# Reviewer 1 Comments and Replies

**Reviewer 1**: I am in no doubt that this paper should be published in AMPPS.   
  
In this manuscript the authors demonstrate that a sizable proportion of commonly used scales do not pass a comprehensive check of structural validity. It is troubling that validity reporting is so often confined to reporting Cronbach’s α and, as this paper suggests, that deeper validity issues are likely often missed due to under-reporting. This paper makes an important contribution to the literature by demonstrating in no uncertain terms that comprehensive assessments of measure validity are necessary and offers insightful recommendations for future research.   
  
I want to commend the authors for balancing their narrative. The authors avoid ‘preaching’ about the way research should be conducted and avoid any tone of accusation about sub-optimal approaches to structural validity in previous research. The authors manage this while simultaneously succeeding to convince readers of importance of measurement validity and the gravity of implications that many scales may not hold up to the standards required for robust research.  
  
This is an important paper with a large potential to influence measurement practices across fields. I have a few minor comments, but leave it up to the authors to determine how useful they will be. Primarily these are clarifications or additional points I think may help readers understand and make the most use from the paper.    
  
A benefit of the study’s sheer scale is the ability to conduct these analyses. Could the authors also comment on how applicable this is within individual studies, or give an idea of the sample size required?

**Authors**: We thank Reviewer 1 for his constructive comments on our paper.

In response to his first point, we would argue that similar analyses to ours could be carried out by other researchers in their studies. That is, we and others (e.g., Flake et al., 2017) advocate for the wider assessment and reporting of structural validity (e.g., see p.27-28; 30-31). Our hope is that after reading this paper, others will see the value and need to do so in their own work. Moreover, in response to Reviewer 1’s fourth comment (*see below*), we now provide easily useable and documented code to help others achieve this in their own work (see Supplementary Materials).

In response to Reviewer 1’s second point (i.e., required sample size) we are reluctant to provide a recommendation for required sample size values for different tests. This is because our main goal was to address Flake et al.’s (2017) question about whether *under-reporting* of validity in the social and personality psychology literature reflects hidden validity or hidden invalidity. Our goal was to simply highlight the widespread nature of a problem rather than make prescriptive recommendations for how others should solve it (i.e., we argue *that* more widespread structural validity assessment shouldbe done without prescribing *how* it should be done; see p.8):

“Let us be clear: our goal was not to make a final or absolute determination on the (in)validity of any of the scales we assessed; to make a binary determination of their (in)validity; or even to present our analytic strategy as a prescriptive set of standards for future work. This is not to say that our results cannot provide input into the ongoing process of validating these scales. Rather, our primary goal was to test the issue highlighted by Flake and colleagues (2017) - namely - whether the widespread under-reporting of structural validity information reflects hidden validity or, more worryingly, hidden invalidity.”

Our decision to avoid making prescriptive recommendations was based on the experience of others: in a recent paper, Simmons, Nelson, and Simonsohn (2018) reflect on their 2011 ‘False-Positive Psychology’ paper in which they also highlighted a problem in the wider literature (i.e., *p*-hacking). Reflecting back on that paper, they lament the fact that they specified specific sample-size values because doing so led readers to focus on precisely the wrong aspect of their paper (i.e., more on the suggestion for *n* > 20 observations per cell) and less on their central argument (that you cannot consistently get underpowered studies to produce significant results without *p*-hacking). We do not want to fall into that same trap (i.e., encourage others to mindlessly adopt some default strategy or criterion for validity assessment) and have it detract from our core point that underreporting likely hides invalidity, and so greater reporting is necessary. This point has also been added to the manuscript (see p.35).

Nevertheless, we do acknowledge that questions such as required sample size will likely be on many readers’ minds. We therefore point to a number of papers that provide a more thorough treatment of the process of carrying out structural validity assessment (e.g., Flake & Fried, 2019, see p.33).

**Reviewer 1**: On page 6 the authors comment that e.g. test-retest has not been reported for many of these measures – could the authors give an approximation of the rates of reporting – or is this following largely from the Flake et al. paper?

**Authors**: Our original claim on P.6 of the manuscript was that the structure of our dataset “*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*.” What we were originally trying to communicate here is that although past studies may have assessed the structural validity of individual measures, they typically do not assess those measures with (a) a large range of different metrics in a single study, (b) include tests of stability (test-retest reliability) with multiple or long delay ranges (e.g., up to 1 year), and do so (c) in a single study with comparable sample sizes as used here. This claim was partially based on Flake et al.’s paper as well as our own (unsystematic) review of the literature.

Rather than making hard claims about whether such tests have or have not been reported for what measures, which would require a systematic review for each scale, we have revised the manuscript to better qualify our original claim (see p.7):

“Second, the dataset’s structure allowed us to apply a large range of structural validity metrics to the same measure in the same study, include tests of stability (test-retest reliability) based on multiple delay ranges (immediate vs. up to 1 year later).”

**Reviewer 1**: A strength of the study is the large sample size. It may be also useful to mention the approximate minimum sample size required to apply these validity checks. For example, this may help others interested in determining whether they have a sufficient sample to test measurement invariance in an existing dataset or whether a targeted validation study is needed.

**Authors**: see our response to the first comment.   
  
**Reviewer 1**: I commend the authors on providing the full analysis code. This has the dual benefit of ‘checking other’s work’, but also importantly, will provide a useful resource and inspiration to others (myself included) to report their complete – and exact – analytical pathway. Early in the paper I could not help but think “I hope by the end of this paper that the authors show me how to do this, and I hope it’s not too complex because that will put off non-programmers”. I spent half of the time taken for this review looking through the code and supplemental materials – it is an excellent resource. To build on this, I wonder if the authors would consider creating an abridged version for a single scale that readers with minimal programming experience could use on their own dataset? This should be a relatively simple task that would have the benefit of providing readers who agree with the need for comprehensive validity assessments to then apply the approach to their own data. I understand that the code is all there, but a short tutorial/example to use the code would be an excellent resource.

**Authors**: Based on Reviewer 1’s suggestions we now include an abridged version of the code for a single scale that readers can use in their own dataset (see Supplementary Materials). We have been careful to annotate this as being for illustrative and educational purposes in order to minimize the risk that our practices are taken as prescriptive recommendations (a concern highlighted by both Reviewer 2 and 3).

**Reviewer 1**: The descriptions of each validity assessment are detailed and accessible. On page 5 when they are raised initially I wondered if readers might wonder why each is important or require further description – but, this is provided excellently in the methods. I wonder if, for impatient readers like myself, it might be helpful to include a sentence stating that these concepts will be described fully in the methods section.

**Authors**: Based on Reviewer 1’s suggestion we have now included a short sentence indicating that these concepts will be unpacked in the method section (see p.6).

“(*see the Method section for a more detailed treatment of different types of structural validity assessment)*.”  
  
**Reviewer 1**: On page 23 - “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”. Is this a potential trade off in developing scales? Or is the point more that researchers tend to concern themselves with α and not consider e.g. measurement invariance? It may be useful expanding on this point slightly for reader clarity. 

**Authors**: We thank Reviewer 1 for this question, which prompted us to give this issue deeper thought. Based on his suggestions we now better elaborate on our original point (see revisions on p.27-28).

“To illustrate this idea more clearly, imagine that a researcher sets out to develop a new scale assessing ‘negative automatic thoughts’ within depression. After constructing her scale she attempts to determine how ‘reliable’ it is, calculates Cronbach’s α, and obtains a value of α = .60. As things currently stand, reviewers and users of the scale may comment that this value is problematically low. She may then spend her limited time and resources attempting to improve α so that it tips over the commonly used and sought after (yet arbitrary) .70 cutoff (e.g., by excluding or rewording items and testing a new version of the scale). As a consequence, she is therefore less likely to spend her finite resources assessing and attempting to improve other aspects of structural validity, such as measurement invariance between groups. Yet doing so may have a larger pay-off than chasing α: without meeting measurement invariance, subsequent research using the scale with a certain group (e.g., depressed individuals pre- and post-therapeutic intervention) may incorrectly infer that those groups differ in terms of the latent variable (e.g., automatic thoughts in depression), when in fact they simply interpret the items differently across the two measurement time points. For example, the therapeutic intervention may not serve to decrease the frequency of automatic thoughts (i.e., produce changes in the underlying latent variable), but instead increase participants’ introspective abilities to more accurately report on the frequency of those thoughts (i.e., changes only in the measurement properties of the scale). This may lead to problematic or incorrect inferences that the intervention is effective in decreasing negative automatic thoughts in depression when in fact it is not.

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

**Reviewer 1**: I enjoyed the discussion of v-hacking and v-ignorance as barriers to limiting the ultimate goal of a more rigorous field-wide approach to validity evidence. I do hope the terms coined achieve popular use. I wondered where on this scale omitting to report reliability/validity within a single study falls on this scale? Or relatedly, using previous reliability/validity estimates and generalizing them to other samples? It might be worth highlighting these kinds of issues here in order to also highlight that this evidence would be reported (ideally) on a study by study basis.

**Authors**: In this paper we argue that the concept of v-hacking should be considered as a set of scientific behaviours that are distinct from, but analogous to, *p*-hacking, in that they represent questionable research practices in the assessment of validity rather than statistical inference. In order to briefly explore this concept, we make a broad distinction between v-hacking and v-ignorance, which represent two important subclasses of scientific behavior that have serious consequences for the confidence we have in research findings. We see many parallels between *p* and *v*-hacking and many of the questionable reporting, design, and analytic practices that lead to the former also apply to the latter.

How questionable or problematic a given behaviour is should be an empirical question. Given the backlash in some areas against a perceived moralization or accusation around the utilization of the concept of *p*-hacking, we think it is not useful to present a continuum of specific practices as problematic in the absence of data to support these claims. We also feel this is beyond the scope of the current manuscript. That said, we are finalizing a manuscript that unpacks these two concepts in greater detail, and quantifies how the relative influence of different v-hacking and v-ignorance practices impact the validity of conclusions. These questionable practices can be used to engage in what we are referring to as ‘measurement alchemy’ (i.e., transforming invalid into seemingly valid measures).

As mentioned above, we acknowledge that it is useful to describe the many possible forms of questionable measurement practices, as has been done by a recent paper that we now cite in the General Discussion (e.g., Flake & Fried, 2019, p.33).

**Reviewer 1**: Perhaps unsurprisingly, I agree with the four recommendations that should help to ‘immunize research against these biases’. Another might be the development of tools and resources to conduct a comprehensive validity assessment with relative ease, thus breaking down the barrier of v-ignorance (this aligns with my comment above on the authors’ very useful code). Another, related to the point about journals, would be facilitating publication of validation studies – particularly in the case of difficulties to publish reports based on low-validity measures. 

**Authors**: We strongly agree with Reviewer 1. As mentioned in our previous comment, we unpack comparable (and other) claims in our upcoming manuscript (and agree with similar arguments made in Flake and Fried’s recent “measurement-schmeasurement” preprint). We share Reviewer 1’s hopes to see these changes in the future.

# Reviewer 2 Comments and Replies

**Reviewer 2**: This paper examines the structural validity of 15 commonly used psychological measures. The paper has a number of important strengths. Its topic is interesting, and I certainly agree that measurement issues should receive more attention in social and personality psychology (and not just in measurement-focused papers!). The data set includes both a large number of measures and a large number of participants. The presentation is clear and engaging, and efficiently summarizes a large number of results. I also have some important concerns about the paper, but it may be possible to address these in a thoughtful revision.  
  
(A quick note up front: My review uses examples and citations concerning the Big Five personality traits, because this is the measurement literature that I am most familiar with. However, I believe that these points generalize to measures of other constructs. Also, I am not suggesting that the authors should cite all of the papers referenced in my review; these are simply meant as possibilities for further reading when revising the current manuscript.)  
  
Main Points  
  
1. The authors distinguish between three aspects of construct validity: substantive, structural, and external. They then focus on the second of these phases, and conclude that many commonly used measures in social and personality psychology do not meet one or more proposed benchmarks for adequate structural validity.  
  
When discussing the aspects of construct validity, I think it is important to consider that there is often a tension between structural validity and the other aspects, and it is not clear that this tension should always be resolved by prioritizing structural validity. This tension arises because commonly used structural validity indicators focus on internal consistency and simple factor structure, and this focus tends to pull researchers toward the development of scales with homogeneous, and potentially quite redundant, item content. This narrow content, in turn, can undermine both substantive and external validity.  
  
For example, consider the following hypothetical, eight-item Extraversion scale:  
  
I am talkative.  
I talk a lot.  
I enjoy talking.  
I like talking to people.  
I am rarely quiet.  
I am outgoing.  
I am sociable.  
I state my thoughts and feelings.  
  
I am confident that this measure would have higher internal consistency, a more unidimensional factor structure, and stricter measurement invariance than the BFI extraversion scale, due to the narrower range of item content, and lack of negatively keyed items, on the hypothetical scale. However, I am also confident that most personality psychologists would judge the BFI extraversion scale as having greater substantive validity, and I suspect that it would also have greater external validity across a representative set of extraversion-related criteria (i.e., not just frequency of talking).  
  
To be clear, I am not arguing that structural validity is unimportant; I think it is definitely important. But I also think it is important to acknowledge the tension between structural validity and other aspects of construct validity. To address this concern, I recommend that the authors (a) discuss this issue in both the introduction and discussion sections, and (b) provide advice to researchers about how to address this tension when developing or evaluating a measure.

**Authors**: We very much agree with Reviewer 2. There is a clear tension between structural and other aspects of construct validity, and exclusive focus on one aspect of validity is insufficient for holistic conclusions to be made about a given measure. We are explicit about this when we cite Flake et al. (2017), who emphasizes the interrelated nature of the phases. We also acknowledge this issue in the General Discussion (see p.30-31) (also see our response to a similar comment by Reviewer 1):

“This includes attending to all three interrelated phases of validation (substantive, structural, external; Flake et al., 2017). Although we focused on the second phase, all phases of this process must be attended to when making a holistic evaluation about a measure’s validity. One phase (e.g., structural) is neither sufficient nor singularly important relative to the other two (e.g., substantive and external), nor should one strive to maximize it at the expense of the others.”

Nevertheless, we believe that if researchers do not test for, or report tests of, one phase of validation (e.g., as is often the case with structural validity) then one does not have the necessary information to make informed conclusions about construct validity in general (i.e., to empirically test specific instances of Reviewer 2’s point about the tension between structural and external validity). It is important to note that the current paper is not concerned with the tension between structural validity and external validity *per se*, nor with the argument that one should necessarily maximize structural validity. It is not our aim, nor do we believe it our responsibility, to “*provide advice to researchers about how to address this tension when developing or evaluating a measure*”. We are now explicit about this on pages 8; 32-33. Rather, the core aim of our paper is to address a question outlined by Flake et al. (2017) on whether under-reporting of structural validity represents hidden validity or invalidity.

**Reviewer 2**: The authors employ a one-size-fits-all modeling approach, which models subscales using hierarchical factors, does not allow cross-loadings, and does not include method factors. This approach has the advantage of maintaining consistency across all of the various measures evaluated. However, it has the disadvantage that some of these modeling constraints seem inappropriate for some of the measures considered. Again using the example of the Big Five and the BFI, (a) it doesn’t seem appropriate to model agreeableness and openness, or extraversion, conscientiousness, and neuroticism, as subscales of a higher-order factor, because these scales are not meant to be combined into a single “personality score,” (b) this modeling approach doesn’t account for well-established and substantively meaningful cross-loadings (e.g., enthusiasm-and-activity-related extraversion items load negatively on neuroticism, depression-related neuroticism items negatively on extraversion, etc.), and (c) this approach doesn’t account for individual differences in acquiescent response style, which have been repeatedly shown to influence the item-level structure of measures with rating-scale or dichotomous response formats (e.g., Rammstedt & Farmer, 2013; Rammstedt, Kemper, & Borg, 2013; Soto, John, Gosling, & Potter, 2008), but can be modeled using method factors (e.g., John, Naumann, & Soto, 2008; Soto & John, 2017a, 2017b).  
  
To address this issue, I recommend (a) testing each subscale in a separate model (at least for measures in which the subscales are not aggregated into a single, total score), and (b) including a method factor for all scales with a mix of positively and negatively worded items; this could be an acquiescence method factor (for examples see John et al., 2008; Soto & John, 2017a) or a negative-item method factor.

**Authors**: Based on Reviewer 2’s suggestions we have altered our analyses to remove those that considered the BFI scales together as a single metric of personality (see updated tables and plots). We thank Reviewer 2 for his domain expertise here.

We have elected not to examine alternative models that include either method factors (e.g., for negatively worded items) or item cross-loadings. We think this is defendable for three reasons, and we thank Reviewer 2 for the opportunity to elaborate on these in the manuscript.

First, we strongly agree that a uniform analytic strategy enabled us to make comparisons between the scales. This also allowed us to address our central research question and was therefore necessary in this context.

Second, specific item cross-loadings are well established and mutually agreed upon for only a very small minority of the scales we employed (i.e., perhaps only the BFI). In most other cases they are not. This can be for one of two reasons: either the ‘true’ measurement model for a given scale is a matter of long debate and there is much disagreement in the literature (e.g., the Rosenberg Self-Esteem scale), or the scale has yet received relatively little scrutiny (e.g., the Bayesian Racism Scale). Apart from the BFI, many of our scales would therefore require multiple exploratory or weakly informed model choices, all of which would serve to focus more on within-scale determinations of validity than between scale inferences about whether underreporting represents hidden validity or invalidity.

Third, models that do not include item cross-loadings or method factors are most reflective of how these scales’ are actually (modally) used in research. Specifically, most researchers do not use data from these scales in the context of SEM analyses, but instead simply calculate sum scores and use these in subsequent analyses. By calculating sum scores, they tacitly endorse simple measurement models with no cross loadings or method factors. This point was recently discussed by Rose, Wagner, Mayer, and Nagengast (2019).

Given that calculating sum scores represents the modal practice for most of these scales (i.e., as demonstrated by Flake et al. 2017), this tacitly endorsed measurement model is most relevant to our research question. We recognize now that this reasoning was not explicated in the original manuscript, and thank Reviewer 2 for the opportunity to do so (see p.13-14):

“For all scales, simple measurement models were employed which did not involve method factors (e.g., negatively worded items) or item cross-loadings. We did so for three interrelated reasons. First, this uniform analytic strategy allowed us to compare rates of (in)validity across scales, in line with our primary research question. Second, with few exceptions (e.g., the BFI), most scale’s ‘true’ measurement model is either a matter of long debate (e.g., the Rosenberg Self-Esteem scale: see Mullen, Gothe, & McAuley, 2013; Salerno, Ingoglia, & Lo Coco, 2017; Supple, Su, Plunkett, Peterson, & Bush, 2013; Tomas & Oliver, 1999), or has received no scrutiny (e.g., the Bayesian Racism Scale). As such, the choice to employ alternative models (e.g., those considering method factors such as negatively worded items or item cross-loadings) would represent exploratory or weakly informed model choices, comparisons among which would detract from our primary research question. Third, most researchers who use these scale simply calculate sum scores and rely on these in their subsequent analyses. In doing so, they are tacitly endorsing simple measurement models with no cross-loadings or method factors (Rose, Wagner, Mayer, & Nagengast, 2019). By adopting similar assumptions here, our findings better reflect how these scales are commonly used and interpreted.”

**Reviewer 2**: Standards for testing measurement invariance, and the practical implications of non-invariance, are still under considerable debate (e.g., Nye, Allemand, Gosling, Potter, & Roberts, 2016; Nye & Drasgow, 2011). It would therefore be quite informative to report the effects of non-invariance on the age and gender differences observed in this sample. For example, what is the standardized mean-level difference (i.e., Cohen’s *d*) for the observed scale scores vs. invariant latent traits vs. partially invariant latent traits (with some constraints relaxed to achieve acceptable fit)?

A detailed treatment of this issue may be beyond the scope of this paper, but a supplementary table reporting these results, plus a brief discussion in the text, would be illuminating and increase the paper’s contribution to the measurement literature.

**Authors**: We agree that the assessment of measurement invariance is a matter of considerable debate. As such, we decided to employ modal existing strategies and recommendations of a recent systematic review of the MI literature (i.e., Putnick & Bornstein, 2016). We agree that unpacking the application of more recent and innovative analytic approaches is beyond the scope of the manuscript itself. However, based on Reviewer 2’s suggestions, we now report the differences in standardized between-groups differences calculated from the observed sum scores vs. latent scores (see Supplementary Materials). We also include a brief discussion of this in the manuscript (see p.18-19). These serve to provide a continuous metric of the impact of measurement invariance, as Reviewer 2 suggests. Additionally, we also elected to provide continuous effect size metrics for the CFA fits. These quantify a) the correlation between observed sum scores and latent scores calculated using the measurement model, and b) the proportion of individuals what would receive incongruent classifications (e.g., as being in the ‘high’ vs. ‘low’ group) on the basis of observed sum scores vs. latent scores.

**Reviewer 2**: p. 4. I believe “Journal of Social and Personality Psychology” should be “Journal of Personality and Social Psychology.”

**Authors**: Typo corrected.  
  
**Reviewer 2**: p. 15. I believe “CLI” should be replaced with “CFI.”

**Authors**: Typo corrected.  
  
**Reviewer 2**: Tables 1-2. References to “extroversion” should be changed to “extraversion.”

**Authors**: Typo corrected

# Reviewer 3 Comments and Replies

**Reviewer 3**: I reviewed “Hidden invalidity among fifteen commonly used measures in social and personality psychology”, which reported on the structural validity evidence for scales collected as a part of project implicit. I am excited to see such a large study of validity evidence of measures that are in use and think this paper will make a much needed and valuable contribution to the literature. I consider this type of work to be urgent for the field and hope it inspires others to undertake such projects. I have listed some suggestions for improving the clarity and accuracy of the manuscript below.

1.      The authors should be more consistent in how they discuss validity, particularly as this article is expected to be read by those who have limited exposure to validity theory. The authors discuss Loevinger’s phases of validation, The Standards, and Cronbach –these approaches to construct validity are framed more as how much evidence is there (?) and in the most recent iteration of the standards, how much evidence for a specific use or purpose?

I don’t expect the authors to provide a thorough lesson on the nuances of validity theory for us (those papers exist, authored by the authors of this paper even!), but they should be careful to present these definitions of construct validation and then, at other points in the paper present validity as a binary choice of good or bad or valid or not.

The authors use terms like “good validity” in the abstract, yet later describe the ambiguous nature of making decisions while conducting validation research. My concrete suggestion here is to discuss validation as the accumulation of different amounts and types of evidence (the phases align with this notion), and if that evidence is strong or weak. So some percent of scales had weak/little evidence, mixed evidence, or strong evidence and keep it consistent throughout.

I find this useful because this is what we do with other scientific claims, the evidence for the effect is strong or weak, or there are these strengths to the design, but these limitations. A study might be valid is some ways but not in others—I see validation research as sharing those name nuances and challenges and I see a need for us to consider measurement in the same way, to pull us away from trying to use α (or any other rule of thumb) to make a binary decision about scale validity.

**Author**: We thank Reviewer 3 for this comment. It helped us to further clarify a point we wanted to make in the paper but which was clearly not coming across on first attempt.

As we see it, Reviewer 3 has raised two distinct issues that we now clarify and correct in our manuscript.

The first is the delineation of *weight of evidence* on the one hand, and the *nature of the conclusions* on the other. For example, as Reviewer 3 highlights, it was unclear if language like “good” and “poor” in our previous manuscript referred to an *absence of evidence* for a scale’s validity (i.e., low weight of evidence, uncertain conclusions), or *evidence of absence of validity* (i.e., high weight of evidence, negative conclusions).

The latter concept (the nature of conclusions) introduces a second issue - namely - the tension between the use of categorical terms such as “good” validity in some sections of the paper and the treatment of validity as a multidimensional, continuous concept elsewhere.

We now address both of these issues in turn.

For instance, we have revised the manuscript to address the first issue:

* First, we now acknowledge the separation between weight of evidence and nature of conclusions. Specifically, we clarify that we believe our results have strong *evidential weight* insofar as they are derived from a large and diverse sample, obtained across follow up periods, speak to a wider than usual variety of structural validity metrics, and consider many different measures. However, we also acknowledge its evidential weaknesses, in that recruitment was from a single population (i.e., an online sample), and that we consider only the structural phase of validity assessment but not the external (see p.25-26).
* Second, we clarify that the use of terms such as ‘Good’ and ‘Questionable’ refer to the *nature of conclusions* about validity rather than the weight of evidence (which we consider to be relatively strong).

This then leads us to the second issue (appropriateness of categorical terms such as ‘Good’, ‘Questionable’, or ‘Poor’ to summarize conclusions about the structural validity of scales). Our revised manuscript clarifies that the function of these labels is to condense multifaceted metrics of validity to categorical conclusions in order to enable decision-making with regard to our core research question (i.e., whether under reporting represents hidden validity or invalidity).

Critically, we modified the manuscript to ensure that others do not treat these categorical conclusions as literally ‘true’ for other research questions (e.g., the adequacy of a scale for use in their future research). For instance, we explicitly state that they should not do so (see p.20), repeatedly acknowledge that our research questions was one of hidden invalidity rather than making absolute validity decisions about specific scales, and clearly distinguish the categorical summary table with labels such as “Good” vs. “Questionable” from the continuous and multifaceted results table, explicating that the latter should serve as the main basis for drawing personal, subjective conclusions about the scales tested (e.g., see p.20):

“In Table 1, we use categorical terms such as ‘Good’, ‘Questionable’, or ‘Poor’ to summarize conclusions about the structural validity of scales based on the cutoff values discussed above for each dimension of validity. These labels serve to condense multifaceted metrics of validity to categorical conclusions in order to enable decision-making with regard to our core research question (i.e., whether under-reporting represents hidden validity or invalidity). This tradeoff between nuance and heuristic value is analogous to the use of *p* values, which are natively continuous, but which are often reduced to a significant versus non-significant dichotomy to facilitate conclusions regarding hypotheses. These categorical labels should not be taken as literally ‘true’ for any other research question than our own (e.g., when assessing the adequacy of a scale for future use). Instead, such questions should be informed by the continuous and multifaceted results reported in Table 2, which offer a more nuanced perspective on structural validity.”

**Reviewer 3**: Related to the last point—in the discussion the authors describe a lack of clear cutoffs and metrics. I agree we need more guidance on how to plan and interpret validation studies and how to ensure validity when using existing scales, etc., but I don’t think creating a set of cutoffs will get us there. Forced objectivity into a process that is inherently theoretically ambiguous (measuring latent variables) is problematic and contributes to the use of rules in defense of poor/thoughtless practices (similar to what we have done with NHST and effect sizes, see: The ouroboros of psychological methodology: The case of effect sizes (mechanical objectivity vs. expertise). I think focusing on transparency of the decision making pathway (particularly via reg reports) and continued methodological research that grabbles with some of these difficult decisions is more useful.

**Authors**: In raising the point about clarity for cut-offs and metrics, we are not arguing that the field necessarily needs to converge on some absolute or final set of cut-offs or metrics when assessing for validity. If anything we agree with Reviewer 3 – the mindless application of such cut-offs and metrics has contributed to many of the problems psychological science has recently found itself in (e.g., acting as if *p* < .05 or BF10 = 3 are some absolute or strong thresholds for statistical conclusions to made). Instead we simply highlight that the absence of any clear or coherent recommendations concerning cut-offs and metrics is a type of questionable measurement practice (see Flake & Fried, 2019: [*https://psyarxiv.com/hs7wm/*](https://psyarxiv.com/hs7wm/)). Or at least a practice that allows for massive research degrees of freedom at the measurement level.

As things currently stand, researchers are free to select from a range of cut-offs and metrics (and find corresponding papers to support that selection) which can in turn lead them to present their measure as having ‘good’ structural validity. Yet if they were to select another set of cut-offs and metrics (and rely on another set of papers to support that conclusion) they may present the same measure as having ‘poor’ structural validity. As we note in our replies to Reviewer 2, we unpack this issue in great depth in a subsequent paper on *v*-hacking (currently in preparation).

For now, we want to stress that we are not arguing for forced objectivity in terms of using a small set of cut-offs and metrics. Instead we are arguing for forced objectivity when it comes to selecting, justifying, and pre-registering which cut-offs and metrics are employed within and between studies in a more transparent and open way.

We have revised our paper to clarify the above point (see p.35).

“A second is for the field to come together and discuss issues such as choice of metrics, implementations, cutoffs, and other experimenter degrees of freedom. Let us be clear here: we are not advocating for the introduction of some set of universally applied cutoffs and metrics. Such an approach may lead researchers to mindlessly employ such values and raises a host of well-known issues (e.g., those associated with treating *p* < .05 or BF10 ≥ 3 as a sacrosanct threshold; for related arguments see Simmons, Nelson, & Simonsohn, 2018). Rather we hope that readers will recognize that massive heterogeneity in the choice of cutoffs and metrics serves to inflate research degrees of freedom, and therefore threatens our confidence in measurement. If the ongoing debate elsewhere around *p* values is any indication (e.g., Benjamin et al., 2018; Lakens et al., 2018), addressing this issue may take time and is unlikely to be trivial.”

**Reviewer 3**: The discussion of reliability could be improved to be more concise. I agree it has effectively become a sole source of validity evidence and fits under the structural validity aspect of Loevinger’s phases of validation. But reliability is not validity and this point should be made very clear, early on in the paper.

There could be more thorough treatment of the issue of reliability and validity tradeoff in this paper in general – what do we lose by maximizing reliability, from a validity perspective? The results of this study highlight a persistent problem, we trick ourselves into thinking a high α gives us some information about the measure’s representation of the construct, when in fact, it gives us none.

**Authors**: It’s worth noting that trade offs between reliability and validity are to be made by researchers faced with choices when developing or refining a measure (e.g., the creation, selection, exclusion, or change in items or response options). Given that the current article is solely concerned with assessing properties of extant scales, and not their modification, the reliability-validity trade off (though important) is beyond the scope of our paper.

**Reviewer 3**: Related to the last point—reliability is discussed in tandem with factor analysis, but it more appropriate to think of the factor analysis as testing a fundamental assumption of the reliability—that you are measuring one thing. If the reliability estimate doesn’t reflect one thing—it loses meaning. What is the reliability of two things? This point is not discussed at all, it would be more accurate to discuss the use of reliability coeffs without assessing the assumptions, and what you find is that there is evidence these assumptions are not met regularly.

**Authors**: We agree with Reviewer 3’s point and have therefore added new content in the revised manuscript that explicates our logic and guards against misinterpretation of the ordering of our results as prescriptive for future research (see p.7-8):

“It is worth noting that the sequential ordering of the tests we carried out, as reported in text and in Tables 1 and 2, was according to the frequency with which they are reported in the literature (see Flake et al., 2017). We adopted this strategy in order to demonstrate the inverse relationship between rate of reporting and hidden invalidity. Note we are not suggesting that other researchers should sequence their analyses or reporting in a similar way. Indeed, as argued elsewhere (Flake et al., 2017) the most common test (α) makes numerous assumptions that can only be assessed by less commonly applied analyses (e.g., within a CFA context).”

**Reviewer 3**: Measurement invariance could be more explicitly defined - the definition of configural is particularly unclear - the “whole model” could mean the intercepts and slopes, be clear it is the specification of which items map to which factors, this would be useful early on because I also see a need for those scales which failed configural invariance to have more space in the discussion. The results indicate that a non-trivial number of scales form different factors across gender and age, yikes! That has massive theoretical and practical implications. Be clear about what measurement invariance tells us theoretically when you define it and in the discussion of the results of your tests in relation to that.

**Authors**: We thank Reviewer 3 for pointing this out. We have tightened these definitions, with reference to a recent systematic review of the field of measurement invariance (see p.17):

“A scale’s capacity to measure the same construct in a comparable way between populations or contexts typically involves three component tests: (1) configural invariance (i.e., equivalence of model form: does the unconstrained model provide adequate fit in each of the groups), (2) metric invariance (or weak factorial invariance; i.e., equivalence of factor loadings), and (3) scalar invariance (or strong factorial invariance; i.e., equivalence of item intercepts or thresholds; Putnick & Bornstein, 2016). These are typically assessed as nested models, whereby the initial measurement model is first fit to each group’s data, a second fit constrains factor loadings to be equivalent, and a third fit constrains item intercepts (or thresholds) to be equivalent. Change in fit metrics between these nested models is then typically used to determine whether each test is passed in sequence. When a scale passes all three tests, one can conclude that correlations between scores on the scale and other external variables have equivalent interpretations between the groups. That is, individuals’ observed scores on the scale are likely to measure the same latent variable and in a comparable way between the groups. Loosely speaking, one accessible interpretation of meeting measurement invariance is that individuals in both subgroups interpret the items in an equivalent manner. Not meeting measurement invariance has important implications for the researcher: it is not possible to meaningfully interpret comparison between the subgroups, nor associations between scores on the scale and external variables.”

**Reviewer 3**: On page 5 the authors discuss α as the sole source of structural validity evidence in the Flake review and that rigorous methodologies are rarely reported - it would be good to note here as well that this occurs for existing scales, but nearly half of the scales from the Flake review lacked any evidence because they were ad-hoc - so not just structural but any validity evidence-- substantive development, etc.

Though substantive stage validity evidence isn’t the focus of this paper, it is important to note that and discuss the importance of it, even for existing scales – and the authors have a scale in their study that, despite having been not developed in any formal sense, is still in use, so this issue is relevant to this study as well.

**Authors**: We have revised our material on p.5. as requested. Specifically, we now say the following:

“Indeed, Flake et al. (2017) found that the problem with validation was actually more severe that it initially appeared. Specifically, they not only found that research with well-known measures over-relied on Cronbach’s α as the sole test of structural validity, but that nearly half of the measures sampled were ad-hoc, and lacked evidence of validity testing at any of the three phases of validation.”

With respect to Reviewer 3’s point on substantive vs. structural validity, we agree. We now acknowledge this point while also highlighting another: that the goal of our paper was not to validate a given scale, but to examine whether under-reporting likely represents hidden invalidity (see changes on p.30-31).

“This includes attending to all three interrelated phases of validation (substantive, structural, external; Flake et al., 2017). Although we focused on the second phase, all phases of this process must be attended to when making a holistic evaluation about a measure’s validity. One phase (e.g., structural) is neither sufficient nor singularly important relative to the other two (e.g., substantive and external), nor should one strive to maximize it at the expense of the others.”

Substantive references are made to the importance of all three phases in multiple other places within the manuscript.

**Reviewer 3**: I was wondering what percent of scales had this existing evidence, specifically were the ones that failed the ones with very little evidence in the existing literature or were the results just not consistent with previous literature (e.g., the scale development paper reported a single factor model but you didn’t find that in your extension). It would be nice to report the frequency of the evidence evaluated here in the original development paper –this is discussed on page 10, but having a frequency table or a column that indicates if this analysis was conducted in the original development paper would be nice—I could see, development paper yes or no?

then yes or no for each column of evidence you conduct as a part of your study. I think this increases the contribution of the paper in two ways: first it makes clear the lack of thorough evaluation of scales in use to the extent that it exists, second it gives us an idea of how consistent results are from the original scale validation to a new, large, online same very explicitly. I recognize this could be a bit of a rabbit hole because some scales have many validation papers, but the authors could limit it to just the first introduction of the scale and make sure to mention that measurement invariance tests may be in subsequent papers, but you had to limit the scope of your review of the existing validation literature to keep the paper manageable.

**Authors**: We agree with Reviewer 3 that assessing for the number and frequency of structural validity tests during the initial development of a measure is a very interesting question. It would provide an interesting analysis of how much evidence for structural validity (or indeed construct validity) that scale developers are typically willing to live with before their measure is thrown out into the wild for use. Similarly, assessing how much evidence currently exists in the literature would provide an indication if a field takes the time to engage in ongoing validation.

However, we are reluctant to create and include this table in the current study, for several reasons. First, we feel that the proposed table would (at least in the context of this paper) serve to paint only a *partial* picture of the current state of affairs, and one that could be misleading or potentially deceptive. A table or column indicating how frequently tests of structural validity are carried out during initial measure development does not tell us anything about the current state of validation for those measures. It may be that a scale was initially introduced with no significant validity testing but has accumulated strong evidence for construct validity in the subsequent years. To only report the former and not the latter may be unfair: as Flake et al. (2017) note, validity evidence is highly cumulative in nature, and including such a column may not accurately depict how evidence for validity has accumulated. To answer such a question would require a separate systematic review for each of the scales we assessed, and undertaking that would be intensive in terms of time and effort (i.e., 15 separate systematic reviews).

Second, adding such a table would serve to introduce yet another question (is there a difference between validation during the development phase relative to use phase) that we do not address in this paper. It may give the impression that we are the first to actually carry out these types of validity tests (which is not the case for several of these scales), or that summarizing the total evidence for the validity of each of these scales was in fact our primary goal (which was not the case). As we mention in response to Reviewer 2, and a previous comment by Reviewer 3, our goal in this paper was test if under-reporting of structural validity in the literature (as identified by Flake et al.) is reflective of hidden validity or hidden invalidity. Our aim was to identify global trends across measures rather than to provide a definitive or final answer to the structural validity of any one measure.

**Reviewer 3**: Chi-square is useful to report even if it is significant because the other fit indices are based off of it, I would tone down the uselessness of chi-square insinuation

**Authors**: We thank Reviewer 3 for this comment. The text now reflects that Chi square values should always be reported (following Putnick & Bornstein, 2016), and that our earlier text intended to imply that the *p* values associated with these Chi square values are not particularly informative given our current sample sizes (i.e., not that the Chi square values are uninformative; see p.16).

“Chi square tests (although, given our sample sizes the *p* values for these are universally significant and therefore uninformative; nonetheless Chi square values should be reported”

**Reviewer 3**: I’m not very familiar with planned missingness designs, but tests of measurement invariance are influenced by the number of items and how strongly the items relate to the factor (factor overdetermination see Power and precision in confirmatory factor analytic tests of measurement invariance by Meade and Bauer 2007 in Structural Equation Modeling). I wonder how this is connected to the results shown in figure 1 for split scales and if this is a limitation of the current work, particularly for the measurement invariance testing

**Authors**: Although the design of the underlying study (the AIID study) followed a planned missing data design, data completeness was employed as an inclusion criterion for our analyses here. The nature of the AIID study was there were two broad subtypes of participants - those who provided complete data and those who provided very partial data. Exclusion of very partial data was therefore used as a mean to exclude low effort responding. All analyses in our manuscript were therefore run on complete data, and therefore should not impact or limit the interpretation of the tests of measurement invariance.

**Reviewer 3**: I also think it would be useful to explain the purpose of planned missingness designs in a sentence or two and reference some work (Todd Little has some publications on this) for those interested in learning more and aren’t familiar with these designs

**Authors**: The AIID study was specifically designed with reuse potential in mind, and indeed is currently being used as the basis for a series of registered reports by over 120 research teams that we are currently coordinating. In order to minimize redundancy between manuscripts, details on the study like those mentioned by Reviewer 3 can be found on the AIID project OSF page (https://osf.io/pcjwf/). A data curation manuscript containing details of the study along with the result of many data processing integrity tests will soon be submitted for review. Once a preprint of this is available, which should be very soon, we will cite it in the current manuscript to address Reviewer 3’s point.

**Reviewer 3**: Page 24—I think you mean the substantive phase.

**Authors**: Corrected.

**Reviewer 3**: On page 4 the authors quote “is the process of integrating…” this quote is in Flake et al., but was from Cronbach.

**Authors**: Corrected.

**Reviewer 3**: On page 4 the authors present the phases of validation—these are from Loevinger and she should be cited at this first mention.

**Authors**: Corrected.