# Changes between different versions of the manuscript

Results changed between the submission of the first and second version of the manuscript. Most notably, the rate at which the tests of measurement invariance were passed decreased from 67% to 4%. These results did not alter our overall conclusions – indeed, they suggest that the issue of hidden invalidity may be even worse than our original manuscript suggested. However, given the magnitude of this change in the reported results, we think it is important to document what changed between the first and second submissions.

## Change 1. Changes to the underlying dataset

Our manuscript makes use of the AIID dataset: a very large public dataset designed with reuse potential in mind. This dataset is curated as a “living dataset”: given the scale and complexity of the dataset, issues and errors will inevitably be detected. By versioning the changes to the processing and validation code and the resulting data, these changes to the dataset can be tracked. Whereas submission 1 employed a pre-release version of the AIID data, submission 2 employs the most recent version (v1.0.4). This resulted in a small reduction of the number of experimental sessions included in the dataset (<1%), which may have contributed to a small numerical change in the results reported in submission 2.

## Change 2. Removal of the “full scales” from the analyses

As noted in both versions of the manuscript, some questionnaires were split into two halves in the original data collection process (e.g., the BFI A/O vs. C/E/N scales). Submission 1 attempted to assess the structural validity of whole questionnaires by combining responses from individuals who completed both halves within one day across two experimental sessions. However, reviewers of submission 1 correctly pointed out that our analytic strategy treated the subscales as reflecting multiple correlated latent variables, which is typically not the case in the literature. That is, the Big 5 OCEAN scales are typically not used to assess one single “personality” latent variable, but 5 distinct personality facets. Upon reflection, we realized that, of the other questionnaires that possess multiple subscales, most are typically analyzed at the subscale level rather than as a set of correlated latent variables. It was therefore both simpler and more appropriate to analyze the questionnaires by each subscale. This strategy is employed in submission 2.

## Change 3. The implementation of tests of measurement invariance

Reviewers of submission 1 asked that we improve our definition of the three tests of measurement invariance. We did so with reference to a recent systematic review of the literature (Putnick & Bornstein, 2016). When doing so, we elected to revisit our implementation of these tests as Putnick and Bornstein define them. Additionally, we did this with reference to the implementation of the measurementInvariance function in the semTools R package, to see how experts in the field have elected to implement these tests.[[1]](#footnote-1) Upon inspecting the source code of this function, we observed differences between our original implementation of configural invariance in submission 1 compared to the semTools package. In our code for submission 1, we fitted two separate CFA models; one to the subset of one group’s data (e.g., men) and a second fit to the other group’s data (e.g., women). Two sets of model fit statistics were then compared against the cut-off criteria, where both had to pass for us to conclude that configural invariance was met. In contrast, the semTools package implemented the test of configural invariance by fitting a single CFA fit to the whole dataset, while passing a parameter to the group argument to define the subgroups within the data. One set of model fit statistics was therefore produced for the combined data, which was compared against cut-off values. In submission 2, we elected to bring our implementation in line with the semTools package, given that it is a popular package written by experts.

## Change 4. Code omission

In the process of revising the manuscript, we realized that the code for submission 1 omitted a number of items from two scales for all analyses (the NFCC A and O subscales). Specifically, the last item from each scale was accidentally omitted from the code that subset the data and model for these subscales. This was corrected in submission 2.

## Change 5. Code refactoring

In the process of implementing the above changes, we undertook a general refactoring of our code to (a) reduce repetition (and therefore any potential for mistakes) and (b) in doing so, increase comprehensibility. No errors were found in the submission 1 code during this process (other than those noted above), but nonetheless this change is worth noting for transparency, given that it could possibly have contributed to the change in the results in some unforeseen way.

1. Note that an earlier pre-submission version of the analyses employed the measurementInvariance function to test measurement invariance. However, poor convergence was found for many scales due to additional implementation choices made within the function. As such, both of our submissions construct the constrained measurement models manually in lavaan using the cfa function and arguments passed to the constraints parameter. [↑](#footnote-ref-1)