Dear Professor Simons,

Please find attached a manuscript titled “Hidden invalidity among fifteen commonly used measures in social and personality psychology”. Our article builds on Flake et al.’s (2017) work “Construct Validation in Social and Personality Research: Current Practice and Recommendations”, which found that articles published in JPSP routinely under-report evidence for the construct validity of the scales used. Flake et al. noted that it was not possible to know whether this issue was merely a problem of underreporting valid measures or, more worrying, a matter of invalid measures being used, unwittingly, due to insufficient consideration of their construct validity.

To address this issue we made use of large-scale dataset consisting of more than 151,000 experimental sessions. We apply a standardized set of best-practices analyses and success criteria to evaluate the structural validity of 15 widely used self-report measures from personality and social psychology along a comprehensive battery of metrics (internal consistency, immediate and delayed test-retest reliability, factor structure, and measurement invariance for median age and gender). Only 60% of the scales we assessed demonstrated good structural validity. We also found that the less commonly a test is reported in the literature (according to Flake et al., 2017), the more likely it was to fail our analyses. This supports the idea that the pattern of underreporting in the field reflects widespread hidden invalidity of the measures used. These results therefore pose a large-scale threat to many research findings in social and personality psychology, but also beyond in other areas of psychology, should they be representative of a broader trend. Specifically, as Flake et al. note, even where statistical effects are replicable, theoretical inferences made based on them can be incorrect if the measures they are based on are not structurally valid. Our manuscript therefore highlights important new information relevant to the replication crisis debate. Critically, our manuscript also highlights the degrees of freedom afforded to researchers in the assessment and reporting of structural validity, and suggests that the issue of validity hacking (*v*-hacking) should be acknowledged and addressed just as the concept of *p*-hacking is currently being tackled.

Due to the large number of structural validity tests we carried out, combined with the large number of scales these analyses were applied to, our manuscript (6632 words) is above the journal’s recommended 5000 word limit. You indicated by email that you would be willing to consider it at this length, but of course we would be willing to cut or shorten sections if you or the reviewers see potential areas where we could do so.

Kind Regards,

Ian Hussey & Sean Hughes