Appendix A - Pilot study

This document contains the methods and results of the pilot study of the *The Dire Disregard of Measurement Invariance Testing in Psychological Science* article.

All analyses were conducted in R (R Core Team, 2021). We used the following packages for data (pre)processing: $dplyr$ (Wickham et al., 2021), $foreign$ (R Core Team, 2022), $haven$ (Wickham & Miller, 2021), $readxl$ (Wickham & Bryan, 2019), and $reshape$ (Wickham, 2007). For the analyses we used the $lavaan$ (Rosseel, 2012), $psych$ (Revelle, 2022), $semPlot$ (Epskamp, 2022), and $semTools$ (Jorgensen et al., 2021) packages.

Deviations from preregistration

Our pilot preregistration contained cutoffs for CFI values to reject certain levels of measurement invariance that were reported as ΔCFI $>$ 0.01 and ΔCFI $>$ 0.02, which indicates we would reject invariance as the CFI estimate increased. The correct values for these cutoffs should be ΔCFI $<$ -0.01 and ΔCFI $<$ -0.02 as we rejected invariance when the CFI value decreased more than 0.01 and 0.02 between models, respectively.

We furthermore regarded a sample size per group or timepoint of *N* = 75 to be too small to perform measurement invariance analyses and planned to exclude these studies from the reproducibility step. However, since our final sample of studies was particularly small and we aimed to investigate the issues we could possibly run into when performing measurement invariance tests, we attempted to test for measurement invariance for all studies that made comparisons across groups or time on a scale and shared usable data.

Method

Reporting

We sampled articles with open data that were previously indexed from the journals Judgment and Decision Making (JDM), PLOS ONE (PLOS), and Psychological Science (PS). The index for JDM contained articles published between April 2011 and December 2014 (223 articles with 426 data sets), for PLOS articles published between 2013 and 2015 (252 articles with 530 data sets), and for PS articles published between 2014 and 2016 (144 articles with 324 data sets). We sampled 20 articles from each journal, for a total of 60 articles.

We coded how often the authors reported comparisons between groups or time on a scale for all articles. A scale had to measure a construct (i.e., we assume a latent trait underlies the scores on the indicators) by multiple indicators (at least three) or subscales. We required this scale is compared on a mean scale score (e.g., total score or sum score), either between groups or over time. Finally, we required the article to interpret the results of this scale comparison as a principal outcome of the study. For all studies that compared a scale across groups or time, we coded if the study reported a measurement invariance test, and if so, which level of measurement invariance held. Additionally, we documented if the scale was previously validated or modified by the authors, the level of measurement of the items, the number of items, the total sample size, and the sample size per group or time point.

Reproducibility

Next, for all studies that compared a scale across groups or time, we attempted to reproduce their reported measurement invariance test if they reported one, and we attempted to perform measurement invariance tests for those studies that compare scales across groups or time but did not report on measurement invariance.

For each study comparing groups or time points on a scale, we indicated whether we could locate and open the data the authors shared. Next, we indicated whether we could identify or construct scale and grouping variables from the data, after which we fit a configural, metric (i.e., loadings equivalent), and scalar (i.e., intercepts equivalent) model to the data. We reported which level of measurement invariance held and, if applicable, whether our result reproduced the measurement invariance result reported by the authors.

We rejected configural invariance if the chi-square test was statistically significant (α = .05), and at least one alternative fit measure was above a specific cutoff value (i.e., RMSEA $>$ .08 and CFI $<$ .95). If configural invariance held, we rejected metric invariance if the chi-square difference between the configural and metric model was statistically significant and at least one alternative fit measure between the configural and the metric step was above specific cutoff values (i.e., ΔRMSEA $>$ 0.03, ΔCFI $<$ -0.02). If metric invariance held, we rejected scalar invariance if the chi-square difference between the metric and scalar model was statistically significant, and at least one alternative fit measure between the metric and scalar step was above specific cutoff values (i.e., ΔRMSEA $>$ 0.01, ΔCFI $<$ -0.01).

Results

For each article, we indicated the number of studies, and within each study, we indicated the number of comparisons made between groups and over time. Five studies were ineligible because no (human) data was collected (e.g., simulation or animal studies). Next, 27 studies were ineligible because no groups were compared. Another 48 studies were excluded because no scale was measured. Finally, one study dropped out because a scale was measured, but no latent trait was assumed to underlie the item scores. This left 18 articles that contained 36 comparisons in which groups or time points were compared on a scale assuming an underlying latent variable. Of the 36 comparisons, four comparisons (across two articles) mentioned measurement invariance, of which one comparison included a test of measurement invariance. This one study reported that metric invariance held in their sample.

We attempted to reproduce the results for the comparison that performed a measurement invariance test and tried to perform a measurement invariance test for the other 35 comparisons that did not. We were able to successfully reproduce the result of the one study that performed a measurement invariance test. Even though we only sampled articles that indicated they shared their data, we could find data sets for only 30 out of 36 comparisons. Of those 30 data sets, 25 had usable and interpretable data (i.e., we could open the file, and the variables had names that were interpretable to us). For 22 comparisons we were able to construct groups based on the data and information provided in the article. We were able to identify the items and construct scales for 16 comparisons. In total, we were able to perform a measurement invariance test for 16 out of 36 comparisons. Of these 16 comparisons, one model did not converge, for four comparisons no level of measurement invariance held (i.e., non-configural), one comparison indicated configural invariance, four comparisons indicated metric invariance, and for six comparisons scalar invariance held.

Discussion

One-third of our sample of psychology articles contained scales used to make comparisons across groups or over time. A tiny proportion of this sample reported on measurement invariance, and of those studies, only one included a test for measurement invariance. In this case, the level of invariance that held was not sufficient to make valid mean comparisons between groups. From our pilot study we conclude that authors do not often assess or report on measurement invariance in psychological articles.

We were able to run measurement invariance tests for about half the comparisons, with the other half being ineligible due to inaccessible data, or data that did not contain enough information to run a measurement invariance test (e.g., individual item scores were omitted and only total scores were shared). For most of the comparisons we could test for measurement invariance, the results indicated some level of invariance held, but for only a small proportion of comparisons the level of measurement invariance was sufficient to attribute group or time differences to actual differences instead of measurement artifacts. We were able to reproduce the result of the only study that performed a measurement invariance test. It appears that it is generally possible to test for measurement invariance based on reported information in the manuscript and shared data. However, the data should be accessible, readable in non-propriety software, explicable (i.e., the meaning of the variables needs to be explicitly clear), and should minimally contain the item scores and grouping variable(s). Moreover, it is necessary for authors to indicate their data preprocessing steps, such as procedures around the exclusion of or missing data.

Changes from the pilot study to the main study

We sampled from three journals from our pilot study, but decided to only sample from PLOS ONE and Psychological Science for the main study. JDM did not contain many studies that compared scales across groups or time, and our aim for the main study was to sample from journals with somewhat more general psychological studies.

Because relatively many articles in the pilot study dropped out because they did not fit our first criterion (i.e., an empirical study where a scale is used to compare an outcome across groups or time) or because of inaccessible data, we decided to not compare open data articles with articles that do not share their data in the main study, but limit our sample to open data articles only. By only focusing on articles that share their data, we have access to a larger sample to investigate both measurement invariance reporting and reproducibility.

Our pilot study indicated that we would search the articles for principal outcomes compared between groups or over time. It was challenging to decide and agree on which tests or outcomes constituted principal outcomes in the articles. Therefore, for our main study, we decided to include all outcomes in an article (excluding supplemental material) that compared a scale over groups or time.

As our pilot sample was relatively small, we decided to run measurement invariance tests for all scales that were compared across groups and time. However, we noted that some studies may be underpowered and a measurement invariance test is not feasible. Therefore, in our main study, we will only run measurement invariance tests for those studies we deem to have enough statistical power to find measurement non-invariance.

We decided to modify the RMSEA and CFI cutoff criteria to reject metric measurement invariance in the main study. In the pilot study we adhered to cutoffs by (Rutkowski & Svetina, 2014), which are suitable for comparisons of more than 10 groups. For our main study, we will adhere to cutoffs by (Chen, 2007) and (Cheung & Rensvold, 2002), as they performed simulation studies with two groups, which is more applicable to our sample. We will furthermore supplement the RMSEA and CFI criteria with two information criteria (i.e., AIC and BIC) to assess measurement invariance, as the performance of fit measures and cutoffs varies across conditions, and using multiple fit measures may help us in assessing measurement invariance more accurately (Putnick & Bornstein, 2016)).

References

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/s15328007sem0902_5

Epskamp, S. (2022). *semPlot: Path diagrams and visual analysis of various SEM packages' output*. https://CRAN.R-project.org/package=semPlot

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2021). *semTools: Useful tools for structural equation modeling*. https://CRAN.R-project.org/package=semTools

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

R Core Team. (2022). *Foreign: Read data stored by 'minitab', 's', 'SAS', 'SPSS', 'stata', 'systat', 'weka', 'dBase', ...* https://CRAN.R-project.org/package=foreign

Revelle, W. (2022). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://www.jstatsoft.org/v48/i02/

Rutkowski, L., & Svetina, D. (2014). Assessing the Hypothesis of Measurement Invariance in the Context of Large-Scale International Surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. http://www.jstatsoft.org/v21/i12/

Wickham, H., & Bryan, J. (2019). *Readxl: Read excel files*. https://CRAN.R-project.org/pa ckage=readxl

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. https://CRAN.R-project.org/package=dplyr

Wickham, H., & Miller, E. (2021). *Haven: Import and export 'SPSS', 'stata' and 'SAS' files*. https://CRAN.R-project.org/package=haven