Supplemental Material A

Description of the 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, 2023). We used the following packages for data (pre)processing: *dplyr* (Wickham, François, et al., 2023), *foreign* (R Core Team, 2023), *haven* (Wickham, Miller, et al., 2023), *readxl* (Wickham & Bryan, 2023), and *reshape* (Wickham, 2007). For the analyses we used the *lavaan* (Rosseel, 2012), *psych* (Revelle, 2023), *semPlot* (Epskamp, 2022), and *semTools* (Jorgensen et al., 2022) packages.

Deviations from preregistration

Our pilot preregistration initially contained cutoffs for CFI values to reject certain levels of measurement invariance that were reported as $\Delta\text{CFI} > 0.01$ and $\Delta\text{CFI} > 0.02$, which indicates we would reject invariance as the CFI estimate increased. The correct values for these cutoffs should be $\Delta\text{CFI} < -0.01$ and $\Delta\text{CFI} < -0.02$. In other words, we would reject invariance when the CFI estimate 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 from three journals: Judgment and Decision Making (JDM), PLOS ONE (PLOS), and Psychological Science (PS). The sample included JDM articles published from April 2011 to December 2014 (223 articles with 426 data sets), PLOS articles published from 2013 to 2015 (252 articles with 530 data sets), and PS articles published from 2014

to 2016 (144 articles with 324 data sets). We randomly sampled 20 articles from each journal, resulting in a total of 60 articles.

We assessed each article to determine if the authors reported comparisons between groups or time on a scale. To be included, 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. The groups or time points had to be compared on a mean scale score (e.g., total score or sum score). Additionally, the article needed to interpret the results of this scale comparison as the primary outcome of the study. For all studies that met these criteria, we coded whether the study reported a measurement invariance test, and if so, which level of measurement invariance held. We also documented whether the scale had been 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

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

For each study comparing groups or time points on a scale, we first determined whether we could locate and open the data shared by the authors. Next, we attempted to identify or construct scale and grouping variables from the data. We then fit a configural, metric (i.e., loadings equivalent), and scalar (i.e., intercepts equivalent) model to the data and indicated which level of measurement invariance held. If the study reported a measurement invariance test, we also noted whether our results were consistent with those reported by the authors.

We rejected configural invariance if the χ^2 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 χ^2 difference between the configural and metric model was statistically significant, and at least one alternative fit measure between the configural and 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 time points. Out of the total number of studies, five were excluded because they did not collect any human data (e.g., simulation or animal studies), whereas 27 studies were ineligible because they did not compare any groups. Additionally, 48 studies were removed because they did not measure any scale. Lastly, one study was dropped because it measured a scale but did not assume any latent trait underlying the item scores. Consequently, 18 articles were left, containing 36 comparisons in which groups or time points were compared on a scale assuming an underlying latent variable.

Out of the 36 comparisons, only four comparisons (across two articles) mentioned measurement invariance, and one comparison included a test of measurement invariance. This study reported that metric invariance held in their sample. We attempted to reproduce the results for the comparison that tested measurement invariance and also attempted to test measurement invariance for the other 35 comparisons. We successfully reproduced the results of the one study that tested measurement invariance.

Despite selecting articles that indicated data sharing, we could only find data sets for 30 out of 36 comparisons. Out of these 30 data sets, 25 had usable and interpretable data, allowing us to open the file, and the variables had names that we could interpret. We could construct groups for 22 comparisons based on the data and information provided in the articles, and identify the items and construct scales for 16 comparisons.

Overall, we were able to perform a measurement invariance test for 16 out of 36 comparisons. Out of these 16 comparisons, one model did not converge, for four no level of measurement invariance held (i.e., configural non-invariance), one comparison indicated configural invariance, four comparisons indicated metric invariance, and for six comparisons scalar invariance held.

Discussion

We found that one third of the articles in our sample of psychology articles used scales to compare groups or time points. However, only a small fraction of these articles reported on measurement invariance, and out of those articles, only one included a test for measurement invariance. In this case, the level of invariance was insufficient for valid mean comparisons between groups. Based on our pilot study, we concluded that authors do not often assess or report on measurement invariance in psychological articles.

We were able to conduct measurement invariance tests for about half of the comparisons, with the other half being ineligible due to inaccessible or incomplete data. For the comparisons that we could test for measurement invariance, most indicated support for some level of invariance, but only a small proportion of comparisons showed sufficient levels of measurement invariance to attribute group or time differences to actual differences rather than measurement artifacts (i.e., scalar invariance). We were able to reproduce the results of the only study that performed a measurement invariance test.

We found 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. This is because the JDM journal did not have many studies that compared scales across groups or time, and we aimed to include journals with more general psychological studies.

We encountered a significant number of articles that did not meet our first criterion (i.e., an empirical study comparing outcomes across groups or time) or had inaccessible data, resulting in their exclusion from the pilot study. Thus, for the main study, we decided to limit our sample to only open data articles, rather than comparing them with those that do not share their data. By

only focusing on articles that share their data, we have access to a larger sample to investigate both measurement invariance reporting and reproducibility.

In our pilot study, we attempted to identify principal outcomes that compared scales between groups or over time. However, 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.

Because our pilot sample was small, we tested for measurement invariance across all scales compared between 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 (2013), 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 also supplement the RMSEA and CFI criteria with two information criteria (AIC and BIC) to more accurately assess measurement invariance, given the varying performance of fit measures and cutoffs across conditions (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. (2022). *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. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Revelle, W. (2023). *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://doi.org/10.18637/jss.v048.i02

Rutkowski, L., & Svetina, D. (2013). 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). https://www.jstatsoft.org/v21/i12/

Wickham, H., & Bryan, J. (2023). *Readxl: Read excel files*. https://CRAN.R-project.org/package=readxl

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *Dplyr: A grammar of*

data manipulation. https://CRAN.R-project.org/package=dplyr

Wickham, H., Miller, E., & Smith, D. (2023). *Haven: Import and export 'SPSS', 'stata' and 'SAS' files*. https://CRAN.R-project.org/package=haven