Preregistration Pilot study (anonymized)

Measurement Invariance Testing in Psychology: A Systematic Review

Description

Note: this is the preregistration for the pilot study, in which we take a random sample from articles within psychology that shared their data and code these articles to answer the research questions below. Our aim with the pilot study is to investigate what kind of scales researchers use in the psychological literature and how they report on the measurement of those scales. Moreover, for those articles that do use a scale and compare it between groups or over time, we would like to know if we could conduct (if the original authors did not report it) or reproduce (if the original authors reported it) a measurement invariance check of that scale based on the data the original authors shared and the information they reported in the article or appendices.

We also would like to investigate whether there are procedures or variables in our preregistration, codebook and flowchart that might not be realistic based on the articles we have. Based on this pilot study we will make additions to our current preregistration, and explicate which information we will no longer collect.

Research questions:

To what extent do empirical articles in psychology use scales to make comparisons across groups or over time?

To what extent is measurement invariance assessed in psychology?

To what extent does measurement invariance hold in psychology?

For those studies in psychology that share their data, is it possible to assess measurement invariance based on the information reported and data shared?

Sub questions:

- Do studies in the psychology literature that use multiple indicators to measure an underlying scale between groups or over time report on measurement invariance?

- How do these studies report on measurement invariance?

- How often is measurement (non-)invariance found in these studies? For those studies that share their data and fall within our inclusion criteria, meaning a construct is measured and compared between groups or over time:

- Does the data allow us to check for measurement invariance (e.g., clear reporting on variables, data in a readable format, do we need materials or syntax as well)?

- Do the reported and reproduced results on measurement invariance coincide?

- How often is measurement (non-)invariance found in studies that share their data?

- How often are studies underpowered (e.g., small sample size), making MI checks hard or impossible?

Hypotheses

We expect for those studies that measure a scale with multiple indicators between groups or over time and for which we reasonably can assume a measurement model, many do not check for measurement invariance of the scale, but we do not have a specific prediction on the extent to which studies (fail to) check measurement invariance. We expect the reporting on measurement invariance is inconsistent across studies (Vandenberg & Lance, 2000). We expect that a reasonable portion of the shared data has insufficient reporting that will hinder us in checking for measurement invariance. More specifically, we expect the data to be insufficiently documented regarding indicators that are included in the measure, the handling of missing data, and syntax (instructions) to analyze the data. We expect reporting quality to be higher in studies that share their data compared to studies that do not share their data.

Design Plan

Study type

Meta-Analysis - A systematic review of published studies.

Blinding

No blinding is involved in this study.

Is there any additional blinding in this study?

NA

Study design

NA

No files selected

Randomization

No randomization

Sampling Plan

Existing Data

Registration prior to accessing the data

Explanation of existing data

For our open dataset pilot sample, we will use articles from the journals Judgment and Decision Making (JDM), PLoS ONE (PLOS) and Psychological Science (PS) that have been indexed previously. Note that our unit of analysis is the number of articles (k = 619), not the datasets. The collection consists of information on three journals: JDM: all articles published between april 2011 and december 2014. In total 223 articles with 526 datasets. PO: articles published between 2013 and 2015. In total 252 articles with 537 datasets PS: articles published between 2014 and 2016. In total 144 articles with 363 datasets. For our pilot study, we only use articles that shared their data. For the main study, we would like to compare these articles to articles within psychology that do not share their data. We will update our preregistration accordingly. Neither of the authors have studied the collection of articles sharing open data before the start of this pilot.

Data collection procedures

We only use existing data (articles) for this study.

No files selected

Sample size

The total number of articles that share their data from the three journals is 619. For our pilot study, we will take a sample of k = 60 from the three journals, with k = 20 from JDM, k = 20 from PO, and k = 20 from PS. We aim for a newer sample (2018, 2019) from these new journals for the main study.

Sample size rationale

We expect the coding of the articles to be time-consuming, and investigating the data and trying to reproduce results or run analyses to be even more time-consuming. However, we also expect many articles to drop out of our initial selection because they do not measure a scale across groups or over time. We have a stopping rule of one hour, and believe that coding k = 60 articles is doable in a relatively short time, but also gives us enough information on what to expect from other articles.

Stopping rule

Our stopping rule for coding individual studies will be one hour. This entails that we want to be able to code all variables and (re)run a measurement invariance analysis for one study within an hour. We believe reporting should be transparent and thorough enough to enable the investigation of measurement invariance of a construct in a study within an hour. However, we expect that this will sometimes be impossible because documentation regarding the steps followed in analyses is insufficient, and the data are not transparent enough to ensure reproducibility. If this is the case, we will document this in the codebook.

Variables

Manipulated variables

Not applicable.

- pilot-study.R

Measured variables

Included in this question is a codebook with all measured variables per study, and basic information on the sampled articles. Since our study is descriptive, main outcomes will entail (but are not limited to): The percentage of studies that use a scale to compare across groups or over time The percentage of studies that report on measurement invariance The percentage of studies that report on finding measurement non-invariance The percentage of studies for which we can reproduce the measurement invariance check The percentage of studies for which our reproduced result on measurement invariance coincides with the reported result.

- pilot-codebook.xlsx

Indices

Not applicable

No files selected

Analysis Plan

Statistical models

Our pilot is a descriptive study and will not require any statistical analyses. However, we will use statistical analyses to reproduce results of articles sharing their data. First, we will combine the indicators that measure the scale or construct based on the documentation and (possible) shared syntax, as well as collect the participant scores either in groups or in a group over time based on the documentation and (possible) shared syntax. Second, we will test this construct on measurement invariance between groups or over time. For this second step, we will (among possible others), use the lavaan package (version 0.6-5) in R (R Core Team, 2018; Rosseel, 2019) to perform measurement invariance in steps and to check for model fit.

No files selected

Transformations

We will transform existing data based on what is reported in the article or additional documentation (e.g., combining items in a construct, deleting participants, missing data). If

this documentation is unavailable or incomplete, we will make inferences based on the information that is available and will document accordingly. Disagreements will be solved by discussion, if needed by a third person. We do not expect any other data that have to be transformed.

Inference criteria

The criteria used to reject the fit of models that test measurement invariance are:

Configural:

- significant chi-square test (alpha = .05);

- at least one alternative fit measure (RMSEA, CFI) above these cut-off values (RMSEA > .08, CFI < .95)

Metric:

- significant chi-square test (alpha = .05);

- At least one difference in alternative fit measures (RMSEA, CFI) between the configural and the metric step above these cut-off values (ΔRMSEA > 0.03, ΔCFI > 0.02)

Scalar:

- significant chi-square test (alpha = .05);

- At least one difference in alternative fit measures (RMSEA, CFI) between the metric and scalar step above these cut-off values (ΔRMSEA > 0.01, ΔCFI > 0.01)

Data exclusion

For any reporting on measurement invariance or analyzing data, we will exclude all articles: - for which the principal outcome(s) do not measure a scale that do not compare the principal outcome(s) between groups or over time - for which the sample size per group or timepoint is N < 75, since this sample size is too small to perform measurement invariance analyses. We define a scale by a construct that is measured by multiple indicators or subscales. We require that the article tests this scale on a scale score (e.g., sum score), either between groups or over time. Additionally, we require the article to interpret the results of this scale as a relevant outcome within the study. For both article sets, we will report on measurement invariance and attempt to analyze data for all articles for which:

- An empirical study is conducted

- The article mentions the principal outcome(s) is a scale - The scale is measured by multiple indicators (at least 3)

- A total score or sum score is calculated for the scale

- The scale can reasonably be assumed to have a measurement model; a latent trait is assumed to underlie the scores on the indicators. An example: the IQ index is measured, and the authors interpret it in terms of intelligence.

- A comparison is made between either new (e.g. control and treatment) or existing (e.g. American and European) groups, or a comparison is made in one group at two timepoints

Missing data

If we deal with incomplete or missing data within the studies, we will use the reporting in the main article as well as additional documentation to code all variables as best as possible. If needed, we will make inferences and document these accordingly. Disagreements will be solved by discussion or if needed by a third person. We will cease the coding of articles for which the time to code all variables and to (re)run measurement invariance analyses exceeds an hour. We do not expect to have to exclude any data for our final analyses after coding.

Exploratory analysis

Not applicable

Other

Other

No response