Preregistration Main study (anonymized)

Measurement Invariance Testing in Psychology: A Systematic Review

Hypotheses

We have specific research questions related to the reporting in articles, and specific research questions related to the reproducibility of articles.

Main research question:

- To what extent is measurement invariance assessed in articles that use psychological scales?

Research questions (reporting):

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

- If scales are used to make comparisons across groups or over time:

- To what extent is measurement invariance of scales assessed? - To what extent is measurement invariance of scales found?

Research questions (reproducibility and data availability):

If articles share their data:

- For those studies that check for measurement invariance, can we reproduce the results based on the information reported and data shared?

- For those studies that do not check for measurement invariance (but should), is it possible to assess measurement invariance based on the information reported in the article and data shared online?

HYPOTHESES

We have not formulated specific hypotheses that lend themselves to hypothesis testing. Next, we formulate general expectations. In the exploratory section we have formulated for which outcomes we may perform additional exploratory hypothesis tests. 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, 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 information we need to assess whether measurement invariance is performed or whether we are able to (re)run analyses is insufficiently reported on, which will hinder us in checking for measurement invariance. More specifically, we expect insufficient documentation on:

- indicators that are included in the measure;

- which subjects are included in which groups;

- the handling of missing data;

- transformation of data;

- availability of item scores (i.e., only sum scores on measures are shared);

- syntax (instructions) to analyze the data.

Design Plan

Study type

Observational Study - Data is collected from study subjects that are not randomly assigned to a treatment. This includes surveys, “natural experiments,” and regression discontinuity designs.

Blinding

No blinding is involved in this study.

Is there any additional blinding in this study?

No response

Study design

Not applicable. More information will be added in the next questions.

No files selected

Randomization

Not applicable.

Sampling Plan

Existing Data

Registration prior to accessing the data

Explanation of existing data

Our preregistration for the pilot states we would like to analyze articles that do not share their data and those that do share their data. After conducting the pilot study we decided to only include articles that shared their data for our main study, because we found that relatively many articles drop out because they do not fit our first criterion (i.e., an empirical study where a scale is used to compare an outcome across groups or time). By only focusing on articles that share their data, we have access to a larger sample to extract relevant information from, on both measurement invariance reporting and reproducibility. For our sample, we will use articles from the journals PLoS ONE (PO) and Psychological Science (PS).

In our pilot study we also included articles from the Journal of Judgment and Decision Making (JDM). We decided to focus only on the other two journals for our main study because JDM is not a general psychological science journal and the number of articles that fit our criterion of an empirical article using a scale with multiple indicators between groups was relatively small.

Note that our unit of analysis in our main study is the number of scales we find, not the number of articles or the datasets. However, we do not know yet how many scales are used within each article.

The collection of articles consists of: PS: all articles published in 2018 and 2019 (Ntotal = 368) that share their data online (N = 213; N2018 = 114, N2019 = 99).

To get an equal sample for both journals, we will randomly sample 213 articles from PO from the same years (weighted based on PS, i.e., N2018 = 114, N2019 = 99).

None of the authors have studied the collection of articles sharing open data before the start of this study. The index of articles in PS was made by different researchers, who will not analyze the data. The random sample of articles in PO will be drawn by the main authors after preregistration. The code we use to sample the PO articles is added to the next question of this preregistration (articleindex.R).

Data collection procedures

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

- articleindex.R
- sample-ps.txt

Sample size

The total number of articles from the two journals is 426 (213 Psychological Science, 213 PLOSONE).

Sample size rationale

We expect investigating the data and trying to reproduce results or run analyses to be considerably more time-consuming than the coding on the reporting of the articles. 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 did not want to exceed 500 articles due to time constraints. Because we did want to have a large and relatively representative sample, we chose the two most recent years we could find data on, 2018 and 2019. The number of articles containing open datasets for PS for 2018 and 2019 combined was 213, so we decided to match this number with articles of PO, making a total of 426 articles.

Stopping rule

Because investigating the data and trying to reproduce results often was not a clear-cut process and required much discussion in the pilot study, we decided for the main study to divide the coding process into two steps. First, each coder (first authors) separately codes all articles regarding their reporting on scales and measurement invariance (i.e., the reporting step). Disagreements will be solved by a third person. Second, we will go through all articles again and investigate and (re)run the data together (i.e., the reproducibility step). For the reporting step, based on what was observed during the pilot study we decided not to have a stopping rule. Most articles could be coded within a reasonable timeframe. For the reproducibility step, our stopping rule will be one hour. This entails that we want to be able to code all variables in our codebook and (re)run a measurement invariance analysis for one article 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 information is not transparent enough to ensure reproducibility. If this is the case, we will document this in the codebook.

Variables

Manipulated variables

Not applicable.

No files selected

Measured variables

Included in this question is a codebook with all measured variables per article, study and scale. We also included our coding protocol.

In our pilot study we decided on only including the main outcome from a study, instead of all the comparisons that were made between groups or time on outcomes. For the main study we decided to include any outcome on a scale that is compared between groups or time and reported in the manuscript (i.e., not in the supplementary material), since it was relatively hard to decide during the pilot study what exactly constituted a main outcome.

We define a scale by a construct that is measured by multiple indicators or subscales, and for which a total score or sum score is calculated for the scale. We further require the scale to be amenable to psychometric analysis (e.g., reliability, FA, IRT). We require that the article tests scores on this scale either between groups or over time. This also includes comparisons made in which a reflective factor was not necessarily assumed (i.e., t-tests, (M)ANOVA), but does require the authors to interpret the scale as a reflective factor.

We will use the reporting in the article as the guideline. For example, if we encounter situations where a Principal Component Analysis is used to analyze the data in an article, but the authors report on the scale as it being a reflective scale, we will interpret this scale as reflective. We will make a note in our codebook when this occurs.

Additionally, if the authors report running an analysis on a certain sample size, but the degrees of freedom of the test indicate another sample size, we will use the reported sample size as a guideline, and base our reanalyses on the reported statistics. We do so because we think written text gets read and interpreted more (by the authors, peer-reviewers, and readers) than statistics results or results in tables, making it more likely that the text is in line with what the authors intended to convey. We will make a note in our codebook when the reported sample size and reported degrees of freedom differ. Additionally, if we find deviating results in text compared to tables, we will use the information reported in text as the guideline. We will make a note in our codebook when this occurs.

Generally, we will use the information as reported in the article as a guideline over any results we find in the data. For example, if the article describes a scale with three measures, but we obtain a dataset with a scale with five measures, we will use the three measures for analyses. We will make a note in our codebook when this occurs.

Our study makes use of only descriptive statistics. Outcomes will entail (but are not limited to):

- The proportion of articles/studies that use a scale to compare across groups or over time

- The proportion of articles/studies/scales that use an already existing scale (i.e., validated in a previous study or in this study) or a new scale (i.e., a scale that has been altered or constructed by the authors without validation).

- The proportion of articles/studies that report on measurement (non)invariance

- The proportion of articles/studies/scales for which we can reproduce the measurement invariance check

- The proportion of articles/studies/scales for which our reproduced result on measurement invariance coincides with the reported result.

- The proportion of articles/studies/scales for which we can conduct a measurement invariance check

- The level of measurement invariance most often reported in articles that use scales.

- Reasons for irreproducibility or inability to run a measurement invariance check.

- main-codebook.xlsx
- variables-description.html
- MI-Flowchart 28-07-20.pdf

Indices

Not applicable.

No files selected

Analysis Plan

Statistical models

Our main study is a descriptive study and will not require any statistical analyses. We may decide on performing on additional exploratory statistical analyses if we obtain enough articles to have an adequately powered study. This procedure is reported in the exploratory analyses section below.

We will use statistical analyses to (re)produce 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 for measurement invariance. Based on the pilot study we do expect most scales to be composed of ordinal (3 to 5 categories) or continuous (more than 5 categories) items, for which measurement invariance can be tested using multiple group categorical confirmatory factor analysis (MG-CCFA), for ordinal items, and multiple group confirmatory factor analysis (MG-CFA), for continuous items. However, if we will encounter scales composed of dichotomous items we will use a multiple group item response theory (MG-IRT)-based approach to test for MI. For this second step, we will (among possible others), use the lavaan package (version 0.6-6) in R (R Core Team, 2018; Rosseel, 2019) to perform measurement invariance in steps and to check for model fit.

We will perform a measurement invariance check on the scales in the article, between the groups or times under investigation. We will use descriptives to report our outcomes. We do not expect to run any follow-up analyses.

We have added the code that we will use to perform these measurement invariance checks to this question. Note that often we need to recode the data in a way that is specific to that dataset. As such, this code is only part of the code we will need to (re)produce results.

We have also added the analysis script of the pilot study that we will use as a framework for the main study.

- mi-analysis.R

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 needs to be transformed.

Inference criteria

In our pilot study, articles made comparisons ranging from between two to between five groups, with most comparing two groups. When fewer than 10 groups are compared, we will use the following criteria to make inferences about model fit and level of measurement invariance (based on Cheung & Rensvold, 2002 and Chen, 2007, who both use simulations with two groups):

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

Configural:

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

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

Metric:

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

- at least two out of these four alternative fit measures (AFI) between the configural and the metric step are either above these cut-off values (ΔRMSEA > 0.01, ΔCFI > 0.01) or have lower estimates than in the configural step (AIC, BIC).

Scalar:

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

- at least two out of these four alternative fit measures (AFI) between the configural and the metric step are either above these cut-off values (ΔRMSEA > 0.01, ΔCFI > 0.01) or have lower estimates than in the metric step (AIC, BIC).

Note that these guidelines for the change in RMSEA and CFI (Chen, 2007) are based on situations with an adequate total sample size (N > 300). The guidelines for cutoffs for studies with smaller sample sizes are more stringent. However, since the cutoffs mentioned before are also in line with general practices and are more lenient, we adhere to these in our main study.

If more than 10 groups are compared we will use the following criteria, based on Rutkowski & Svetina, 2013):

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 > .10, CFI < .95)

Metric:

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

- at least two out of these four alternative fit measures (AFI) between the configural and the metric step are either above these cut-off values (ΔRMSEA > 0.03, ΔCFI > 0.02) or have lower estimates than in the configural step (AIC, BIC).

Scalar:

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

- at least two out of these four alternative fit measures (AFI) between the configural and the metric step are either above these cut-off values (ΔRMSEA > 0.01, ΔCFI > 0.01) or have lower estimates than in the configural step (AIC, BIC).

If we perform a MG-IRT analysis, we will test for measurement invariance at the item-level. We will perform a Likelihood Ratio Test (LRT) per item in the scale to check for measurement non-invariance on the metric, scalar and residual level. This is a df = 1 test. The chi-square test will be rejected after a statistically significant likelihood ratio test (LRT) (alpha = .05, chi-square > 3.64). Since multiple tests are conducted (i.e., one for each item), a Bonferroni correction will be used (0.05 divided by the number of tests). In order to keep the coding coherent with the MG-(C)CFA analysis, scalar, metric or residual invariance will be rejected if at least one item is non-invariant.

Data exclusion

For all articles and studies within articles, we will code whether a (reflective) scale is used and whether group comparisons are made. If no scale is used or no group/time comparisons are made, we will exclude this article or study within an article from all further coding.

For the reporting step, we will only include articles that use a scale and compare groups or time (see previous paragraph). We will not exclude any articles or studies within articles while coding this step.

For the reproducibility and data analysis step, there might be reasons to exclude certain scale comparisons while coding. One of them is data availability: it might be the case that the data is unretrievable, the datafile does not open, or the data is not interpretable. If this is the case, we will cease coding and will not perform a measurement invariance check.

A second reason is insufficient power. Power to detect measurement invariance depends on the sample size, correlations between indicators and the amount of invariance between groups. Since we do not know the amount of invariance that will be found. We will decide on the minimum sample size to run a measurement invariance analysis after the reporting step of the coding. Specifically, we will analyze all sample sizes in the article to get an idea of the lower bound of sample sizes we encountered, and what kind of models are used.

We will then run a simulation study with this type of model and the lowest sample sizes, while under different levels of loading and intercept non-invariance between groups, and when correlations between variables are of medium size (r = .40-.50). Our aim is to investigate what the statistical power would be to find measurement noninvariance given these characteristics. Afterwards, we will decide on a minimum cutoff for sample size. This entails we will only run measurement invariance analyses on samples that exceed the sample size related to a statistical power estimate of 0.70. We will note in our manuscript that power to detect invariance in situations with a relatively small sample size is small and should be interpreted with caution.

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 (i.e., supplementary materials) 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 (re)run measurement invariance analyses exceeds an hour.

If no link can be found to the data accompanying the article, we will search for the data by using the author names and the article title and searching on google and Web of Science. We will not mail the authors for their data. If we cannot find a link to the data, we will only include this article in the reporting step, and skip the coding on the reproducibility section.

Exploratory analysis

In our main study we make use of outcomes interpreted in proportions, included (but not limited to) the following:

- The proportion of articles/studies that use a scale to compare across groups or over time

- The proportion of articles/studies/scales that use an already existing scale (i.e., validated in a previous study or in this study) or a new scale (i.e., a scale that has been altered or constructed by the authors without validation).

- The proportion of articles/studies that report on measurement (non)invariance

- The proportion of articles/studies/scales for which we can reproduce the measurement invariance check

- The proportion of articles/studies/scales for which our reproduced result on measurement invariance coincides with the reported result.

- The proportion of articles/studies/scales for which we can conduct a measurement invariance check

- The proportion of articles/studies/scales that report on measurement invariance and find measurement invariance compared to the proportion of articles/studies/scales that do not report on measurement invariance and find measurement non-invariance

It would be possible for us to perform one-sample (or two-sample) proportion hypothesis tests to determine whether there is a statistically significant difference in our sampled proportion outcome compared to a population proportion or compared to our pilot study results. However, because we do not know how many articles, studies, or scales (depending on the unit of analysis) can be used, these analyses may be underpowered. We may decide on also doing hypothesis testing on our descriptive outcomes, and will report these as exploratory in the article.

Other

Other

No response