The title

First Author1 & Ernst-August Doelle1,2

1 Wilhelm-Wundt-University

2 Konstanz Business School

Author note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line. Enter author note here.

The authors made the following contributions. First Author: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Ernst-August Doelle: Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to First Author, Postal address. E-mail: [my@email.com](mailto:my@email.com)

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline. Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines. One sentence clearly stating the **general problem** being addressed by this particular study. One sentence summarizing the main result (with the words “**here we show**” or their equivalent). Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge. One or two sentences to put the results into a more **general context**. Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

*Keywords:* keywords

*Word count:* X

The title

Psychological research involves the difficult task of assessing non-observable phenomena, such as depression or meaning in life, as a measurement proxy for testing hypotheses. Researchers develop surveys or instruments to estimate underlying constructs of interest (DeVellis & Thorpe, 2022), with item relationships often validated with latent variable modeling (i.e., structural equation modeling [SEM]) or item response theory (IRT, Byrne, 2001). Entire journals, such as *Assessment,* are devoted to the publication of scale development and (re)-assessment across populations - a necessary avenue given that development information for many scales is not reported in other journal articles (Barry et al., 2014; Weidman et al., 2017). Scale development manuscripts are crucial to interpretation of studies that use these measures and determining the usefulness of measured scores (Flake & Fried, 2020).

Confirmatory factor analysis (CFA) is a common analysis that examines the relationship between observed variables and latent constructions, helping determine the overall model fit to the data collected (**kline2015?**). Researchers specify the measurement model that demonstrates the proposed relation between measured items and latent variables and use fit indices such as the Tucker Lewis Index (TLI, **tucker1973?**), Comparative Fit Index (CFI, **bentler1990?**), Root Mean Square Error of Approximation (RMSEA, **steiger1990?**), and standardized root mean residual (SRMR, **bentler1995?**) to determine model fit. Generalized measures are designed, in theory, to provide the same assessment for different populations, but researchers can extend their investigation of item and latent variable properties by examining scale functioning across groups or subpopulations within the data. Multigroup CFA (MGCFA, Meredith, 1993) examines different components of a CFA model for measurement invariance. Measurement invariance (or equivalence) implies that the instrument provides the same latent variable measurement for all populations. Equivalence in measurement is desirable, as it allows for the same interpretation of latent variable scores across groups, as well as knowing that group differences are not attributable to differences the way the measure operates within groups.

Measurement invariance is tested by sequentially examining models with group equality constraints. There are several versions of the procedure, but the terminology and steps from (**brown2005?**) are commonly found within studies on psychological constructs. First, groups are combined together within a *configural* model to assess for the same factor structure across groups. The factor loadings for each group are then set to be equal for the assessment of *metric* invariance, implying that each group has the same strength of item relation to the latent variable. *Scalar* invariance is this examined by constraining the item intercepts across groups, which implies that each group shows the same item means. Last, *strict* invariance may be tested by constraining the item residuals to be equal across groups, suggesting that each group shows the same variance across items. Each step is compared to the previous step to determine if the equality constraints result in a degradation of fit, often by using CFI or RMSEA (**cheung2002?**). Scales that demonstrate measurement invariance allow researchers to use scores to make group comparisons without adjustments for measurement differences (**millsap2011?**).

Non-invariance implies that individuals in separate populations interpret items differently (Cheung & Rensvold, 2000; Dong & Dumas, 2020; Liu et al., 2017; Wicherts et al., 2005), which may affect the overall latent variable score. Thus, it can be difficult to know if group differences are due to population differences or measurement. Being unaware of non-invariance in measures could lead to incorrect interpretations of group differences (Van De Schoot et al., 2015), and these results could potentially explain the replication or lack-thereof for results across studies (Maassen et al., 2023). For example, (**trent2013?**) examined the Revised Child Anxiety and Depression Scale (RCADS, **chorpita2000?**), an instrument that examines depression and anxiety disorders in children. They determined that the six-factor structure was appropriate for White and Black youth (*configural*) and that the factor loadings were also equal across groups (*metric*). They found that children differed in their item intercepts (*scalar*) for five items across the scale, which suggested that the same latent construct would represent different scale scores on those items. These results often suggest partial invariance, wherein the scale often generalizes minus a few items (**cheung1999?**).

Partial invariance is critical for valid interpretations because without it, researchers risk conflating true group differences with artifacts of measurement bias. Failure to identify and address non-invariance can lead to inaccurate conclusions about group-level comparisons and potentially undermine the validity of psychological research. Despite its importance, researchers frequently overlook measurement invariance testing, as highlighted by Maassen et al. (2023). Their study examined the reporting of measurement invariance tests across *Judgment and Decision Making*, *PLoS One*, and *Psychological Science* and found that 1) very few papers included measurement invariance tests; 2) none of those reported tests could be reproduced; and 3) very little measurement invariance was found across when new tests could be examined.

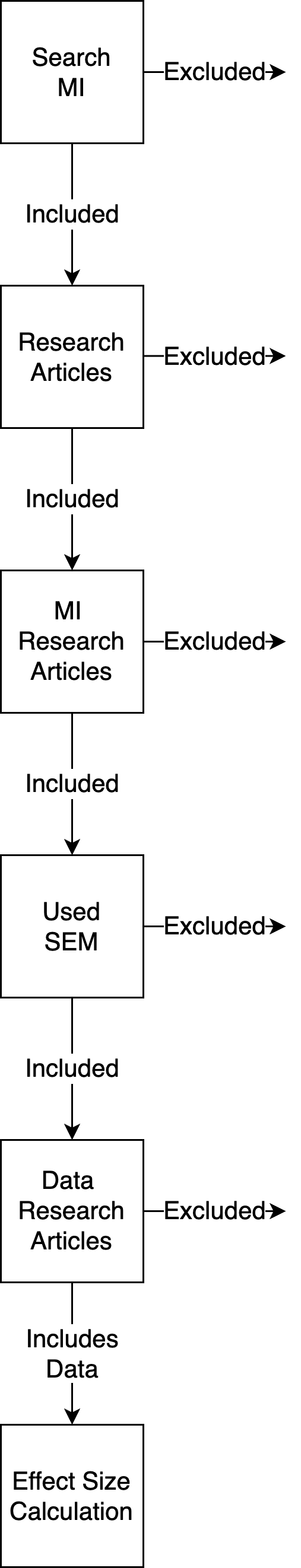
This study demonstrates the need for an investigation of measurement invariance within a journal that specifically targets assessments as the area of publication, such as *Assessment*. Given differences in cultural, experience, language skills, and more, we may not expect all measurements to show invariance across populations. Partial invariance extends multigroup testing of measurement invariance to show exactly where and how many parameters are non-invariant (Byrne et al., 1989; Meredith, 1993). Understanding these items can lead to further investigation into group differences, new interpretation guidelines, or scale improvement to eliminate identified differences. Measurement invariance testing suffers from the same black-and-white judgment criteria found in traditional null hypothesis testing and *p*-values with cutoff criteria and rules of thumb (Marsh et al., 2004; Putnick & Bornstein, 2016). was developed as an effect size for measurement invariance for group differences in observed variables, which is affected by both factor loadings and item intercepts (Nye & Drasgow, 2011).

visualizemi is a new *R* package that calculates the replication rate of the overall model steps in measurement invariance (as compared to randomized data), as well as the effect sizes for individual parameters, separating loadings, intercepts, residuals, and so on (Buchanan, 2024). This package can be used to determine the overall model level replication effect sizes, individual parameter replication effect sizes, and effect sizes for non-invariance for each parameter separately (i.e., traditional Cohen’s based on parameter size differences). The replication effect sizes are calculated by creating bootstrapped samples from the data and comparing their model results to the same data with randomized group labels. The implication is that randomizing group labels should not affect the results if groups were truly equal in their measurement. The effect size is calculated as a version of Cohen’s which is the standardized difference in proportions between the bootstrapped results and randomized results. As the effect size increases, we would expect to find non-invariance between groups in a new sample, as the bootstrapped data shows consistent differences between groups.

At the moment, it is difficult to know what effect sizes one may expect to find for measurement invariance and what may be a level of measurement invariance to worry about (i.e., moving away black and white decision making on invariance). Researchers may be able to define a smallest effect size of interest in measurement invariance given the new publications on effect sizes within this area (Anvari & Lakens, 2021; Lakens et al., 2018). In this registered report, we propose to examine studies published within *Assessment* that report measurement invariance. We will create a database of studies that report measurement invariance and code these articles for the type of groups tested, steps of measurement invariance performed, and results obtained. This searchable database can be useful for further meta-research on measurement invariance or simple search for individuals searching for measurement instruments. Next, we will reproduce measurement invariance tests for publications with sufficient data and calculate the effect sizes for model and parameter level invariance. This information will be provided in the database to allow researchers to gauge what they may expect if they use a questionnaire or wish to replicate/extend previous work. We will provide the overall summary of effect sizes within measurement invariance tests and comment on the distributions of effect sizes found within the literature. We will end by providing researchers with suggestions for ways to determine their smallest effect size of interest for pre-registration or practical assessment.

# Proposed Method

## Database Curation



*Figure* 1. A flow chart of potential exclusions to create the database of measurement invariance.

Using *Assessment*’s online search feature, we will search for **measurement invariance** allowing for either term to be present in the manuscript for inclusion in the first round of papers (which would also cover related terms such as **measurement equivalence**). As of December 2024, this search returns over 600 articles from 1994 (i.e., the journal’s establishment date) to 2024 publications. The data will then be filtered to only include research articles under the article type filter present on the journal website. These citations will be exported to the Zotero group created for this project found at: <https://www.zotero.org/groups/5407184/measurement_invariance_assessment>. Figure 1 displays the filtering and coding process to create the measurement invariance database. We chose this journal primarily due to journal focus, and the number of available articles indicates a large sample size of articles to investigate measurement invariance.

### Screening procedure.

**Inclusion**: Each article will be coded for inclusion in the measurement invariance database and for potential further analysis by filling out a coding survey created for this purpose. The coding scale is included in Appendix A. We created the original article inclusion form using students enrolled in a structural equation modeling course, and pilot tested several articles to determine the scope of the project (see results below). We then tested if ChatGPT could respond to the survey questions and answer these questions accurately if given a copy of the research article. ChatGPT 4.0 appears to be able to answer questions accurately, and therefore, we will use the 4.0 model (or higher depending on time of study) to initially code articles. The example prompt and output from ChatGPT for an article with multiple invariance tests is included in Appendix A. The survey includes items to determine if the article includes measurement invariance, how the measurement was analyzed (i.e., structural equation modeling, item response theory, or other), and if the study included participant data (i.e., no simulation studies). Only research articles that used structural equation modeling to examine measurement invariance on participant data will be included in the database.

Measurement Invariance Variables: The next portion of the coding survey includes information about the measurement invariance test(s) in the manuscript. Each measurement test will be coded separately. ChatGPT will include the name and citation of the scale assessed, what groups are compared in the measurement invariance test, and the steps performed in the measurement invariance assessment. Once these steps are selected, ChatGPT will order them based on the manuscript, list the type of invariance claimed, and list the fit index used for determination of invariance, along with the rule (i.e., CFI, CFI > .01, RMSEA, RMSEA > .015 (**cheung2002?**)). If the data appears to be included, links for the repository will be extracted.

### Coders.

Coders will be recruited from author networks targeting researchers with self-identified experience in structural equation modeling. We will randomly select 10% of the articles processed by ChatGPT to check for accuracy. The coder will be given the responses from ChatGPT. They will note any incorrect answers or potential issues with output. The lead team will analyze these results to determine if consistent issues with particular questions answered by ChatGPT can be found. If these results indicate consistent inappropriate answers (i.e., more than 30% of the item is flagged completely incorrect), the lead team and coders will recode that item for all papers. For answers that are flagged as 30% partially incorrect (i.e., half correct, misleading, or otherwise not quite right), we will tweak the ChatGPT prompt until coders agree the issue has been fixed. Each coder will be asked to complete at least ten articles for credit (i.e., the data curation CRediT category, Allen et al., 2014). We will report these coding statistics for the randomly selected articles.

### Pilot Test.

A pilot test of the coding procedure was completed with the lead author’s structural equation modeling teaching course[^This test was completed before ChatGPT tests but provides information about the number of expected articles and scope of the project]. Each coder was assigned a specific issue of Volume 30 of *Assessment*, and they examined all articles within that issue. Approximately 38.30 percent of articles within those issues included measurement invariance (*n* = 120 unique articles coded). Of those articles, 66.70 % used structural equation modeling, and of those articles, 89.20 percent used participant data. 45 measurement invariance tests were coded, and approximately 42.20 percent included data for testing.

For the proposed project, we might expect that a smaller proportion will be excluded for only discussing measurement invariance, as the coding class project examined all articles within an issue, rather than just examining results returned from the online search. If the other percentages are approximate, then we might expect that 400 articles would use structural equation modeling, 357 would include participant data, and 151 may have available data for examination. Given the recent trends in transparency and openness, this number is likely an overestimate as the pilot included only new articles.

## Database Analysis

The final coding of articles and their measurement invariance tests will be presented as a database of measurement invariance results. These results can be re-used in other meta-analytic studies. We will present the following summary statistics:

* Prevalence of Measurement Invariance
  + Total number of articles published to date
  + Total number of invariance related articles
  + Split of articles into structural equation modeling, item response theory, and other analyses types
  + Number of articles that analyzed participant data (versus simulation only studies)

We will incorporate this information in Figure 1 similar to a Prisma schema to show the number of articles included in each step of the screening procedure.

* Measurement Invariance Test Statistics
* Commonly reported scales
* Frequency of number of groups compared
* Frequency of data inclusion for reproducibility

The statistics from this section will be calculated based on the number of papers coded.

* Commonly occurring group comparisons: we will code this variable into overall category labels. Variable group labels will be determined based on results from the study, but we expect certain categories such as age, gender/sex, language, country, and race/ethnicity.
* Frequency of each type of measurement invariance assessment (i.e., number of equal form, equal item intercepts, etc.)
* Commonly used fit statistics and rules of thumb for invariance testing
* Frequency of each type of measurement invariance claimed

The statistics from this section will be calculated based on the total number of measurement invariance tests coded. If a paper includes multiple invariance tests, it will therefore be included multiple times in the overall assessment of measurement invariance claimed (for example).

## Effect Size Calculation

Each measurement invariance test that included data will then be coded using the template in Appendix 2. Using the visualizemi package (Buchanan, 2024), coders will program a multigroup confirmatory factor analysis using the steps outlined from the research article and the provided data. Each model will first be examined for convergence across the same measurement invariance steps from the published articles, only using information present within the article or supplemental material for factor structure and any modifications made to scale structure (i.e., correlated residuals). Models that do not converge will be noted, and no more analysis on that data will be performed. Each model will then be tested for replication effect sizes rates at the model and parameter level. For example, an article that claimed measurement invariance for equal form, loadings, and intercepts will be tested with these same steps to determine the effect size of potential replication for each of those steps.

The visualizemi package calculates the effect size of potential replication by bootstrapping the original data with replacement and compares these results to the same bootstrapped dataset that has had the group labels randomized. Table 1 shows the results of model level bootstrapping using the Holzinger-Swineford data embedded in the *lavaan* package [CITE]. The factor loadings (*metric* invariance) were non-invariant when compared to a *configural* model 90% of the time while the randomized data was non-invariant only 10% of the time. The effect size represents the standardized difference between these two proportions and the amount of expected non-invariance if the study was replicated with similar data. The range of statistics is , so the column represents the scaled -1 to 1 for easier interpretation. At zero, we would expect to find that the bootstrapped data and randomized data do not show differences (thus, should be invariant), and as the statistic approaches 1, the more likely we are to find non-invariance in new samples.

Table 1

| model | non\_invariant | random\_non\_invariant | h\_nmi | h\_mi | h\_nmi\_p | h\_mi\_p |
| --- | --- | --- | --- | --- | --- | --- |
| intercepts | 0.15 | 0.01 | 0.59 | -0.59 | 0.19 | -0.19 |
| loadings | 0.80 | 0.06 | 1.74 | -1.74 | 0.55 | -0.55 |
| residuals | 0.04 | 0.06 | -0.11 | 0.11 | -0.03 | 0.03 |

*Note.*

If a model showed any partial invariance, we will then use the partial bootstrapping function in visualizemi to calculate the effect size of replicating partial invariance and the difference between parameter estimates. Given the HolzingerSwineford1939 data shown in Table 1, we would examine the results of partial invariance at the loadings step to determine where the potential group differences lie. Each factor loading would be individually relaxed one at a time and the bootstrapping procedure is examined for each loading separately. Table 2 demonstrates the results of the bootstrapping for the HolzingerSwineford1939 data, including the number of non-invariant results for each individual factor loading. We will report for each parameter, along with for the group size differences between groups for bootstrapped estimates. The effect size difference between parameter estimates is calculated by using for independent samples groups. These estimates are calculated on the standardized loadings so they are comparable across different scale examples.

Table 2

| term | boot\_1 | boot\_2 | random\_1 | random\_2 | boot\_fit | random\_fit | boot\_difference | random\_difference | boot\_index\_difference | random\_index\_difference |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| visual = x1 | 0.91 | 0.92 | 0.88 | 0.93 | 0.85 | 0.92 | -0.02 | -0.05 | FALSE | FALSE |
| visual = x3 | 0.52 | 0.63 | 0.65 | 0.57 | 0.85 | 0.92 | -0.11 | 0.08 | FALSE | FALSE |
| textual = x4 | 0.86 | 0.83 | 0.82 | 0.84 | 0.85 | 0.92 | 0.03 | -0.01 | FALSE | FALSE |
| textual = x5 | 0.87 | 0.91 | 0.91 | 0.88 | 0.85 | 0.92 | -0.03 | 0.02 | FALSE | FALSE |
| textual = x6 | 0.87 | 0.86 | 0.87 | 0.83 | 0.85 | 0.92 | 0.00 | 0.04 | FALSE | FALSE |
| speed = x7 | 0.54 | 0.61 | 0.53 | 0.59 | 0.85 | 0.92 | -0.07 | -0.06 | FALSE | FALSE |

*Note.*

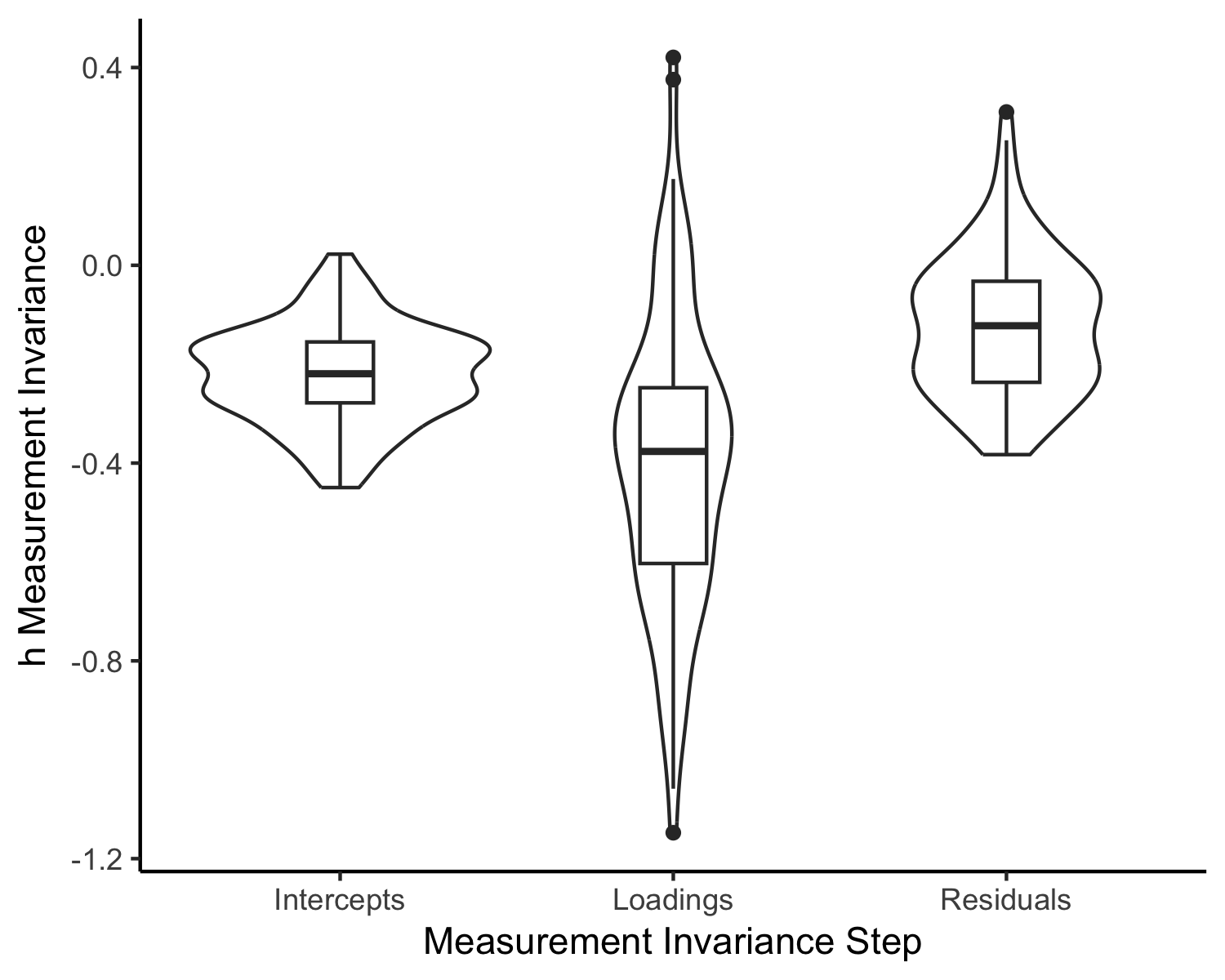
Each measurement invariance comparison found in the database curation (with data) will be programmed by one coder and checked by another coder using a small number of bootstrap simulations. If a disagreement arises during the code checking process, which cannot be resolved by the coders, the lead author will make the final decision. After both coders and the lead author (if applicable) agree that the coding matches the article data, the model and parameter level simulations will be run over 1000 simulations on a high performance computing server to ensure consistency in versions of *R* and packages. The exact versions will be reported for reproducibility. The results will be exported in both an HTML format and Rdata for further use. These files will be available on our repository at <https://github.com/doomlab/assessment-squared>.

### Model Level Results.

We will compile the results based on measurement invariance steps (i.e., configural, metric, scalar, etc.), and we will report the following (simulated data below):

* Mean, standard deviation, sample size for each measurement invariance step
* 25, 50, 75% quantile values
* Visualization of the distribution of effects - for example, Figure 2

These results are based on the total number of programmed measurement invariance analyses, as not all articles from the database curation are expected to include data.



*Figure* 2. A visualization of the effect sizes for measurement invariance at the model level.

### Parameter Level Results.

For the parameter level results, the bootstrapping function returns similar information for each of the parameters in the final step of the model as described above. We will present:

* Mean, standard deviation, sample size for each type of parameter (loadings, intercepts, residuals, etc.)
* 25, 50, 75% quantile values
* Visualization of the distribution of effects

Next, the function additionally returns effect sizes for group differences on each parameter, . This value is calculated as the average mean difference of the parameter estimates for each group divided by their pooled standard error across bootstraps. is calculated for both the bootstrapped data and the randomized data. We will calculate a “normalized” effect size for each parameter by subtracting for the bootstrapped data minus the for the randomized data. Because group order is arbitrary in our analysis, we will take the absolute value of the effect size for final reporting. From this dataset, we will report:

* Mean, standard deviation, sample size for each type of parameter (loadings, intercepts, residuals, etc.)
* 25, 50, 75% quantile values
* Visualization of the distribution of effects

## Results Interpretation

Our study is exploratory to determine the landscape of measurement invariance effect size using the premier journal outlet in clinical psychology for such publications. We make no predictions on the direction or size of the results. Instead, we will present the database for future reuse and examine the results for any consistent patterns or findings. By understanding the range of published values[[1]](#footnote-41), researchers can use these results to gauge their study results against. We will comment on potential ideas for estimating the smallest effect of interest for measurement invariance.

# References

Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature*, *508*(7496), 312–313. <https://doi.org/10.1038/508312a>

Anvari, F., & Lakens, D. (2021). Using anchor-based methods to determine the smallest effect size of interest. *Journal of Experimental Social Psychology*, *96*, 104159. <https://doi.org/10.1016/j.jesp.2021.104159>

Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and Reliability Reporting Practices in the Field of Health Education and Behavior: A Review of Seven Journals. *Health Education & Behavior*, *41*(1), 12–18. <https://doi.org/10.1177/1090198113483139>

Buchanan, E. M. (2024). Visualizemi: Visualization, Effect Size, and Replication of Measurement Invariance for Registered Reports. *Assessment*, 10731911241280763. <https://doi.org/10.1177/10731911241280763>

Byrne, B. M. (2001). Structural Equation Modeling With AMOS, EQS, and LISREL: Comparative Approaches to Testing for the Factorial Validity of a Measuring Instrument. *International Journal of Testing*, *1*(1), 55–86. <https://doi.org/10.1207/S15327574IJT0101_4>

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. <https://doi.org/10.1037/0033-2909.105.3.456>

Cheung, G. W., & Rensvold, R. B. (2000). Assessing Extreme and Acquiescence Response Sets in Cross-Cultural Research Using Structural Equations Modeling. *Journal of Cross-Cultural Psychology*, *31*(2), 187–212. <https://doi.org/10.1177/0022022100031002003>

DeVellis, R. F., & Thorpe, C. T. (2022). *Scale development: Theory and applications* (Fifth edition). SAGE Publications, Inc.

Dong, Y., & Dumas, D. (2020). Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and Individual Differences*, *160*, 109956. <https://doi.org/10.1016/j.paid.2020.109956>

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456465. <https://doi.org/10.1177/2515245920952393>

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. <https://doi.org/10.1177/2515245918770963>

Liu, M., Harbaugh, A. G., Harring, J. R., & Hancock, G. R. (2017). The effect of extreme response and non-extreme response styles on testing measurement invariance. *Frontiers in Psychology*, *8*, 726. <https://doi.org/10.3389/fpsyg.2017.00726>

Maassen, E., D’Urso, E. D., Van Assen, M. A. L. M., Nuijten, M. B., De Roover, K., & Wicherts, J. M. (2023). The dire disregard of measurement invariance testing in psychological science. *Psychological Methods*. <https://doi.org/10.1037/met0000624>

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing hu and bentler’s (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(3), 320–341. <https://doi.org/10.1207/s15328007sem1103_2>

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. <https://doi.org/10.1007/BF02294825>

Nye, C. D., & Drasgow, F. (2011). Effect size indices for analyses of measurement equivalence: Understanding the practical importance of differences between groups. *Journal of Applied Psychology*, *96*(5), 966–980. <https://doi.org/10.1037/a0022955>

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>

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology*, *6*. <https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01064>

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2017). The jingle and jangle of emotion assessment: Imprecise measurement, casual scale usage, and conceptual fuzziness in emotion research. *Emotion*, *17*(2), 267–295. <https://doi.org/10.1037/emo0000226>

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype Threat and Group Differences in Test Performance: A Question of Measurement Invariance. *Journal of Personality and Social Psychology*, *89*(5), 696–716. <https://doi.org/10.1037/0022-3514.89.5.696>

# (APPENDIX) Appendix

# Appendix A

## Coding Survey

Note: this document was created in qualtrics and then downloaded into word format to upload into ChatGPT. We formatted it for ease of reading here.

### Article Coding Form.

You will use this form to code articles for the Assessment of Assessment Project. You should fill out this form once for each article. If the article has multiple measurement invariance tests (i.e. once for sex and once for age), then you would fill out the article once for each test of measurement invariance within an article.

Note: the survey will end if your article does not have the required components we are looking for in the article. If you think you did something wrong, please ask or simply recode it.

### doi\_link.

Enter the article doi as an html link (i.e., <https://doi.org/10.1177/107319119400100401>)

### invariance\_yes.

Does this article report measurement invariance?

Search the document (control + F) for equal groups, invariance, multigroup. Notes:

* Code yes even if the data is simulated
* Code yes if item response theory differential item functioning is used - Code other for unsure or unclear or list your own reason
  + Yes (1)
  + No (2)
  + Other (3) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### cfa\_irt.

Does this article use structural equation modeling (confirmatory factor analysis) or item response theory?

* Structural Equation Modeling (1)
* Item Response Theory (2)
* Neither: List Analysis (3) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### real\_data.

Does this article include real data (i.e., no simulation studies)?

* Yes (1)
* No (2)
* Unclear (3) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### data.

Is the data accessible? Look for supplemental documents, links to files, and pages of correlation/covariance matrices.

* Yes (1)
* Matrices included (2)
* Contact author (3)
* No (4)
* Unclear or broken links (5) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

### scale\_name.

What is the name of the instrument/scale they are testing? Please include the entire name, not the abbreviation.

### scale\_ref.

What is the citation of the scale they are testing? Copy from the reference section.

### group\_number.

How many groups are compared?

### group\_names.

What sample groups are they comparing in this analysis? List groups the way they are described (i.e., do not correct Male/Female to Men/Women, use the names from the paper). Separate groups with a comma.

### what\_steps.

What steps of measurement invariance did they test? Please click all that were used. The names in parentheses are sometimes used to describe each step.

* Equal form (configural) (1)
* Equal item loadings (metric) (2)
* Equal item intercepts/means (scalar) (3)
* Equal item thresholds (4)
* Equal item residuals (strict) (5)
* Equal item residual covariances (8)
* Equal latent means (population means) (6)
* Equal latent variances (7)
* Equal latent covariances (9)
* Equal regressions (10)

### step\_order.

Carry Forward Selected Choices from “What steps of measurement invariance did they test? Please click all that were used. The names in parentheses are sometimes used to describe each step.”

What order did they test the steps in? Please drag and drop them into the order found in the paper (often in a table).

* Equal form (configural) (1)
* Equal item loadings (metric) (2)
* Equal item intercepts/means (scalar) (3)
* Equal item thresholds (4)
* Equal item residuals (strict) (5)
* Equal item residual covariances (6)
* Equal latent means (population means) (7)
* Equal latent variances (8)
* Equal latent covariances (9)
* Equal regressions (10)

### metric\_mi.

What metric did they use to assess measurement invariance? (i.e., CFI, RMSEA, other fit measures, you can use abbreviations of fit indices)

### rule\_mi.

What rule did they use for measurement invariance? (i.e., change in CFI < .01, chi-square difference test, etc.). List the rule as what would be considered invariant (i.e., passes the test, groups are considered equal).

### invariance\_claim.

What type of invariance did they claim? Use the step name and their words for the type of invariance. For example, fully invariant to residuals/strict, partially scalar invariant, non-invariant, etc.)

## ChatGPT Results

### Prompt.

Given the survey questions in Article\_Coding\_Assessment.docx, answer the survey questions for the panayiotou paper as one answer per row in a tabular format. (attached survey above + pdf)

### Answer.

Here’s a table summarizing the answers for the survey questions based on the Panayiotou et al. paper:

| Survey Question | Answer |
| --- | --- |
| DOI Link | <https://doi.org/10.1177/10731911221119533> |
| Does the article report MI? | Yes |
| Does the article use CFA/IRT? | Structural Equation Modeling |
| Does the article include real data? | Yes |
| Is the data accessible? | Unclear or broken links (mention of OSF link: <https://osf.io/cqxrd/>, but not verified in text) |
| Scale name | Revised University of California Los Angeles Loneliness Scale (UCLA-LSR) |
| Scale reference | Russell, D., Peplau, L. A., & Ferguson, M. L. (1978). Development of a measure of loneliness. Journal of Personality Assessment, 42(3), 290–294. |
| Number of groups compared | Seven |
| Group names | 18–25, 26–35, 36–45, 46–55, 56–65, 66–75, 76+ |
| Steps of MI tested | Equal form (configural), Equal item loadings (metric), Equal item intercepts/means (scalar), Partial scalar invariance |
| Order of steps tested | Configural -> Metric -> Scalar |
| Metric used for MI | CFI, RMSEA, SRMR |
| Rule for MI | CFI difference < 0.01 |
| Invariance claim | Partial scalar invariance for the UCLA-LSR with varying results across item intercepts |

This table consolidates relevant information. If you need deeper details on any specific part, let me know!

### Prompt.

Do the same thing for the zager paper attached (attached pdf)

### Answer.

Here’s a table summarizing the answers for the survey questions based on the Zager Kocjan et al. paper:

– table included but see note below –

This table consolidates the key survey answers based on the paper. Let me know if you need additional details or refinements!

### Prompt.

In the tabular analysis of the Zager paper, reformat the table so that the number of groups, names of groups, and invariance claimed for groups are broken out into separate rows

### Answer.

Here is the reformatted table for the Zager Kocjan et al. paper:

| Survey Question | Answer |
| --- | --- |
| DOI Link | <https://doi.org/10.1177/1073191121993558> |
| Does the article report MI? | Yes |
| Does the article use CFA/IRT? | Yes, uses Multi-Group Confirmatory Factor Analysis (MG-CFA) |
| Does the article include real data? | Yes |
| Is the data accessible? | No explicit mention of accessible data or shared datasets. |
| Scale name | Subjective Happiness Scale (SHS) |
| Scale reference | Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. Social Indicators Research, 46(2), 137–155. |
| Number of groups compared (Countries) | 9 |
| Names of groups (Countries) | Australia, Colombia, Czech Republic, England, Hungary, Mexico, New Zealand, Slovenia, USA |
| Invariance claimed (Countries) | Configural and metric invariance achieved; partial scalar invariance with varying intercepts for Item 2 |
| Number of groups compared (Genders) | 2 |
| Names of groups (Genders) | Male, Female |
| Invariance claimed (Genders) | Configural, metric, and scalar invariance achieved |
| Number of groups compared (Age) | 4 |
| Names of groups (Age) | 17–30, 31–45, 46–60, 61–75 |
| Invariance claimed (Age) | Configural, metric, and scalar invariance achieved |
| Number of groups compared (Time) | 5 |
| Names of groups (Time) | Time points 1–5 (3-month intervals over 1 year) |
| Invariance claimed (Time) | Configural, metric, and scalar invariance achieved |
| Steps of MI tested | Configural, Metric, Scalar |
| Order of steps tested | Configural -> Metric -> Scalar |
| Metric used for MI | CFI, RMSEA, SRMR |
| Rule for MI | CFI difference ≤ 0.01, ΔRMSEA ≤ 0.015, ΔSRMR ≤ 0.030 |

This breakout provides separate rows for each group type (countries, genders, ages, and time points) and their respective invariance details. Let me know if you need further refinements!

# Example Data Analysis Code

## Libraries

library(lavaan)  
library(visualizemi)  
library(rio)  
# set the seed  
set.seed(43843)

## Data

# import the dataset here

## Model Programming

Program the model in lavaan syntax including all correlated errors or other adjustments noted by the original authors.

# example model   
HS.model <- ' visual =~ x1 + x2 + x3  
 textual =~ x4 + x5 + x6  
 speed =~ x7 + x8 + x9 '

## Use MGCFA function

Use the mgcfa function to run the proposed model steps. This model should replicate their model excluding any partial invariance steps. If they say they will run “residuals” but then do not because of partial invariance, only include up to the step that was invariant or partially invariant.

# cfa model  
saved\_mgcfa <- mgcfa(model = HS.model,  
 # dataset in data frame   
 data = HolzingerSwineford1939,   
 # grouping variable column as a character  
 group = "sex",  
 # lavaan syntax for group constraints  
 group.equal = c("loadings", "intercepts", "residuals"),   
 # any other lavaan cfa arguments  
 meanstructure = TRUE)

## Review Model Outputs

Review model outputs to ensure they do not produce heywood cases or other errors.

# review the model outputs to ensure they do not produce errors  
saved\_mgcfa$model\_fit

## # A tibble: 7 × 18  
## agfi AIC BIC cfi chisq npar rmsea rmsea.conf.high srmr tli  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.991 7535. 7647. 0.931 85.3 30 0.0921 0.114 0.0595 0.896  
## 2 0.989 3671. 3761. 0.906 61.4 30 0.103 0.136 0.0749 0.859  
## 3 0.991 3846. 3938. 0.959 44.4 30 0.0741 0.108 0.0525 0.938  
## 4 0.990 7518. 7740. 0.935 106. 60 0.0894 0.113 0.0634 0.903  
## 5 0.989 7526. 7726. 0.919 126. 54 0.0943 0.116 0.0741 0.892  
## 6 0.989 7533. 7711. 0.905 145. 48 0.0970 0.117 0.0753 0.886  
## 7 0.988 7537. 7682. 0.890 167. 39 0.0972 0.116 0.0926 0.886  
## # ℹ 8 more variables: converged <lgl>, estimator <chr>, ngroups <int>,  
## # missing\_method <chr>, nobs <int>, norig <int>, nexcluded <int>, model <chr>

# overall  
saved\_mgcfa$model\_overall

## lavaan 0.6-19 ended normally after 35 iterations  
##   
## Estimator ML  
## Optimization method NLMINB  
## Number of model parameters 30  
##   
## Number of observations 301  
##   
## Model Test User Model:  
##   
## Test statistic 85.306  
## Degrees of freedom 24  
## P-value (Chi-square) 0.000

# groups  
# saved\_mgcfa$group\_models$model.1  
# saved\_mgcfa$group\_models$model.2  
  
# configural  
# saved\_mgcfa$model\_configural  
  
# constraints  
# saved\_mgcfa$invariance\_models$model.loadings  
# saved\_mgcfa$invariance\_models$model.intercepts  
# saved\_mgcfa$invariance\_models$model.residuals

## Replication Model Level

Run the bootstrap\_model function and determine the effect sizes for model level invariance steps. Use the same group.equal constraints as above.

saved\_boot\_model <- bootstrap\_model(  
 # saved configural model to start at  
 saved\_configural = saved\_mgcfa$model\_configural,  
 # dataset for the analysis  
 data = HolzingerSwineford1939,  
 # model lavaan syntax  
 model = HS.model,  
 # group variable in the dataset  
 group = "sex",   
 # number of bootstraps  
 # this is set to a low number for testing  
 nboot = 20,  
 # name of the fit measure you want to use, make sure it's lavaan code  
 invariance\_index = "cfi",  
 # rule for the difference in fit indices  
 invariance\_rule = .01,  
 # what order of steps do you want to test?   
 group.equal = c("loadings", "intercepts", "residuals")  
)

## Finished Bootstrap Number: 1  
## Finished Bootstrap Number: 2  
## Finished Bootstrap Number: 3  
## Finished Bootstrap Number: 4  
## Finished Bootstrap Number: 5  
## Finished Bootstrap Number: 6  
## Finished Bootstrap Number: 7  
## Finished Bootstrap Number: 8  
## Finished Bootstrap Number: 9  
## Finished Bootstrap Number: 10  
## Finished Bootstrap Number: 11  
## Finished Bootstrap Number: 12  
## Finished Bootstrap Number: 13  
## Finished Bootstrap Number: 14  
## Finished Bootstrap Number: 15  
## Finished Bootstrap Number: 16  
## Finished Bootstrap Number: 17  
## Finished Bootstrap Number: 18  
## Finished Bootstrap Number: 19  
## Finished Bootstrap Number: 20

kable(saved\_boot\_model)

| model | non\_invariant | random\_non\_invariant | h\_nmi | h\_mi | h\_nmi\_p | h\_mi\_p |
| --- | --- | --- | --- | --- | --- | --- |
| intercepts | 0.15 | 0.00 | 0.7953988 | -0.7953988 | 0.2531833 | -0.2531833 |
| loadings | 0.75 | 0.10 | 1.4508940 | -1.4508940 | 0.4618339 | -0.4618339 |
| residuals | 0.10 | 0.05 | 0.1924743 | -0.1924743 | 0.0612665 | -0.0612665 |

## Replication Parameter Level / Partial Invariance

If their model was invariant: use the last step as the partial\_step option. If their model was partially invariant, use the step that was indicated as partially invariant (where they had to relax parameters - if they did this on more than one step, use the last one).

saved\_boot\_partial <- bootstrap\_partial(  
 # the model you want to test   
 # use the model before your invariant one  
 # in this case it broke down at loadings  
 saved\_model = saved\_mgcfa$model\_configural,  
 # the dataframe  
 data = HolzingerSwineford1939,  
 # the original model syntax  
 model = HS.model,  
 # the grouping variable column  
 group = "sex",  
 # run more but start low for testing  
 nboot = 10,  
 # what index are you using for invariance?  
 # match this to lavaan's name under fitmeasures()  
 invariance\_index = "cfi",  
 # what rule are you using?  
 invariance\_rule = .01,  
 # what are we comparing against?   
 invariance\_compare = unname(fitmeasures(saved\_mgcfa$model\_configural, "cfi")),  
 # which step you want to estimate effect size for  
 partial\_step = c("loadings"),   
 # which parameters you want to hold constrained  
 group.equal = c("loadings")  
 )  
  
kable(head(saved\_boot\_partial$boot\_DF))  
  
kable(saved\_boot\_partial$boot\_summary)  
  
kable(saved\_boot\_partial$boot\_effects)  
  
saved\_boot\_partial$invariance\_plot  
  
saved\_boot\_partial$effect\_invariance\_plot  
  
saved\_boot\_partial$density\_plot

1. The advantage of studying measurement invariance in this scenario is that both invariant and non-invariant models are typically publishable. While publication bias likely still exists, it may be less than traditional effect size meta-analysis studies. [↑](#footnote-ref-41)