Replication Report

Anna Lohmann[1], Arjan Huizing [2]

[1] Leiden University Medical Center

[2] University of some other place

March 5, 2022

Abstract

<a summary of the replication effort>

Correspondence concerning this replication report should be addressed to:
primary_a.l.lohmann@lumc.nl

1 Introduction

This replication report documents the replication attempt of the simulation study Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. Psychological Methods, 17(3), 354–373. https://doi.org/10.1037/a0029315
Following the definition of Rougier et al. (2017) we understand the replication of a published study as writing and running new code based on the description provided in the original publication with the aim of obtaining the same results.

2 Method

2.1 Information basis

The replication attempt was based on the information provided in the original manuscript as well as the supplemental material accompanying the publication. The main text provided a link to the supplements (http://dx.doi.org/10.1037/a0029315.supp) which referred to the website of the publisher where an additional pdf document with extensive result tables was freely available. `<What sources were used to obtain information? The original article, some appendix, online supplements, other articles from the same authors, code available from the authors personal website?>`

2.2 Data Generating Mechanism

Information provided in the above mentioned sources indicated that the following simulation factors were systematically varied in a full factorial design for generating the artificial data.

Simulation factor	No. levels	Levels
Fixed		
CFA model size	2	10 indicators, 20 indicators
Underlying distribution	2	normal, non-normal
Number of categories	6	2,3,4,5,6,7
Threshold symmetry	5	symmetry, moderate asymmetry, moderate asymmetry alternative, extreme asymmetry
Sample Size	4	100, 150, 350, 600

Data was generated according to these 480 simulation scenarios. This was repeated for 1000 repetitions. The data generating mechanism consisted of two steps. (1) Data was generated based on the underlying distribution, CFA model and sample size. (2) The generated data was categorized based on the given category

thresholds corresponding to a given number of categories and threshold symmetry.

2.2.1 Underlying distribution, CFA model size and Sample Size

The original study indicated that "[c]ontinous data (normal and nonnormal) were generated in EQS (Version 6.1; Bentler,2008) using methods developed by Fleishman (1978) and Vale Maurelli (1983)." We emulated this approach using the `generate()` function from the `simsem` package with the parameter `inDist` set to `NULL` in the normal case and to `simsem::bindDist(skewness = 2, kurtosis = 7)` in the non-normal case. This function also to the CFA model (as matrix …) as well as the sample size as input constituing the first step of the data generation.

2.2.2 Number of categories and Threshold symmetry

`<More detail of how factor 3 was varied and implemented>` After data was generated based on the given CFA model and the underlying distribution the resulting data was categorized into the number of categories for the scenario at hand. For each number of categories and each threshold symmetry Z-scores for category thresholds could be obtained from the first table of the supplemental material.

`<You can add pseudocode or a flowchart to illustrate the data generation or the entire simulation design>`

Data generation can be summarized with the following pseudo code:

```
For 1000 repetitions of each of 480 unique scenarios:

- Sample data according to the given CFA model, sample size as well as underlying distribution of the
  scenario at hand.

- Categorize data into the number of categories for the scenario at hand. Category thresholds depended
  on the threshold symmetry of a given scenario.

- If any sample covariance matrix was not positive definitive, repeat sampling until it is.

- Analyse data using a robust ML approach.

- Analyse data using a robust ULS approach.

- Remove results that ....

- Obtain performance measures.

    * Parameter estimates

    * Bias.

    * Compute ... based on these random elements.

    * Determine ... based on mechanism of current scenario.

- If some condition is > x:

    * Determine ... & resample from corresponding ... model.

- Apply ...
```

2.3 Investigated Methods

The study compares the performance of robust normal theory maximum likelihood (ML) and robust categorical least squares (ULS) methodology for estimating confirmatory factor analysis (CFA) with ordinary variables. The underlying CFA model was fit using each of the two methods under investigation. `<Describe the methods that are investigated and how they are implemented>`

2.3.1 robust normal theory maximum likelihood (ML)

`<Describe how the first method is defined and implemented. You can include equations and or R code. If applicable, mention specialized R packages, their version as well as, parameters of specific functions.>`

2.3.2 robust categorical least squares (ULS)

`<Describe how the second method is defined and implemented. You can include equations and or R code. If applicable, mention specialized R packages, their version as well as, parameters of specific functions.>`

2.4 Performance measures

The two methods described above were compared on various performance measures.

2.4.1 Convergence Failures

2.4.2 Improper solutions

2.4.3 Outliers

The orinal study defined outliers as “any cases that produced a standard error greater than 1.” We implemented this as the robust standard errors listed in the lavaan fit object.

2.4.4 Parameter Estimates

2.4.5 Parameter Bias

2.4.6 Efficiency

2.4.7 Relative bias for robust standard errors

2.4.8 Coverage

2.4.9 Type I error rate

2.4.10 Outliers

`<Describe which performance measures are compared, if applicable mention specialized R packages, their versions, as well as parameter settings of functions.>`

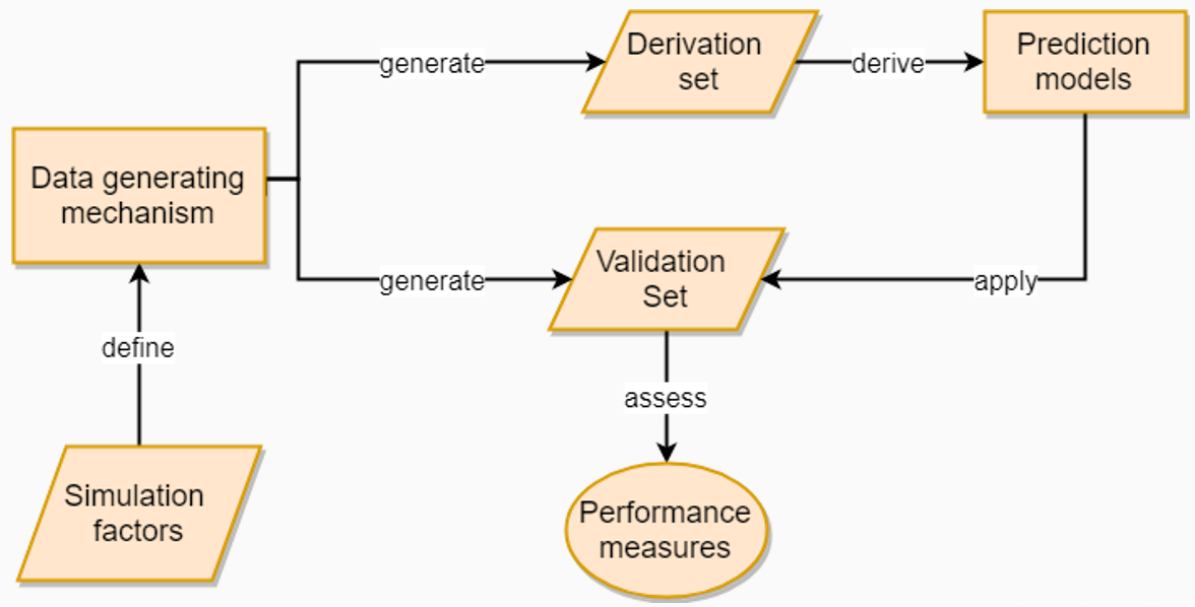

Figure 1: Flow chart of data generating mechanism

The following flowchart describes the simulation design

2.5 Technical implementation

The original simulation study was carried out in EQS (Version 6.1) as well as Mplus (Version 6.11). The authors of the original study report that data generation was carried out in EQS and data analysis was conducted using both EQS as well as MPlus. However, only results from the Mplus analysis are reported. Our replication was implemented using the R programming environment (details regarding software versions can be obtained from the section Reproducibility Information). The corresponding R code can be obtained from https://github.com/replisims/rhemtulla-2012.

The following table provides an overview of replicator degrees of freedom, i.e. decisions that had to be made by the replicators because of insufficient or contradicting information. Issues were resolved by discussion among the replicators. Decisions were based on what the replicators perceived to be the most likely implementation with likeliness estimated by common practice and/or guideline recommendations. Wherever feasible multiple interpretations where implemented.

Issue	Replicator decision	Justification
Data dependence	each scenario is implemented in independently generated data	Best practice (Burton et al. 2006)

2.6 Data basis for Figures 1 and 2

<More details on how the information provided was insufficient, unclear or vague> *"Some weird quote from the original article that you could not make any sense of"* (p.XY) We could not infer whether "for each condition" also included for each model size. We interpreted the text such that this was the case and hence collapsed the distributions across both model sizes.

2.7 Another issue

<More details on how the information provided was insufficient, unclear or vague> *"Some weird quote from the original article that you could not make any sense of"* (p.XY)

3 Results

3.1 Replication of result figures

3.2 Simulation descriptives

<Describe the sampling distribution if any of the simulation parameters were sampled> The original study provides descriptives for the simulated data in two figures. Figure 1 and Figure 2 of the original manuscript

3.3 Replication of result tables

<Compare any tabulated data to the original> Table 1 presents the "Skew and Kurtosis of Observed Categorical Variables by Threshold Distribution, Underlying Distribution, and Number of Categories" (p.363). The "[v]alues in this table were obtained by generating samples of size N = 1,000,000 for each condition and recording the skew and kurtosis of the observed distributions." (p.363) As discussed above we understood "each condition" to also refer to the model size. Our results are hence pooled across model size.

3.4 Replication of results presented in text form

While the vast majority of results is presented in the form of figures, a few outcomes regarding outliers, relative bias of parameter estimates as well as relative bias of robust standard errors are only communicated in text form. <If the text describes any results using words describe how that relates to your findings.>

3.4.1 Outliers

The original study reports the frequency of outliers in the text. There was one outlier in the original study. In our replication we found ...

3.4.2 Relative bias

Figures and tables report absolute bias. Results pertaining to relative bias are only summarized in a more qualitative manner in text form. “As the number of categories increases, ML estimates gradually become less biased and by five categories relative bias is always less than 10%.(p.362)” “When the underlying distribution is non-normal, all cat-LS parameter estimates take on a slightly positive bias (around 4%), except when there are just two categories.” (p.364) “[B]ias is almost never greater that 5% with either method.

3.4.3 Relative bias for robust standard error estimates

“ML standard errors are from 8% to 30% (average = 15%) smaller than empirical standard errors when the sample size is small, and cat-LS standard errors are from 3% to 37% (average 13%) smaller than empirical standard errors when the sample size is small.” “Cat-LS produces better robust standard errors for factor loadings, and ML produces better robust standard errors for factor correlations. This finding isconsistent across number od categories.

4 Discussion

4.1 Replicability

Due to the high amount of details in the original publication and the corresponding supplemental materials the replication was straight forward. The largest amount of time was spent ensuring that the methods used for data generation and analysis did indeed correspond to what was used in the original study. This is, however, in no way the fault of the authors but rather due to limited documentation of the R packages used fro replication. On the contrary the detailed description of the implementation allowed for a close correspondence of methodology which would have otherwise been left to guesswork.

A feature that deserves special praise with regards to facilitating replicability is the high amount of documentation that the authors dedicated to the generation of the simulated data as well as the descriptives of the same. The ability to closely monitor the data generation process and compare features of the simulated data to the original study instilled a great deal of confidence in the replicators and ensured that any potential deviations of results could not be attributed to faulty interpretation and implementation of the data generating mechanism.

Another feature that increased reproducibility was the structure of the manuscript. The very first element of the method section was an overview of the simulation factors. Readability was increased by listing each factor as a separate bullet point. Subsequent sections detailed the implementation of each simulation factor. A separate subheading for each simulation factor made it easy to locate relevant information.

The detailed description of error handling procedures as well as error descriptives …

The large number of result tables presented in the supplemental material is another exemplary reporting practice worth highlighting. While the comparison of hundreds of table cells is not an easy endeavor and the

general interest in these tables likely limited it protects the authors against any allegations of selective reporting and makes the assessment of replicability possible.
A similar structure could be found for the performance measures which were discussed in separate subsections separated by corresponding heading. While very readable as is, we would have however preferred the performance measures to be elaborated on as part of the method section instead of the result section.
The introduction section included the presentation and discussion of several closely related methods as well as findings from previous studies investigating the same. Due to the large amount of information surrounding highly similar methods and their implementation it took us several readings of the introduction to feel confident about having identified the version actually implemented in the study at hand. A clearer separation of the implemented methods (e.g. in a box) would have facilitated isolating the relevant implementation details.
Finally, a major factor facilitating the reproduction process was the availability of specialized SEM software in the R programming environment. As R is frequently used for simulation studies investigating SEM methodology we were able to build upon a code base that was designed for this very purpose. While such specialized software has the potential of huge time savings on the coding end and additionally is likely to minimize coding errors on the part of the replicator it consumes a significant amount of time to familiarize oneself with the exact parameters underlying the tools. The inexperienced user is at the mercy of the package documentation and the occasional peek under the hood of a given function. Having a code base from related simulation studies available would increase confidence in using such tools and avoid some trial and error while familiarizing oneself with the functionalities.

<Provide a general statement of how you experienced the replication process. Was it easy? What made it easy or difficult?>

4.2 Replicator degrees of freedom

<Here you can discuss the replicator degrees of freedom. What could the authors have done to make it more clear? Do you think the replicator degrees of freedom are so extensive that they could influence the results?>

4.3 Equivalence of results

<How would you judge the overall equivalence of results? Are the orders of magnitude comparable? Are trends in the same direction? Would you draw the same conclusions as the authors based on your replication? Were some results not comparable because of insufficient figure resolution or labeling? Did the authors ommit some results which consequently cannot be compared?>

5 Acknowledgments

<Acknowledge the help of anyone who assisted you in the process>

6 Contributions

Authors made the following contributions according to the CRediT framework https://casrai.org/credit/

Anna Lohmann:

- Data Curation
- Formal Analysis (lead)
- Investigation
- Software
- Visualization (lead)
- Writing - Original Draft Preparation
- Writing - Review & Editing

Arjan Huizing:

- Formal Analysis (supporting)
- Investigation
- Software (supporting)
- Visualization (supporting)
- Validation
- Writing - Review & Editing

References

10 Burton, Andrea, Douglas G. Altman, Patrick Royston, and Roger L. Holder. 2006. “The Design of Simulation Studies in Medical Statistics.” *Statistics in Medicine* 25 (24): 4279–92. https://doi.org/10.1002/sim.2673.

Rougier, Nicolas P., Konrad Hinsen, Frédéric Alexandre, Thomas Arildsen, Lorena A. Barba, Fabien C. Y. Benureau, C. Titus Brown, et al. 2017. “Sustainable Computational Science: The ReScience Initiative.” *PeerJ Computer Science* 3 (December): e142. https://doi.org/10.7717/peerj-cs.142.

Appendix

Additional result

`<insert additional results not reported in the original article or results presented in an alternative way>`

6.1 Code organization

The code and the files associated are organized in the form of a research compendium which can be found in the following git repository `https://github.com/replisims/rhemtulla-2012`

```
## .
## +-- defs.tex
## +-- flowchart.PNG
## +-- Lato-Black.ttf
## +-- Lato-BlackItalic.ttf
## +-- Lato-Bold.ttf
## +-- Lato-BoldItalic.ttf
## +-- Lato-Italic.ttf
## +-- Lato-Regular.ttf
## +-- references.bib
## +-- Replication Report Rhemthulla et al 2012.Rmd
## +-- Replication-Report-Rhemthulla-et-al-2012.log
## +-- Replication-Report-Rhemthulla-et-al-2012.pdf
## +-- Replication-Report-Rhemthulla-et-al-2012.Rmd
## +-- Replication-Report-Rhemthulla-et-al-2012.tex
## +-- UbuntuMono-Bold.ttf
## +-- UbuntuMono-BoldItalic.ttf
## +-- UbuntuMono-Italic.ttf
## \-- UbuntuMono-Regular.ttf
```

- `foldername`: contains `<insert description>`
- `filename`: contains `<insert description>`
- ...

Reproducibility Information

This report was last updated on 2022-03-05 15:03:17. The simulation replication was conducted using the following computational environment and dependencies:

```
## - Session info ---------------------------------------------------------------
##  setting  value
##  version  R version 4.1.2 (2021-11-01)
##  os       Windows 10 x64 (build 19043)
##  system   x86_64, mingw32
##  ui       RTerm
##  language (EN)
##  collate  English_United States.1252
```

```
##  ctype    English_United States.1252
##  tz       Europe/Berlin
##  date     2022-03-05
##  pandoc   2.14.0.3 @ C:/Program Files/RStudio/bin/pandoc/ (via rmarkdown)
##
## - Packages ------------------------------------------------------------------
##  package        * version    date (UTC) lib source
##  assertthat       0.2.1      2019-03-21 [1] CRAN (R 4.1.2)
##  cachem           1.0.6      2021-08-19 [1] CRAN (R 4.1.2)
##  callr            3.7.0      2021-04-20 [1] CRAN (R 4.1.2)
##  cli              3.1.0      2021-10-27 [1] CRAN (R 4.1.2)
##  crayon           1.4.2      2021-10-29 [1] CRAN (R 4.1.2)
##  DBI              1.1.2      2021-12-20 [1] CRAN (R 4.1.2)
##  desc             1.4.0      2021-09-28 [1] CRAN (R 4.1.2)
##  devtools         2.4.3      2021-11-30 [1] CRAN (R 4.1.2)
##  digest           0.6.29     2021-12-01 [1] CRAN (R 4.1.2)
##  dplyr          * 1.0.8      2022-02-08 [1] CRAN (R 4.1.2)
##  ellipsis         0.3.2      2021-04-29 [1] CRAN (R 4.1.2)
##  evaluate         0.14       2019-05-28 [1] CRAN (R 4.1.2)
##  fansi            1.0.2      2022-01-14 [1] CRAN (R 4.1.2)
##  fastmap          1.1.0      2021-01-25 [1] CRAN (R 4.1.2)
##  fs               1.5.2      2021-12-08 [1] CRAN (R 4.1.2)
##  generics         0.1.2      2022-01-31 [1] CRAN (R 4.1.2)
##  glue             1.6.2      2022-02-24 [1] CRAN (R 4.1.2)
##  htmltools        0.5.2      2021-08-25 [1] CRAN (R 4.1.2)
##  knitr          * 1.37       2021-12-16 [1] CRAN (R 4.1.2)
##  lifecycle        1.0.1      2021-09-24 [1] CRAN (R 4.1.2)
##  magrittr         2.0.2      2022-01-26 [1] CRAN (R 4.1.2)
##  memoise          2.0.1      2021-11-26 [1] CRAN (R 4.1.2)
##  pillar           1.7.0      2022-02-01 [1] CRAN (R 4.1.2)
##  pkgbuild         1.3.1      2021-12-20 [1] CRAN (R 4.1.2)
##  pkgconfig        2.0.3      2019-09-22 [1] CRAN (R 4.1.2)
##  pkgload          1.2.4      2021-11-30 [1] CRAN (R 4.1.2)
##  prettyunits      1.1.1      2020-01-24 [1] CRAN (R 4.1.2)
##  processx         3.5.2      2021-04-30 [1] CRAN (R 4.1.2)
##  ps               1.6.0      2021-02-28 [1] CRAN (R 4.1.2)
##  purrr            0.3.4      2020-04-17 [1] CRAN (R 4.1.2)
##  R6               2.5.1      2021-08-19 [1] CRAN (R 4.1.2)
##  remotes          2.4.2      2021-11-30 [1] CRAN (R 4.1.2)
##  RepliSimReport   0.0.0.9000 2022-02-03 [1] Github (replisims/RepliSimReport@5f14003)
##  rlang            1.0.1      2022-02-03 [1] CRAN (R 4.1.2)
##  rmarkdown        2.11       2021-09-14 [1] CRAN (R 4.1.2)
##  rprojroot        2.0.2      2020-11-15 [1] CRAN (R 4.1.2)
##  rstudioapi       0.13       2020-11-12 [1] CRAN (R 4.1.2)
##  sessioninfo      1.2.2      2021-12-06 [1] CRAN (R 4.1.2)
##  stringi          1.7.6      2021-11-29 [1] CRAN (R 4.1.2)
##  stringr          1.4.0      2019-02-10 [1] CRAN (R 4.1.2)
##  testthat         3.1.1      2021-12-03 [1] CRAN (R 4.1.2)
##  tibble           3.1.6      2021-11-07 [1] CRAN (R 4.1.2)
##  tidyselect       1.1.1      2021-04-30 [1] CRAN (R 4.1.2)
##  usethis          2.1.5      2021-12-09 [1] CRAN (R 4.1.2)
##  utf8             1.2.2      2021-07-24 [1] CRAN (R 4.1.2)
##  vctrs            0.3.8      2021-04-29 [1] CRAN (R 4.1.2)
##  withr            2.4.3      2021-11-30 [1] CRAN (R 4.1.2)
##  xfun             0.29       2021-12-14 [1] CRAN (R 4.1.2)
```

```
##  xtable         * 1.8-4      2019-04-21 [1] CRAN (R 4.1.2)
##  yaml             2.2.1      2020-02-01 [1] CRAN (R 4.1.1)
##
##  [1] C:/Users/alohmann/Documents/R/win-library/4.1
##  [2] C:/Program Files/R/R-4.1.2/library
##
## ------------------------------------------------------------------------------
```

The current Git commit details are:

```
## Local:    test C:/Users/alohmann/Dropbox/anna/projects_new/replisims/replications/rhemtulla-2012
## Remote:   test @ origin (https://github.com/replisims/rhemtulla-2012.git)
## Head:     [7b6ab82] 2022-02-28: A lot of stuff
```