Replication Report

Y. Andre Wang[1], Udi Alter [2]

[1] University of Toronto Scarborough

[2] York University

October 24, 2021

Abstract

This replication report documents the replication attempt of the simulation study reported in Flora and Curran (2004). Although the original simulation code was not available to us, the article provided most of the theoretical information and instruction required to replicate the study, save for a few ambiguities (e.g., the exact tau values used to transform continuous data into ordinal data). Moreover, a recent article by Chalmers and Adkins (2020) provided the code for a partial replication of the same study using the SimDesign package in R. We recruited the code provided by Chalmers and Adkins (2020) as a base for more a complete set of simulation studies to replicate tables, graphs, and conditions reported in Flora and Curran (2004). Results from our simulation were overall consistent with the original simulation results. However, a few small differences surfaced: (1) The rates of improper solutions at small sample sizes (N = 100 and 200) using full weighted least squares estimation were higher in the replication than those in the original study; (2) the relative bias of some parameters was consistently positive in the original study but negative in the corresponding conditions in the replication; (3) the magnitude and pattern of relative bias of the factor loadings estimates differed somewhat between the original study and the replication for the two-factor models. One possible explanation for these inconsistencies might be due to the use of different software (the original study used EQS and Mplus, whereas the replication used R). A further investigation into these differences might be warranted. Full results, tables, and figures are presented below.

Correspondence concerning this replication report should be addressed to:
ylawang@ucdavis.edu

1 Introduction

This replication report documents the replication attempt of the simulation study Flora & Curran (2004). Following the definition of @rougier_sustainable_2017-1, 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

We drew from two sources to obtain information needed to run the replication. First, we relied on the descriptions of the simulations as reported in the original article, Flora and Curran (2004; referred to as "F&C" below). Second, we drew from the code written using the SimDesign R package by Chalmers and Adkins (2020; referred to as "C&A" below). Although F&C referred to a technical appendix that ostensibly contains example EQS code for data generation and Mplus code for model estimation (Footnote 13, p. 477; Footnote 14, p. 481), both links to the appendix were broken and we were unable to locate the appendix by visiting either author's personal websites. As part of the demonstration of the SimDesign package, C&A included code that replicates the simulation used to produce Table 6 of F&C; we drew from their code in writing our own simulation code. C&A also reported tau values used to transform continuous data to ordinal data, which were retrieved from Flora's unpublished dissertation but not reported in F&C.

2.2 Data Generating Mechanism

Information provided in the above mentioned sources indicated that the following simulation factors were systematically varied in generating the artificial data.

Simulation factor	No. levels	Levels
Model estimation method	2	Full weighted least squares (WLS), robust WLS
Number of factors	2	1, 2
Sample size	4	100, 200, 500, 1000
Number of indicators per latent factor	2	5, 10
number of categories in ordinal data	2	2,5

Simulation factor	No. levels	Levels
Deviation from normality (skewness and kurtosis)	5	Normal, Low skewness vs. low kurtosis, Low skewness vs. moderate kurtosis, Moderate skewness vs. low kurtosis, Moderate skewness vs. moderate kurtosis
Factor loadings	1	.70
Residual variances of indicators	1	.51
Interfactor correlation (only for models with 2 latent factors)	1	.30

2.2.1 Model Estimation Method

Model estimation method was defined using the 'estimator' argument of the 'cfa' function in the 'lavaan' package, with either "WLS" (for full WLS) or "WLSMV" (for robust WLS).

2.2.2 Number of Latent Factors

Number of latent factors varied between the CFA models. Models 1 and 2 each had one latent factor; models 3 and 4 each had two latent factors.

2.2.3 Sample Size

Random samples of four different sizes were generated: 100, 200, 500, and 1,000.

2.2.4 Number of Indicators Per Latent Factor

Number of indicators per latent factor varied between the CFA models. In models 1 and 3, each latent factor was measured by five indicators; in models 2 and 4, each latent factor was measured by 10 indicators. Thus, the four models were:

- Model 1: 1 latent factor, 5 indicators total
- Model 2: 2 latent factors, 10 indicators total
- Model 3: 1 latent factor, 10 indicators total
- Model 4: 2 latent factors, 20 indicators total.

2.2.5 Number of Categories in Ordinal Data

Number of categories in ordinal data reflects the number of unique values in the ordinal observed distributions that were transformed from continuous data. In two-category conditions, the ordinal data had two unique values (defined as 0 or 1); in five-category conditions, the ordinal data had five unique values (defined as 0, 1, 2, 3, and 4).

2.2.6 Deviation From Normality (Skewness and Kurtosis)

Deviation from normality was defined as a combination of non-zero skewness and non-zero kurtosis. The conditions are as follows:

Condition	Skewness	Kurtosis
Normal	0	0
Low skewness vs. low kurtosis	0.75	1.75
Low skewness vs. moderate kurtosis	0.75	3.75
Moderate skewness vs. low kurtosis	1.25	1.75
Moderate skewness vs. moderate kurtosis	1.25	3.75

2.2.7 Fixed Values

Factor loadings represent the standardized regression coefficients of the latent response variables (i.e., the continuous variables pre-transformation) on the latent factors, and all factor loadings were set to .70. All latent responses variables were standardized to have means of zero and unit variances; thus, the residual variance of each was set to 1 - .7^2 = .51. The latent factors were also set to have unit variances. For models with two latent factors (Models 2 and 4), the interfactor correlation was set to 0.30. Each condition of the simulations was iterated 500 times.

2.3 Compared Methods

The study compares two methods of fitting confirmatory factor analysis (CFA) models to estimate polychoric correlations among ordinal variables (e.g., Likert-type items) across different conditions encountered by applied researchers. The statistical performance of the two methods, fully weighted least squares (WLS) and robust WLS, was compared across various levels of nonnormality of the underlying continuous latent distribution.

2.3.1 Full WLS

The WLS approach was developed for estimating a weighted matrix based on the asymptotic variances and covariances of polychoric correlations that can be used in conjunction with a matrix of polychoric correlations in the estimation of SEM models (e.g., Browne, 1982, 1984; Jöreskog, 1994; B. Muthén, 1984; B. O. Muthén & Satorra, 1995). WLS applies the fitting function:

$$F_{WLS} = [s - \sigma(\theta)]' W^{-1} [s - \sigma(\theta)]$$

In our replication, we implemented full WLS with 'estimator = "WLS" ' argument of the 'cfa' function in the 'lavaan' package.

2.3.2 Robust WLS

The robust WLS approach was developed by B. Muthén et al. (1997) to address the problems faced when using WLS with small to medium samples. Built on the works of Satorra and colleagues (Chou et al., 1991; Satorra, 1992; Satorra & Bentler, 1990), the robust WLS approach obtains parameter estimates by substituting a diagonal matrix, V, for W in the full WLS approach. The diagonal matrix contains the asymptotic variances of the thresholds and polychoric correlation estimates. Calculation of this matrix involves the full weight matrix W (as in the above equation); however, it does not need to be inverted. A robust goodness-of-fit test can be calculated with a mean- and variance-adjusted chi-square test statistic, which similarly involves the full weight matrix W but avoids inversion.

2.4 Performance measures

The primary performance measure was the mean relative bias (RB), defined as:

$$RB = (\frac{\hat{\theta} - \theta}{\theta}) * 100$$

Where $\hat{\theta}$ is the estimated statistic from a given replication, and θ is the corresponding population parameter. Following F&C, we calculated the mean RB across replications within a given condition. We focused on three estimated statistics: chi-square test statistics, polychoric correlations, and factor loadings. For the χ^2 test statistics from full WLS estimation, the RB was calculated with respect to the degree of freedom of each model. For the factor loadings and the polychoric correlations, the RB was calculated with θ as the population value, and $\hat{\theta}$ as the pooled mean of the statistic:

$$Pooled\, Mean = P^{-1} \sum_{i=1}^{P} \bar{\hat{\lambda}}_i$$

In addition, the square root of the mean of the statistic variances was calculated to quantify the variance of the mean statistic estimates:

$$Pooled\, SD = \sqrt{P^{-1} \sum_{i=1}^{P} VAR(\hat{\lambda}_i)}$$

Note that although F&C further considered the RB of standard error estimation (p. 474), we did not examine it in our replication because F&C did not fully report the results of that metric, thus preventing us from being able to directly compare our results to theirs (the link to the technical appendix that would include results on that metric was broken). In addition to RB, rates of improper solutions were assessed as a secondary performance measure. An improper solution was defined as "a nonconverged solution or a solution that converged but resulted in one or more out-of-bound parameters (e.g., Heywood cases)" (p. 473, F&C).

2.5 Technical implementation

While the original simulation study was carried out in EQS (for data generation) and Mplus (for data analysis), 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 GitHub.

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 were implemented.

Issue	Replicator decision	Justification
Lack of access to the original software environment	Conducted the replication in R	Success of the replication should not depend on the software used
Tau values not reported or misreported in the original article	Followed the tau values reported in Chalmers and Adkins (2020)	Tau values for transforming continuous data into two-category ordinal data were not explicitly reported; tau values for five-category ordinal data produced distributions inconsistent with the original article
Some results were not readily available in the primary replication	Ran 2 additional simulation studies for Tables 2 and 3, respectively	Results from Tables 2 and 3 were not directly available from the replication of the primary simulation study
Too many models fail to converge in some conditions	Decided to omit those conditions from the replication	Could not obtain parameter estimates to assess estimator performance because of nonconverging models; consistent with decision in C&A

2.6 Lack of Access to the Original Software Environment

The original simulation studies reported in F&C used EQS (Version 5.7b; Bentler, 1995) for data generation and Mplus (Version 2.01; L. K. Muthén & Muthén, 1998) for data analysis. Because the script for these software environments was unavailable to us (see Section 2.1 above), and because we did not have access to EQS, a proprietary software, we decided to implement both data generation and data analysis in R (see

Appendix for details). Specifically, we relied on two core packages, lavaan and SimDesign, for the bulk of our replication efforts.

2.7 Tau Values not Reported or Misreported in the Original Article

Tau values are thresholds used for transforming continuous data into ordinal data. F&C generated ordinal data from continuous data with various degrees of nonnormality; however, the tau values used to do so were not directly reported in the original article. For two-category ordinal data, the tau value was not reported. Although one can reasonably assume that the tau value for transforming normal, standardized distribution into ordinal data is 0 (the mean of such a distribution), it is unclear whether the tau value for nonnormal distribution should still be 0 or not. Based on C&A, we chose tau = 0 for all transformations from continuous data to two-category ordinal data. For five-category ordinal data, F&C referred to tau values reported in Muthén and Kaplan (1985) in Footnote 9, noting that "[t]his same threshold set was also applied to each nonnormal y*distribution." Yet C&A suggested that this was not accurate: For the moderate skewness vs. low kurtosis condition (skewness = 1.25, kurtosis = 1.75), the first threshold was not -1.645 (p. 176; Muthén & Kaplan, 1985), but -1.125. C&A noted that this modification was included in Flora's (2002) unpublished dissertation, which we do not have access to. We confirmed that this modification was likely made by F&C: Using the modified threshold gave us results that closely matched those of F&C for that condition in Table 1, but using the threshold reported in Muthén and Kaplan (1985) did not. Therefore, we decided to use the modified tau values as reported by C&A in our replications. (The tau values are presumably reported in the original script; as we note in Section 2.1, however, we did not have access to the original script and thus were unable to confirm them.)

2.8 Some Results Were not Readily Available in the Primary Replication

F&C reported one primary simulation study, the results from which were reported in Tables 2-11 (results reported in Table 1 was from a separate simulation study, for which we conducted a separate replication). We however found that the results reflecting those reported in Tables 2 and 3 were not readily available in our primary replication conducted in SimDesign. In Table 2, F&C examined the accuracy of polychoric correlation estimates by generating bivariate data for each of two population correlation values, before estimating CFA models on those polychoric correlations. In contrast, because our primary replication in SimDesign generated data directly from the CFA models, the polychoric correlation estimates were not directly available in the output. Therefore, we conducted a separate replication to examine the accuracy of the polychoric correlation estimates. In Table 3, F&C examined the rate of improper solutions at small sample sizes (N = 100 and N = 200). Improper solutions were defined as model solutions that either do not converge or contain out-of-bound estimates (i.e., Heywood cases). The safeguarding features of SimDesign prevented us from running conditions in which a high rate of nonconvergence occurred. Therefore, we conducted a separate replication to examine the rate of improper solutions at small sample sizes without using SimDesign.

2.9 Too Many Models Fail to Converge in Some Conditions

F&C reported results for Model 4 at N = 200 using full WLS in Tables 7 and 11. In our replication, however, we found that the rate of nonconvergence for these conditions were too high (see Table 3 of our replication below). Because we could not obtain model fit statistics or parameter estimates from nonconverging models, we excluded these conditions from our replication results reported below.

3 Results

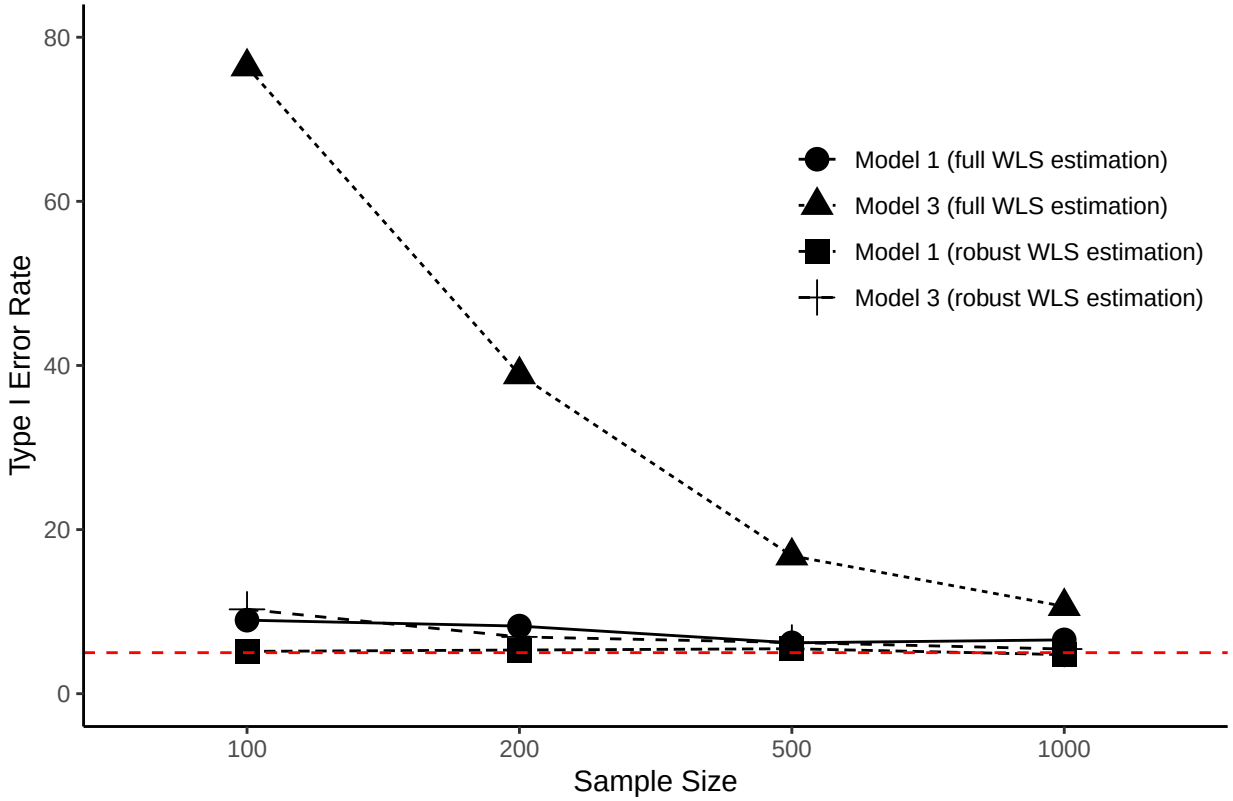

FIGURE 1: Generate summary data for Figure 6 in F&C

TABLE 1: Skewness and Kurtosis of Univariate Latent Response, y* Distributions, and Five-Category Ordinal,y Distributions

Condition	y* Skewness	y* Kurtosis	y Skewness	y Kurtosis
Normal	0.00	0.00	0.00	-0.01
Low skewness vs. low kurtosis	0.75	1.75	0.31	0.21
Low skewness vs. moderate kurtosis	0.75	3.75	0.23	0.59
Moderate skewness vs. low kurtosis	1.25	1.75	0.50	-0.07
Moderate skewness vs. moderate kurtosis 1	1.25	3.75	0.48	0.35

TABLE 2: Means, Standard Deviations, and Relative Bias of Polychoric Correlation Estimates

N	s	k	M(rho=.147/2cat)	SD(rho=.147/2cat)	RB(rho=.147/2cat)	M(rho=.147/5cat)	SD(rho=.147/5cat)	RB(rho=.147/5cat)	M(rho=.49/2cat)	SD(rho=.49/2cat)	RB(rho=.49/2cat)	M(rho=.49/5cat)	SD(rho=.49/5cat)	RB(rho=.49/5cat)
100	0.00	0.00	0.15	0.16	0.63	0.14	0.11	-2.79	0.49	0.12	-0.22	0.49	0.09	0.98
100	0.75	1.75	0.15	0.15	3.10	0.16	0.12	6.70	0.50	0.13	1.42	0.50	0.09	1.27
100	0.75	3.75	0.14	0.15	-3.19	0.15	0.11	4.72	0.50	0.13	2.46	0.51	0.10	4.85
100	1.25	1.75	0.16	0.15	9.36	0.15	0.11	2.59	0.51	0.12	4.95	0.50	0.09	1.66
100	1.25	3.75	0.15	0.15	4.44	0.16	0.11	8.14	0.51	0.13	4.37	0.52	0.09	6.68
200	0.00	0.00	0.15	0.11	0.57	0.15	0.08	3.04	0.48	0.10	-2.58	0.49	0.06	0.14
200	0.75	1.75	0.15	0.11	3.30	0.15	0.08	2.21	0.50	0.09	2.16	0.50	0.06	1.12
200	0.75	3.75	0.16	0.11	7.63	0.15	0.08	0.49	0.50	0.09	2.70	0.50	0.06	1.75
200	1.25	1.75	0.17	0.10	14.07	0.15	0.08	3.51	0.52	0.09	5.86	0.49	0.06	0.48
200	1.25	3.75	0.16	0.11	5.99	0.15	0.08	2.68	0.50	0.09	1.79	0.51	0.07	4.03
500	0.00	0.00	0.14	0.07	-2.64	0.14	0.05	-1.46	0.49	0.06	-0.54	0.49	0.04	-0.03
500	0.75	1.75	0.15	0.07	2.00	0.15	0.05	4.12	0.50	0.06	1.67	0.49	0.04	0.97
500	0.75	3.75	0.15	0.07	1.00	0.15	0.05	3.57	0.51	0.06	3.59	0.51	0.04	3.36
500	1.25	1.75	0.16	0.07	7.74	0.15	0.05	2.71	0.52	0.06	5.68	0.50	0.04	1.02
500	1.25	3.75	0.15	0.07	-0.25	0.16	0.05	7.56	0.51	0.06	3.71	0.51	0.04	3.81
1000	0.00	0.00	0.15	0.05	-0.22	0.15	0.04	-0.61	0.49	0.04	-0.24	0.49	0.03	-0.01
1000	0.75	1.75	0.15	0.05	0.57	0.15	0.03	2.71	0.50	0.04	1.54	0.50	0.03	1.63
1000	0.75	3.75	0.15	0.05	4.28	0.15	0.04	4.88	0.50	0.04	2.80	0.51	0.03	3.09
1000	1.25	1.75	0.16	0.05	9.17	0.15	0.03	3.87	0.52	0.04	5.95	0.50	0.03	1.18
1000	1.25	3.75	0.15	0.05	4.97	0.16	0.04	7.04	0.51	0.04	4.17	0.51	0.03	3.70

TABLE 3: Rates of Improper Solutions Obtained With Full WLS Estimation

N	s	k	2cat:imp.mod1	2cat:nc.mod1	2cat:imp.mod2	2cat:nc.mod2	2cat:imp.mod3	2cat:nc.mod3	2cat:imp.mod4	2cat:nc.mod4	5cat:imp.mod1	5cat:nc.mod1	5cat:imp.mod2	5cat:nc.mod2	5cat:imp.mod3	5cat:nc.mod3	5cat:imp.mod4	5cat:nc.mod4
100	0	0	2.2	0	26.6	0.0	36.2	1.0	100	100	0.8	0.8	2.4	2.2	1.8	0.4	100	100
100	0.75	1.75	1.4	0	27.8	0.2	41.0	0.6	100	100	5.0	5.0	19.8	19.2	6.2	5.2	100	100
100	0.75	3.75	1.4	0	30.4	0.0	38.6	0.0	100	100	3.0	3.0	12.6	12.4	4.4	3.2	100	100
100	1.25	1.75	2.2	0	38.6	0.0	47.4	0.4	100	100	2.6	2.6	10.2	10.0	6.0	2.8	100	100
100	1.25	3.75	1.6	0	31.6	0.0	40.6	0.4	100	100	12.6	12.6	47.6	47.4	18.4	15.4	100	100
200	0	0	0.0	0	0.0	0.0	0.4	0.0	100	100	0.0	0.0	0.0	0.0	0.0	0.0	100	100
200	0.75	1.75	0.0	0	0.0	0.0	0.6	0.0	100	100	0.2	0.2	0.4	0.4	0.0	0.0	100	100
200	0.75	3.75	0.0	0	0.0	0.0	1.0	0.0	100	100	0.0	0.0	0.2	0.2	0.2	0.2	100	100
200	1.25	1.75	0.0	0	0.0	0.0	0.4	0.0	100	100	0.0	0.0	0.0	0.0	0.0	0.0	100	100
200	1.25	3.75	0.0	0	0.2	0.0	0.4	0.0	100	100	1.2	1.2	6.4	6.4	1.6	1.6	100	100

Note:

imp = rate of improper solutions.

TABLE 4: Chi-Square Test Statistics for Model 1 (Five Indicators, One Factor)

N	s	k	WLS M	WLS SD	WLS RB	WLS % reject	Robust WLS RB	Robust WLS % reject
100	0.00	0.00	5.75	3.91	15.01	9.4	1.81	6.0
100	0.75	1.75	5.54	3.88	10.71	8.4	9.06	4.8
100	0.75	3.75	5.66	3.91	13.29	11.2	-1.41	4.0
100	1.25	1.75	6.03	3.75	20.59	9.2	4.13	5.2
100	1.25	3.75	5.34	3.60	6.87	6.6	-2.06	5.8
200	0.00	0.00	5.33	3.43	6.53	7.2	-0.41	3.6
200	0.75	1.75	5.52	3.82	10.47	9.2	1.37	5.6
200	0.75	3.75	5.18	3.49	3.60	7.0	-4.61	3.2
200	1.25	1.75	5.74	3.87	14.74	8.6	13.57	8.2
200	1.25	3.75	5.84	3.99	16.79	9.2	6.59	6.0
500	0.00	0.00	5.17	3.47	3.43	6.8	-2.73	4.2
500	0.75	1.75	4.98	3.37	-0.32	5.2	0.91	5.8
500	0.75	3.75	4.84	2.94	-3.14	3.8	-2.79	5.0
500	1.25	1.75	5.26	3.44	5.29	7.2	6.16	7.0
500	1.25	3.75	5.50	3.50	10.00	8.0	1.55	5.4
1000	0.00	0.00	5.10	3.12	1.94	5.2	-1.13	4.0
1000	0.75	1.75	5.18	3.42	3.61	7.2	-5.19	4.0
1000	0.75	3.75	5.14	3.51	2.83	5.8	0.29	4.0
1000	1.25	1.75	5.68	3.62	13.52	8.2	5.68	6.8
1000	1.25	3.75	5.16	3.50	3.20	6.4	-1.14	5.0

TABLE 5: Chi-Square Test Statistics for Model 2 (10 Indicators, One Factor)

N	s	k	WLS M	WLS SD	WLS RB	WLS % reject	Robust WLS RB	Robust WLS % reject
100	0.00	0.00	57.50	18.15	64.28	62.2	6.53	5.2
100	0.75	1.75	57.71	19.48	64.89	62.2	5.26	7.4
100	0.75	3.75	59.69	18.50	70.54	66.8	7.34	8.8
100	1.25	1.75	61.43	20.01	75.51	69.4	9.23	8.6
100	1.25	3.75	58.77	18.93	67.91	63.4	6.36	7.4
200	0.00	0.00	44.47	12.07	27.07	30.4	4.02	6.8
200	0.75	1.75	44.57	12.52	27.35	27.8	2.41	4.8
200	0.75	3.75	44.61	12.78	27.46	30.4	2.14	5.6
200	1.25	1.75	47.78	13.69	36.52	39.4	9.77	9.2
200	1.25	3.75	45.08	12.10	28.80	32.2	3.12	5.6
500	0.00	0.00	37.53	9.24	7.22	9.6	2.43	7.0
500	0.75	1.75	38.75	10.34	10.71	14.2	1.49	5.2
500	0.75	3.75	38.45	9.81	9.86	11.6	-0.69	4.6
500	1.25	1.75	41.92	10.98	19.77	22.0	7.68	7.2
500	1.25	3.75	39.27	9.96	12.21	15.0	1.09	5.0
1000	0.00	0.00	36.27	8.71	3.62	7.2	0.26	6.0
1000	0.75	1.75	37.49	8.82	7.11	8.2	0.99	4.0
1000	0.75	3.75	37.19	9.63	6.25	11.0	2.02	5.2
1000	1.25	1.75	39.39	9.66	12.56	12.6	6.31	8.4
1000	1.25	3.75	37.18	9.11	6.22	9.4	-1.95	3.8

TABLE 6: Chi-Square Test Statistics for Model 3 (10 Indicators, Two Correlated Factors)

N	s	k	WLS M	WLS SD	WLS RB	WLS % reject	Robust WLS RB	Robust WLS % reject
100	0.00	0.00	64.08	21.15	88.48	73.6	9.82	11.8
100	0.75	1.75	65.10	22.35	91.47	76.0	11.17	13.2
100	0.75	3.75	67.22	23.86	97.70	77.2	9.29	8.8
100	1.25	1.75	63.75	19.66	87.49	79.6	7.04	8.0
100	1.25	3.75	63.42	21.13	86.53	75.6	11.23	9.6
200	0.00	0.00	45.79	13.00	34.69	38.2	5.18	5.2
200	0.75	1.75	46.40	13.97	36.48	39.2	6.77	10.0
200	0.75	3.75	46.90	13.78	37.95	40.4	4.39	5.8
200	1.25	1.75	46.08	13.99	35.53	36.0	2.94	6.0
200	1.25	3.75	47.30	14.43	39.10	40.6	6.42	7.6
500	0.00	0.00	38.16	10.37	12.23	15.2	1.06	4.8
500	0.75	1.75	38.68	10.16	13.77	16.8	-0.30	5.6
500	0.75	3.75	39.29	9.76	15.57	16.4	2.23	6.2
500	1.25	1.75	38.53	10.60	13.31	17.4	3.25	7.0
500	1.25	3.75	38.51	10.57	13.25	18.2	3.01	7.6
1000	0.00	0.00	36.19	9.08	6.44	10.4	0.24	4.8
1000	0.75	1.75	36.94	9.75	8.64	10.6	2.20	5.8
1000	0.75	3.75	36.58	9.08	7.59	10.6	-1.36	4.8
1000	1.25	1.75	36.98	9.23	8.76	10.4	0.73	6.0
1000	1.25	3.75	36.06	9.11	6.05	11.2	0.77	5.8

TABLE 7: Chi-Square Test Statistics for Model 4 (20 Indicators, Two Correlated Factors)

N	s	k	WLS M	WLS SD	WLS RB	WLS % reject	Robust WLS RB	Robust WLS % reject
100	0.00	0.00	NA	NA	NA	NA	7.13	10.2
100	0.75	1.75	NA	NA	NA	NA	8.17	13.0
100	0.75	3.75	NA	NA	NA	NA	7.50	12.0
100	1.25	1.75	NA	NA	NA	NA	6.86	10.2
100	1.25	3.75	NA	NA	NA	NA	7.78	12.2
200	0.00	0.00	NA	NA	NA	NA	4.41	8.0
200	0.75	1.75	NA	NA	NA	NA	4.39	7.2
200	0.75	3.75	NA	NA	NA	NA	5.07	9.4
200	1.25	1.75	NA	NA	NA	NA	4.91	10.0
200	1.25	3.75	NA	NA	NA	NA	4.56	7.8
500	0.00	0.00	273.13	35.55	61.61	98.8	1.02	3.2
500	0.75	1.75	276.96	40.14	63.88	98.4	1.25	5.8
500	0.75	3.75	280.14	39.88	65.76	97.8	1.90	5.8
500	1.25	1.75	276.66	38.17	63.70	98.2	2.62	6.8
500	1.25	3.75	278.76	39.99	64.94	99.2	2.28	9.2
1000	0.00	0.00	210.63	25.88	24.63	64.8	1.08	7.6
1000	0.75	1.75	210.68	25.32	24.66	66.4	0.69	5.6
1000	0.75	3.75	213.58	27.08	26.38	64.0	0.24	6.6
1000	1.25	1.75	215.01	25.81	27.23	72.8	1.59	6.0
1000	1.25	3.75	210.53	25.69	24.58	63.6	0.62	5.8

TABLE 8: Mean Factor Loadings for Model 1 (Five Indicators, One Factor)

N	s	k	WLS pooled M	WLS pooled SD	WLS RB	Robust WLS pooled M	Robust WLS pooled SD	Robust WLS RB
100	0.00	0.00	0.71	0.08	1.88	0.70	0.08	0.60
100	0.75	1.75	0.72	0.08	3.32	0.71	0.08	1.18
100	0.75	3.75	0.73	0.08	4.12	0.72	0.08	2.44
100	1.25	1.75	0.72	0.08	3.56	0.71	0.08	1.12
100	1.25	3.75	0.73	0.08	4.09	0.72	0.08	2.92
200	0.00	0.00	0.71	0.05	1.57	0.70	0.05	0.13
200	0.75	1.75	0.72	0.06	2.45	0.71	0.06	0.95
200	0.75	3.75	0.72	0.06	3.30	0.71	0.05	1.78
200	1.25	1.75	0.72	0.05	2.29	0.70	0.05	0.62
200	1.25	3.75	0.73	0.05	3.64	0.72	0.05	2.35
500	0.00	0.00	0.70	0.03	0.67	0.70	0.03	0.07
500	0.75	1.75	0.71	0.03	1.37	0.71	0.03	1.04
500	0.75	3.75	0.71	0.04	2.06	0.71	0.03	1.50
500	1.25	1.75	0.71	0.03	1.08	0.70	0.03	0.65
500	1.25	3.75	0.72	0.03	2.65	0.71	0.03	1.81
1000	0.00	0.00	0.70	0.02	0.41	0.70	0.02	0.15
1000	0.75	1.75	0.71	0.02	1.07	0.71	0.02	0.88
1000	0.75	3.75	0.71	0.02	1.63	0.71	0.02	1.46
1000	1.25	1.75	0.71	0.02	0.75	0.70	0.02	0.59
1000	1.25	3.75	0.72	0.02	2.22	0.71	0.02	1.99

TABLE 9: Mean Factor Loadings for Model 2 (10 Indicators, One Factor)

N	s	k	WLS pooled M	WLS pooled SD	WLS RB	Robust WLS pooled M	Robust WLS pooled SD	Robust WLS RB
100	0.00	0.00	0.78	0.08	11.22	0.70	0.07	0.49
100	0.75	1.75	0.78	0.08	11.43	0.71	0.07	1.86
100	0.75	3.75	0.78	0.08	11.83	0.72	0.07	2.37
100	1.25	1.75	0.78	0.08	10.79	0.71	0.07	1.39
100	1.25	3.75	0.79	0.08	12.99	0.72	0.07	2.93
200	0.00	0.00	0.75	0.05	6.68	0.70	0.05	0.37
200	0.75	1.75	0.75	0.05	7.39	0.71	0.05	0.87
200	0.75	3.75	0.76	0.05	8.46	0.71	0.05	1.77
200	1.25	1.75	0.74	0.05	6.36	0.71	0.05	1.09
200	1.25	3.75	0.76	0.05	8.65	0.71	0.05	2.13
500	0.00	0.00	0.72	0.03	2.89	0.70	0.03	0.23
500	0.75	1.75	0.73	0.03	3.68	0.71	0.03	0.99
500	0.75	3.75	0.73	0.03	4.34	0.71	0.03	1.56
500	1.25	1.75	0.72	0.03	3.10	0.70	0.03	0.54
500	1.25	3.75	0.73	0.03	4.84	0.71	0.03	2.10
1000	0.00	0.00	0.71	0.02	1.43	0.70	0.02	-0.01
1000	0.75	1.75	0.72	0.02	2.26	0.71	0.02	0.81
1000	0.75	3.75	0.72	0.02	2.99	0.71	0.02	1.56
1000	1.25	1.75	0.71	0.02	1.91	0.70	0.02	0.52
1000	1.25	3.75	0.72	0.02	3.33	0.71	0.02	2.02

TABLE 10: Mean Parameter Estimates for Model 3 (10 Indicators, Two Factors)

N	s	k	WLS lambda M	WLS lambda SD	WLS lambda RB	WLS phi21 M	WLS phi21 SD	WLS phi21 RB	Robust lambda M	Robust lambda SD	Robust lambda RB	Robust phi21 M	Robust phi21 SD	Robust phi21 RB
100	0.00	0.00	0.68	0.12	-2.24	0.35	0.19	16.17	0.64	0.09	-8.65	0.31	0.14	3.93
100	0.75	1.75	0.69	0.12	-1.90	0.36	0.20	20.54	0.65	0.09	-7.42	0.31	0.14	2.30
100	0.75	3.75	0.69	0.12	-1.41	0.36	0.20	21.16	0.65	0.09	-7.33	0.32	0.13	6.86
100	1.25	1.75	0.70	0.12	-0.65	0.37	0.20	24.86	0.65	0.09	-7.31	0.32	0.13	6.43
100	1.25	3.75	0.70	0.12	-0.64	0.35	0.20	15.42	0.65	0.09	-6.51	0.33	0.13	8.74
200	0.00	0.00	0.66	0.07	-5.04	0.33	0.11	10.93	0.64	0.07	-8.78	0.31	0.10	3.33
200	0.75	1.75	0.67	0.07	-4.43	0.32	0.11	7.76	0.64	0.07	-7.86	0.31	0.09	3.18
200	0.75	3.75	0.67	0.07	-3.58	0.33	0.11	10.92	0.65	0.07	-7.25	0.31	0.09	3.82
200	1.25	1.75	0.67	0.07	-3.74	0.36	0.11	18.79	0.65	0.06	-7.60	0.31	0.09	4.44
200	1.25	3.75	0.68	0.07	-3.00	0.34	0.10	14.22	0.65	0.07	-6.78	0.31	0.09	3.35
500	0.00	0.00	0.65	0.04	-7.40	0.32	0.06	5.27	0.64	0.04	-8.98	0.30	0.06	0.56
500	0.75	1.75	0.65	0.04	-6.56	0.32	0.06	5.75	0.64	0.04	-8.20	0.30	0.06	0.20
500	0.75	3.75	0.66	0.04	-5.87	0.32	0.06	6.57	0.65	0.04	-7.46	0.31	0.06	3.01
500	1.25	1.75	0.66	0.04	-6.08	0.33	0.06	9.12	0.65	0.04	-7.46	0.31	0.05	3.36
500	1.25	3.75	0.66	0.04	-5.49	0.32	0.06	5.46	0.65	0.04	-7.10	0.31	0.06	4.45
1000	0.00	0.00	0.64	0.03	-8.16	0.30	0.04	1.65	0.64	0.03	-8.95	0.30	0.04	0.77
1000	0.75	1.75	0.65	0.03	-7.32	0.31	0.04	3.24	0.64	0.03	-8.12	0.30	0.04	1.37
1000	0.75	3.75	0.65	0.03	-6.69	0.31	0.04	2.21	0.65	0.03	-7.46	0.30	0.04	0.88
1000	1.25	1.75	0.65	0.03	-6.97	0.32	0.04	5.99	0.65	0.03	-7.61	0.31	0.04	3.11
1000	1.25	3.75	0.66	0.03	-6.21	0.31	0.04	4.08	0.65	0.03	-6.92	0.31	0.04	2.17

TABLE 11: Mean Parameter Estimates for Model 4 (20 Indicators, Two Factors)

N	s	k	WLS lambda M	WLS lambda SD	WLS lambda RB	WLS phi21 M	WLS phi21 SD	WLS phi21 RB	Robust lambda M	Robust lambda SD	Robust lambda RB	Robust phi21 M	Robust phi21 SD	Robust phi21 RB
100	0.00	0.00	NA	NA	NA	NA	NA	NA	0.64	0.08	-8.41	0.32	0.12	5.24
100	0.75	1.75	NA	NA	NA	NA	NA	NA	0.65	0.08	-7.65	0.32	0.11	5.99
100	0.75	3.75	NA	NA	NA	NA	NA	NA	0.65	0.08	-7.04	0.32	0.11	6.76
100	1.25	1.75	NA	NA	NA	NA	NA	NA	0.65	0.08	-7.19	0.32	0.11	5.53
100	1.25	3.75	NA	NA	NA	NA	NA	NA	0.65	0.08	-6.46	0.32	0.11	5.39
200	0.00	0.00	NA	NA	NA	NA	NA	NA	0.64	0.06	-8.72	0.30	0.08	0.51
200	0.75	1.75	NA	NA	NA	NA	NA	NA	0.64	0.06	-8.06	0.31	0.07	3.70
200	0.75	3.75	NA	NA	NA	NA	NA	NA	0.65	0.06	-7.34	0.31	0.08	3.33
200	1.25	1.75	NA	NA	NA	NA	NA	NA	0.65	0.05	-7.36	0.31	0.08	3.94
200	1.25	3.75	NA	NA	NA	NA	NA	NA	0.65	0.06	-6.85	0.31	0.08	4.77
500	0.00	0.00	0.68	0.04	-3.08	0.35	0.07	17.02	0.64	0.04	-8.90	0.30	0.05	0.81
500	0.75	1.75	0.69	0.04	-2.13	0.36	0.07	19.62	0.64	0.03	-8.07	0.31	0.05	2.65
500	0.75	3.75	0.69	0.04	-1.69	0.35	0.06	17.83	0.65	0.04	-7.46	0.31	0.05	1.99
500	1.25	1.75	0.69	0.04	-1.67	0.36	0.07	21.02	0.65	0.03	-7.52	0.31	0.05	4.18
500	1.25	3.75	0.69	0.04	-1.00	0.37	0.07	22.18	0.65	0.03	-6.97	0.31	0.05	2.59
1000	0.00	0.00	0.66	0.03	-5.99	0.32	0.04	8.10	0.64	0.03	-8.94	0.30	0.04	-0.09
1000	0.75	1.75	0.66	0.03	-5.08	0.33	0.04	9.87	0.64	0.02	-8.17	0.30	0.03	1.51
1000	0.75	3.75	0.67	0.03	-4.50	0.33	0.04	9.09	0.65	0.03	-7.56	0.30	0.03	0.67
1000	1.25	1.75	0.67	0.03	-4.62	0.34	0.04	13.05	0.65	0.02	-7.75	0.31	0.03	3.56
1000	1.25	3.75	0.67	0.03	-4.02	0.33	0.04	10.69	0.65	0.02	-6.99	0.31	0.03	2.39

4 Discussion

4.1 Replicability

The replication process was mostly straightforward. F&C had provided almost all of the information required to execute our replication. C&A provided further insight into replicating the original simulation in R using the SimDesign package, including sample R code that became the base script of our primary replication study. Together, these two sources were informative in conducting our replication. The key challenges were: (1) The links to the technical appendix and the code in F&C were broken, so we could not directly compare our simulation code to theirs; (2) some information, such as tau values used to transform continuous data to categorical data, was not directly reported in the paper; and (3) in part because we used a different software environment, we decided to conduct additional simulation studies to fully assess the original results reported in F&C.

4.2 Replicator degrees of freedom

Our replication could benefit from additional information that was not available to us at the time of replication, including the exact tau values used to transform continuous data into categorical ordinal data. As discussed above, they did not report the tau value used for the two-category conditions, and although they referred to the source of tau values used for the five-category conditions in a Footnote, they did not disclose whether those values were used in all of those conditions. Based on C&A and our own replication attempt, we believe that the first threshold in the conditions with moderate skewness and low kurtosis was altered. In addition, access to the original appendix and code would have helped reduce our replicator degrees of freedom. To be fair to the authors, however, sharing code via permalink was uncommon and few viable solutions were available when the paper was published. We do not believe that the replicator degrees of freedom are so intensive that they should influence the results.

4.3 Equivalence of results

Results from our simulation were overall consistent with the original simulation results, and we would draw similar conclusions as the original authors. First, we almost exactly replicated the skewness and kurtosis of the five-category ordinal distributions (Table 1 of F&C), and the means, SDs, and RB of polychoric correlation estimates were comparable to the original estimates as well (Table 2 of F&C). Next, we observed the same pattern of chi-square test statistics across all four models (i.e., means, SDs, RB, and Type I error rates; Tables 4–7 of F&C), except for one set of conditions (N = 200 for Model 4 with full WLS; see discussion below). The orders of magnitude were comparable, and the trends were in the same direction. On parameter estimates, we also observed largely similar patterns of factor loading estimates across full WLS and robust WLS for Models 1 and 2 (with one factor), and the impact of sample size, skewness and kurtosis on those estimates

followed the same pattern as the original simulation study as well; likewise, results on the interfactor correlations for Models 3 and 4 were also highly consistent with those in F&C.

We did observe three sets of notable differences. First, the rates of improper solutions obtained with full WLS at small sample sizes (N = 100 or 200) in our replication were often higher than those in the original study, largely due to nonconvergence (Table 3). Although our rates of improper solutions matched those in F&C for Models 1–3 with two-category data, we observed 100% nonconvergence for Model 4 with two-category data; in contrast, F&C found 65–90% of the solutions were improper, with 5–35% due to nonconvergence. In addition, although F&C found minimal rates of improper solutions (0–2%) across levels of nonnormality for Models 1–3 with five-category data, we found that increasing levels of skewness and kurtosis had major impact on rates of improper solutions for those models. Specifically, as skewness and kurtosis increased, so did the rate of improper solutions, reaching 12.6%, 47.6%, and 18.4% for Models 1–3, respectively. Lastly, for Model 4 with five-category data, F&C found 100% nonconvergence at *N*= 100 and around 20–30% improper solutions (with 0% nonconvergence) at *N*= 200; in contrast, we found 100% nonconvergence at both_N_= 100 and *N*= 200. Because of the nonconvergence at *N*= 200 for Model 4 with five-category data using full WLS, we could not assess the performance of full WLS in those conditions (see our replicated Tables 7 and 11).

Second, in Models 3 and 4 (both with two correlated factors) with five-category data, the factor loading estimates consistently showed negative RB (i.e., the estimates were lower than the population value of .7), whereas those in F&C consistently showed positive RB.

Third, the magnitude and pattern of RB on parameter estimates also differed somewhat from F&C, particularly on the factor loading estimates for Models 3 and 4 (i.e., models with two factors). Whereas F&C found that RB of parameter estimates increased with increasing model size, particularly with full WLS estimation, our replication found only sporadic evidence for this pattern (RB of interfactor correlation estimates was greater in the bigger Model 4 vs. the smaller Model 3; RB of factor loadings using robust WLS estimation was greater in Models 3 and 4 vs. Models 1 and 2; however, RB of factor loadings using full WLS estimation was mostly unrelated to model size). In addition, F&C noted that RB in parameter estimates was notably smaller with robust vs. full WLS; in our replication, this was not the case for factor loading estimates in Models 3 and 4 (if anything, the robust WLS estimation performed worse, at trivial to moderate levels of RB, rather than at trivial levels of RB as F&C observed).

It is unclear to us why these discrepancies occurred. One possible explanation for these discrepancies is the different software environments used for data generation analysis and data analysis. A further investigation into these discrepancies might be warranted. We do not believe these discrepancies crucially undermine the conclusions of F&C. Some discrepancies, assuming they reflect substantive differences, might be worth keeping in mind, especially for those using R and lavaan for data analysis (e.g., in two-factor CFA, full WLS might perform worse at smaller sample sizes than F&C suggested, and factor loadings might be underestimated rather than overestimated).

Finally, it is worth noting that the current replication does not address the design of the simulation study itself: That is, how the original authors operationalized the research question. For example, a different team

of researchers might choose different models to test the performance of the two estimation methods, set different levels of factor loadings (or examine how the strength of factor loadings might affect the performance of the methods), or examine different mechanisms of generating nonnormal data (for recent critiques of the Vale-Maurelli method of generating nonnormal data, see Olvera Astivia & Zumbo, 2015; Foldnes & Grønneberg, 2019; see Foldnes & Grønneberg, 2021, for a simulation study on the performance of estimators using Piecewise Linear Transforms, a new nonnormal data generation method).

References

- Chalmers, R. Philip, Adkins, Mark C. (2020). Writing Effective and Reliable Monte Carlo Simulations with the SimDesign Package, *The Quantitative Methods for Psychology*, *16*(4), 248-280. https://doi.org/10.20982/tqmp.16.4.p248
- Chalmers, R. P. (2020). SimDesign: Structure for Organizing Monte Carlo Simulation Designs. R package version 2.1. Retrieved from https://CRAN.R-project.org/ package=SimDesign
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. https://doi.org/10.1037/1082-989X.9.4.466
- Foldnes, N., & Grønneberg, S. (2019). On Identification and Non-normal Simulation in Ordinal Covariance and Item Response Models. *Psychometrika*, *84*(4), 1000–1017. https://doi.org/10.1007/s11336-019-09688-z
- Foldnes, N., & Grønneberg, S. (2021). The sensitivity of structural equation modeling with ordinal data to underlying non-normality and observed distributional forms. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000385
- Olvera Astivia, O. L., & Zumbo, B. D. (2015). A Cautionary Note on the Use of the Vale and Maurelli Method to Generate Multivariate, Nonnormal Data for Simulation Purposes. *Educational and Psychological Measurement*, *75*(4), 541–567. https://doi.org/10.1177/0013164414548894
- Rougier, N. P., Hinsen, K., Alexandre, F., Arildsen, T., Barba, L. A., Benureau, F. C., … & Zito, T. (2017). Sustainable computational science: the ReScience initiative. *PeerJ Computer Science*, *3*, e142.

Acknowledgments

We would like to thank and acknowledge the contribution of Dr. R. Philip Chalmers and Mark Christopher Adkins for providing a large portion of the replication code as well as support in using the SimDesign package (Chalmers, 2020).