A commentary on Lv and Maeda (2019)

Jak, S.; Cheung, M.W.-L.

DOI
10.1080/10705511.2019.1688155
Publication date
2020
Document Version
Final published version
Published in
Structural Equation Modeling

Structural Equation Modeling: A Multidisciplinary Journal, 27: 438–441, 2020
ISSN: 1070-5511 print / 1532-8007 online
DOI: https://doi.org/10.1080/10705511.2019.1688155

A Commentary on Lv and Maeda (2019)

Suzanne Jak[1] and Mike W.-L. Cheung[2]
[1]*University of Amsterdam*
[2]*National University of Singapore*

Meta-analytic structural equation modeling (MASEM) is a statistical technique to fit hypothesized models on the combined data of multiple independent studies. Lv and Maeda (2019) present a simulation study on the performance of three fixed-effects correlation-based MASEM methods with varying levels of data missing completely at random (MCAR). In this commentary, we discuss several coding errors and other issues that we identified, which demonstrate that Lv and Maeda did not evaluate any of the three intended methods. Furthermore, the authors report very surprising results and offer specific recommendations for the application of the three methods; these actions compel us to express our concerns regarding the validity of the conclusions provided by Lv and Maeda.

Keywords: Meta-analytic structural equation modeling, MASEM, meta-analysis

Correlation-based meta-analytic structural equation modeling (MASEM) involves fitting models to a pooled population correlation matrix that is estimated on the basis of correlation coefficients that are reported by multiple independent studies (Cheung & Cheung, 2016). MASEM typically consists of two stages (Viswesvaran & Ones, 1995). In Stage 1, the correlation matrices are combined to form a pooled correlation matrix. In Stage 2, a structural equation model is fit to the pooled matrix from Stage 1. MASEM methods differ in the way that the correlation matrices are pooled (Stage 1) or in the method of fitting the structural equation model in question (Stage 2).

Lv and Maeda (2019) reported on the performance of three correlation-based MASEM methods, featuring varying levels of data missing completely at random (MCAR), with simulated data under a fixed-effects model. The three methods are denoted by W-COV[1] GLS with pairwise deletion (PD), W-COV GLS with multiple imputation (MI), and Two-Stage SEM (TSSEM). Based on their study, the authors provide specific recommendations and conclusions such as "[t]he findings demonstrated the superiority of using W-COV GLS with MI and the necessity of including full matrices for TSSEM and W-COV GLS with PD" (p. 13) and "[t]he inclusion of at least 14 studies with an average within-study sample size equal to or larger than 200 is required for the application of MASEM with TSSEM, W-COV GLS with PD or MI." (p. 12).

We have reasons to believe that there are errors in the simulation study. Based on the available R code listed in the appendices and sent to us by the authors, we suspect that the authors did not evaluate any of the three methods. Moreover, we identified an error in the code that in all likelihood leads to miscalculations of the bias in the parameter estimates. In the next section, we discuss the identified issues one by one.

Correspondence should be addressed to Suzanne Jak, Research Institute of Child Development and Education, University of Amsterdam, Nieuwe Achtergracht 127, Amsterdam, The Netherlands. E-mail: S.Jak@uva.nl

[1] W-COV GLS stands for "weighted covariance GLS." W-COV GLS uses a weighted average of the individual correlation coefficients across studies to estimate the sampling variance and covariances in the individual studies. Next, the correlation matrices are pooled, taking into account the estimated sampling variance and covariances in the individual studies.

THE STUDY DID NOT EVALUATE MULTIPLE IMPUTATION, ONLY SINGLE IMPUTATION

To help researchers applying multiple imputation, Lv and Maeda included R code in Appendix A to illustrate W-COV GLS with MI. Surprisingly, the code shows that the authors actually used single imputation instead of multiple imputation. Although the authors reported that they used 40 imputations to complete the correlation matrices in the primary studies, the R code in Appendix A shows that they used the complete() function of the mice package (van Buuren & Groothuis-Oudshoorn, 2011) without any further arguments. By default, the complete() function returns the first imputed dataset. As a result, the other 39 datasets were not analyzed. The same error was also found in the R code that the authors sent to us. Therefore, the results obtained with these specifications prohibit any valid conclusion about multiple imputation.

THE STUDY DID NOT EVALUATE GLS AT STAGE 2

Appendix A shows that Lv and Maeda indeed used W-COV GLS estimation to combine correlation matrices in Stage 1. In Stage 2, the authors used the lavaan package (Rosseel, 2012) to fit the CFA on the pooled correlation matrix as if the pooled matrix were an observed covariance matrix, using the sum of the sample sizes of the individual studies as the total sample size. In the article, this approach is claimed to be the GLS approach of Becker (1992, 1995), which is not correct. The original GLS approach uses partitioning of the pooled correlations in combination with the asymptotic sampling covariance matrix of the pooled correlation matrix to fit path models at Stage 2. Factor models cannot be estimated using the original GLS method. Researchers can obtain nearly identical results as with the original GLS approach for all types of models using the wls() function implemented in the metaSEM package at Stage 2 (Cheung, 2015b). However, this is also not what the authors used.

The method as implemented by Lv and Maeda combines W-COV GLS at Stage 1 with the procedures used in the so-called "naïve univariate" method at Stage 2 (see Jak & Cheung, 2018). As such, the Stage 2 procedure used does not take into account the uncertainty in the estimates of the pooled correlations from Stage 1 and treats the correlation matrix as if it were a covariance matrix.[2] Results obtained with these settings are therefore not suitable to draw conclusions about the performance of W-COV GLS as proposed by Becker (1992, 1995) and evaluated by Zhang (2011).

[2] Note that Furlow and Beretvas (2005) coined the term W-COV GLS and also fitted the Stage 2 model without taking into account the asymptotic covariance matrix of the pooled correlations, but they did take into account that the input matrix was a correlation matrix and not a covariance matrix.

THE STUDY DID NOT EVALUATE FIXED-EFFECTS TWO-STAGE SEM AT STAGE 1

Because the specificities of the TSSEM method used were not provided in the article (i.e., it is unclear whether TSSEM was applied with a weighted or unweighted asymptotic covariance matrix for the primary studies), we emailed the authors to request the complete syntax of the simulation study. Although we did not receive the full code, the authors were kind enough to share some parts of the code. It showed that TSSEM was used to evaluate the power of the homogeneity tests, but not for further analyses. The employed tssem1() function in combination with the arguments method = "REM" and RE.type = "zero" implemented in the metaSEM package fits a random-effects model with the between-studies variance fixed at zero (Cheung, 2014). This approach is very similar to fixed-effects W-COV GLS but very different from the multiple-group SEM approach of Cheung and Chan (2005), which was explained in the manuscript. Readers may refer to Cheung (2015a, Section 7.5.2) for a detailed explanation of the differences between these two approaches. Results obtained with these settings are therefore not suitable to draw conclusions about the performance of the fixed-effects TSSEM according to Cheung and Chan (2005).

IDENTIFICATION CONSTRAINTS LED TO A MISCALCULATION OF BIAS IN PARAMETER ESTIMATES

Lv and Maeda reported about biased parameter estimates for multiple conditions in their simulation study, including "extremely biased estimates." This finding is surprising; as with MCAR data, the probability of missing data on a variable Y is unrelated to the value of Y itself or to the values of any other variables in the data set (Allison, 2001). Therefore, MCAR data is not expected to lead to biased parameter estimates (Enders & Bandalos, 2001). Indeed, earlier simulation studies evaluating MASEM parameters under MCAR found negligible bias (Cheung, 2000; Furlow & Beretvas, 2005). Note, in a very similar study on the effect of MCAR correlations in fixed effects MASEM, negligible bias was found in parameter estimates for univariate MASEM, W-COV GLS, and TSSEM, even when 70% of the correlation coefficients were missing (Jak & Cheung, 2018).

It appears that the applied identification constraints led to a miscalculation of bias in parameter estimates. Appendix A and the code that the authors sent to us show that the factor loadings for items X1, X5, and X9 were fixed at 0.7 for identification purposes in all conditions. In the conditions with unequal factor loadings, the population values for these factor loadings were actually 0.6. Fixing factor loadings to values different from the population values leads to rescaling of the model parameters, resulting in seemingly nonzero bias

when calculating the parameter bias by plugging in Lv and Maeda's (2019, p. 7, eq. 9) population values. Fixing the factor loadings to 0.7, while the population values are 0.6, leads to correctly estimated factor loadings of 0.817, 0.933, and 1.050, respectively, instead of 0.7, 0.8, and 0.9 for the remaining three factor loadings per factor. In the conditions with equal factor loadings, the population values for the first factor loadings per factor were indeed 0.7. This may explain part of the differences in the amount of parameter bias found between conditions with equal factor loadings versus conditions with unequal factor loadings.

It is important to note that even given the mistakenly implemented methods that were actually evaluated in this study, we would not expect to find systematic bias in parameter estimates with data missing under MCAR. Because Lv and Maeda reported parameter bias for conditions with equal factor loadings as well, it is highly likely that the study contains more errors than we could identify based on the available information. For example, the code that the authors sent to us did not contain the syntax to calculate the relative bias in parameter estimates, so we were not able to evaluate the calculations.

THE STUDY DOES NOT REPORT ANY RESULTS

Even though many of the obtained results may be of limited value given the issues discussed above, it is remarkable that the article does not report the actual results. The actual amounts of bias and error rates per condition are not included in the article itself, nor in an appendix or supplementary file. Instead, the article contains tables with descriptive overall statements per method and condition, such as "slightly decreased by *pc*" or "dropped dramatically when $pc > 0$" relating to Type 1 errors, or "unbiased, extremely [sic] outliers at $pm = .67$" relating to parameter bias. Without the actual results, readers can only guess the values that qualify as extreme, dramatic, or outlying and must speculate the direction of the bias.

REANALYSIS LEADS TO DIFFERENT RESULTS

In order to evaluate whether the simulation study would indeed have shown different results if it was executed correctly, we generated 2000 datasets according to the conditions that Lv and Maeda reported in Appendix A. That is, we generated data for $k = 10$ studies, with an average sample size of $n = 200$, including two studies with complete data ($pc = .20$) and eight studies that featured two missing variables ($pm = .17$), for a factor model with equal factor loadings. According to Table 5 in the article, this specification should lead to biased parameter estimates and biased standard errors for all methods. We did not evaluate the Type I error rate because it was not clear which null hypothesis Lv and Maeda evaluated (which values for which parameters[3]).

The W-COV GLS with multiple imputation using $m = 40$ imputations took 22 min per replication, meaning that analyzing 2000 replications would take 30 days. The long computational time is probably caused by the imputation being extremely difficult with 66 variables (correlation coefficients) to impute based on only 10 subjects (studies). Therefore, we evaluated only 200 replications for W-COV GLS with MI. The authors discussed combining the *m* imputed correlations for each missing correlation using Rubin's rules (Rubin, 1987, p. 76). The code in Appendix A also suggests that the authors intended to combine the imputed correlation matrices before running the MASEM analyses. In our simulation, we fitted the MASEM to each of the imputed datasets instead, and applied Rubin's rules to combine the Stage 2 estimates (factor loadings and factor covariances). Combining the model estimates, and not the multiply-imputed datasets, is preferred because it takes the between-imputation variance of the Stage 2 parameter estimates into account (van Buuren & Groothuis-Oudshoorn, 2011).

The simulation code and exact results are available at https://osf.io/wfdhn. We found less than 5% bias in all parameter estimates for all methods. GLS with PD and TSSEM both produced adequate standard errors, with standard error bias within 5% for all parameters. GLS with MI resulted in large standard errors for all parameters, with positive bias ranging from 28% to 65%[4]. It is not possible to contrast these numbers with the results of the original study because the exact results are not reported. However, it is clear that we did not find biased parameter estimates and biased standard errors for all three methods.

CLOSING REMARKS

In addition to the issues described above and the lack of results in the manuscript, we found several smaller errors. Appendix A shows output from fitting a factor model on nine indicators whereas the dataset and the specified model contain 12 indicators. Table 7 contains the results of the parameter β_{21}, but the population model contains no parameter β_{21}. Figure 2 includes the acronym *pf*, which is not explained in the article.

[3] The description provided in article is "We counted the frequency of the results in which the null hypothesis of parameter estimates was incorrectly rejected over the 2000 replications." From the R code, it seems that the authors tested whether the parameter estimates differed significantly from the population values, but we are not sure if different tests were performed in parts of the unavailable R code.

[4] With multiple imputation, the standard errors are actually expected to be larger than the standard deviations of the associated sample estimates, because the sample estimates will approximately follow a t-distribution with degrees of freedom dependent on the amount of missing data (van Buuren & Groothuis-Oudshoorn, 2011).

More importantly, the complete simulation code is not available online and was not available upon request. We suspect that more problems could be detected if the complete code for the simulation study were to be inspected. Therefore, we suggest that the complete simulation code and the actual results should be publicly available in an accessible format for all future simulation studies, so that readers can verify the exact specifications of data generation and the fitted models.

In our opinion, there is clear evidence that the findings in Lv and Maeda (2019) are untrustworthy, not as a result of willful misconduct but as the result of honest error. Their simulation study did not evaluate any of the claimed methods (W-COV GLS with PD, W-COV GLS with MI, nor fixed-effects TSSEM), yet contains very specific recommendations for the application of these methods. Moreover, given the surprising results regarding parameter bias, we suspect that more problems with the simulation study remain undetected. We think this is very worrying and harmful to the literature. We wrote this commentary in the hope that the unwarranted conclusions and recommendations in Lv and Maeda can somehow be rectified.

ACKNOWLEGEMENTS

The authors thank Jing Lv for answering some questions related to their paper and Hannelies de Jonge for providing feedback on earlier versions of this commentary.

FUNDING

Suzanne Jak was supported by the Netherlands Organisation for Scientific Research under Grant [NWO-VENI-451-16-001].

REFERENCES

Allison, P. D. (2001). *Missing data* (Vol. 136). Thousand Oaks, CA: Sage publications.

Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics*, *17*, 341–362. doi:10.2307/1165128

Becker, B. J. (1995). Corrections to "using results from replicated studies to estimate linear models." *Journal of Educational and Behavioral Statistics*, *20*, 100–102. doi:10.2307/1165390

Cheung, M. W.-L. (2014). Fixed-and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavior Research Methods*, *46*, 29–40. doi:10.3758/s13428-013-0361-y

Cheung, M. W.-L. (2015a). *Meta-analysis: A structural equation modeling approach*. Chichester, UK: John Wiley & Sons, Inc.

Cheung, M. W.-L. (2015b). metaSEM: An R package for meta-analysis using structural equation modeling. *Frontiers in Psychology*, *5*, 1521. doi:10.3389/fpsyg.2014.01521

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods*, *10*, 40–64. doi:10.1037/1082-989X.10.1.40

Cheung, M. W.-L., & Cheung, S. F. (2016). Random-effects models for meta-analytic structural equation modeling: Review, issues, and illustrations. *Research Synthesis Methods*, *7*, 140–155. doi:10.1002/jrsm.1166

Cheung, S. F. (2000). *Examining solutions to two practical issues in meta-analysis: Dependent correlations and missing data in correlation matrices* (Unpublished doctoral dissertation). Hong Kong, China: The Chinese University of Hong Kong.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, *8*, 430–457. doi:10.1207/S15328007SEM0803_5

Furlow, C. F., & Beretvas, S. N. (2005). Meta-analytic methods of pooling correlation matrices for structural equation modeling under different patterns of missing data. *Psychological Methods*, *10*, 227–254. doi:10.1037/1082-989X.10.2.227

Jak, S., & Cheung, M. W.-L. (2018). Accounting for missing correlation coefficients in fixed-effects MASEM. *Multivariate Behavioral Research*, *53*, 1–14. doi:10.1080/00273171.2017.1375886

Lv, J., & Maeda, Y. (2019). Evaluation of the efficacy of meta-analytic structural equation modeling with missing correlations. *Structural Equation Modeling*, 1–24. Advance online publication. doi:10.1080/10705511.2019.1646651.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley and Sons.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67.

Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining psychometric meta-analysis and structural equations modeling. *Personnel Psychology*, *48*, 865–885. doi:10.1111/j.1744-6570.1995.

Zhang, Y. (2011). *Meta-analytic structural equation modeling: Comparison of the multivariate methods* (Doctoral dissertation). Retrieved from http://purl.flvc.org/fsu/fd/FSU_migr_etd-053.