Meta-analytic structural equation modelling with missing correlations

Jak, S.; Oort, F.J.; Roorda, D.L.; Koomen, H.M.Y.

Publication date
2013

Document Version
Final published version

Published in
Netherlands Journal of Psychology

Meta-analytic structural equation modelling with missing correlations

Cheung and Chan (2005) proposed a two-stage method to conduct meta-analytic structural equation modelling (MASEM). MASEM refers to the technique of fitting structural equation models to pooled correlation or covariance matrices from several studies. Unfortunately, researchers do not always report all correlations between the variables of interest. In this paper, we propose a method to deal with missing correlations in the two-stage approach. We illustrate the proposed model with a meta-analysis of teacher-child relationships variables from 99 studies. In addition, using simulated data, we show that our method leads to more precise parameter estimates than the existing approach.

Where: Netherlands Journal of Psychology, Volume 67, 132-139

Received: 14 August 2012; Accepted: 20 November 2012

Keywords: Meta-analytic structural equation modelling, two-stage approach, missing correlations

Authors: Suzanne Jak*, Frans J. Oort*, Debora L. Roorda** and Helma M. Y. Koomen*

*Research Institute of Child Development and Education, University of Amsterdam, the Netherlands, **School Psychology and Child and Adolescent Development, University of Leuven, Belgium

Correspondence to: Suzanne Jak, Methods and Statistics Department of the Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, the Netherlands, e-mail: S.Jak@uva.nl

Meta-analytic structural equation modelling (MASEM) refers to the technique of fitting structural equation models to correlation or covariance matrices from several studies. A well-known approach to conduct MASEM is the two-stage approach of Cheung and Chan (2005). In the first stage, correlation matrices are tested for homogeneity across studies. If the matrices are not significantly different from each other, they are combined to form a pooled correlation matrix. In the second stage, the pooled correlation matrix is taken as the observed matrix in an SEM analysis. User-friendly software to apply the two-stage method is available in the R-Package metaSEM (Cheung, 2011), which utilises the OpenMx package (Boker et al., 2011). MetaSEM gives parameter estimates with standard errors, a chi-square measure of fit, and likelihood based confidence intervals (see Neale & Miller, 1997) for parameters at both stages of the analysis.

Ideally, researchers always report the correlations between *all* variables in their study. However, often not all correlations between the research variables are given in a paper. Sometimes, the missing correlations can be derived from other statistics that the authors do provide, such as regression coefficients. However, this is not always possible, for example when two variables are both outcome variables in regression analyses. The two-stage approach incorporates studies with missing variables, but a way to handle missing correlations has not yet been proposed. As a consequence, for each missing correlation, one of the two variables associated with the correlation has to be treated as missing. We will refer to this method as the *omitted variables approach* (OV approach).

In the present paper, we propose a method to deal with missing correlations in the two-stage approach. This method involves adding one parameter to the model for each missing correlation. We will refer to this method as the *omitted correlations approach* (OC approach). After outlining the method, we illustrate its use with a meta-analysis of teacher-child relationships variables.

MASEM with missing variables and missing correlations

Meta-analysis combines the results from several studies. For MASEM, correlation matrices of the variables of interest will be collected from several studies. The analysis has two stages.

Stage 1: Pooling correlation matrices from several studies

Let R_g be the p_g x p_g sample correlation matrix and p_g be the number of observed variables in the gth study. Some observed correlation matrices may have missing correlations. Moreover, not all studies may include all variables. The correlation matrices for the first three studies may be:

$$R_1 = \begin{bmatrix} 1 & & \\ r_{21_1} & 1 & \\ r_{31_1} & r_{32_1} & 1 \end{bmatrix} \quad R_2 = \begin{bmatrix} 1 & \\ r_{32_2} & 1 \end{bmatrix} \quad R_3 = \begin{bmatrix} 1 & & \\ r_{21_3} & 1 & \\ r_{31_3} & - & 1 \end{bmatrix}.$$

Here, Study 1 has all variables and correlations, Study 2 misses a variable, and Study 3 has all variables but misses a correlation. The OV approach accounts for missing variables, but not for missing correlations. In the OC approach, we account for missing correlations by adding a new matrix (matrix C_g in Equation 1) to the model. We substitute an arbitrary value (e.g. zero) in the observed matrix R_g for a missing correlation. We obtain an estimate of the population correlation matrix R_{pop} of all p variables by fitting a multigroup SEM model, in which the model for each group (study) is:

$$\Sigma_g = D_g\,(M_g\,R_{pop}\,M_g^t)\,D_g^t + C_g\,. \qquad (1)$$

In this model, R_{pop} is the p x p population correlation matrix with diag(R_{pop}) = I, matrix Mg is a p_g x p selection matrix that filters out the missing variables in study g. Matrix M_g is constructed by taking a p x p identity matrix and removing the rows corresponding to the missing variables in study g. D_g is a p_g x p_g diagonal matrix that accounts for differences in variances across the g studies. New in the OC approach is the addition of Matrix C_g, which is used to account for missing correlations. Matrix C_g is a symmetric p_g x p_g correction matrix, with fixed zeros for all present correlations and a free parameter for the missing correlations in study g.

With the identification constraint diag(R_{pop}) = I and a free D_g matrix, the hypothesis that is being tested is equality of covariances (not of correlations) as the variances in each study do not necessarily equal unity. The homogeneity of covariance matrices (covariances and variances) can be tested by constraining the elements of the diagonal matrix D_g to be equal across studies, so that $D_g = D$ for all g (Cheung & Chan, 2005). The unity of variances can be tested by constraining the elements of the diagonal matrix D_g to unity, $D_g = I$ for all g.

The model in Equation 1 is identical to the model in Stage 1 of Cheung and Chan's two-stage approach, except for the correction matrices C_g. In matrix C_g, the free parameter for each missing correlation will take on a value that minimises the difference between the arbitrary chosen value for the missing correlation in the observed matrix R_g (e.g., zero), and the estimate in R_{pop} for the corresponding correlation. A chi-square measure of fit for the model in Equation 1 is obtained by comparing its -2 log likelihood with the -2 log likelihood of the saturated model. The saturated model is given by:

$$\Sigma_g = D_g\,R_g\,D_g^t\,. \qquad (2)$$

The difference between the -2 log likelihoods follows a chi-square distribution with degrees of freedom equal to the difference in numbers of parameters between the two models. When the chi-square test turns out significant, then the hypothesis of homogeneity of covariances is rejected. The chi-square statistic can also be used to calculate approximate fit indices such as the Root Mean Square Error of Approximation (RMSEA, Steiger & Lind, 1980).

When the model fit is not acceptable, then the hypothesis of homogeneity of covariances is not tenable, so that the estimation of R_{pop} is not valid. Researchers may then create clusters of more similar studies, and construct separate pooled correlation matrices for all clusters of studies. Alternatively, a random effects model could be used, which estimates variances (and covariances) between the pooled correlation coefficients across studies. In this paper we do not consider random effects models.

The model from Equation 1 can be fitted using maximum likelihood estimation with any structural equation modelling program. However, writing the syntax can be very laborious in some programs. OpenMx (Boker et al., 2011) is a very flexible R-package, allowing the use of all R functions (R Development Core Team, 2011).

Stage 2: Fitting structural equation models

At Stage 2, the pooled correlation matrix from Stage 1 is used as the input matrix in an SEM analysis. Cheung and Chan (2005) propose using weighted least squares estimation at this stage. Weighted least squares estimation takes the asymptotic covariance matrix of the correlation coefficients from Stage 1 as the weight matrix in the fit function. Some correlation coefficients at Stage 1 are estimated using information from more studies than other correlation coefficients. As a result, coefficients that are based on more studies will have smaller variance in the asymptotic covariance matrix, and thus get more weight in the estimation process than coefficients that are based on less studies.

SEM generally requires the use of covariance matrices, however, the input matrix at Stage 2 is a correlation matrix. Treating the correlation matrix as a covariance matrix leads to incorrect results when estimating confidence intervals or when testing specific hypotheses (Cudeck, 1989). To obtain correct results at Stage 2, we add a so-called estimation constraint. This constraint enforces the diagonal of the model implied correlation matrix to identity.

Illustrative example

Data

Roorda, Koomen, Spilt and Oort (2011) collected 99 studies that reported correlations between positive teacher-student relations and negative teacher-student relations on the one hand and student engagement and student achievement on the other hand. Correlations between positive teacher-student relations and negative teacher-student relations were collected afterwards for the present paper. Of these studies, 63 were conducted at primary schools and 36 at secondary schools. In total, there where 129,184 respondents (sample sizes ranging from 42 to 39,553). Based on leading theories about teacher-student relations (Connell & Wellborn, 1991; Pianta, 1999), teacher-student relations were considered as exogenous variables and engagement and achievement as endogenous variables. Out of the 99 studies, 20 studies missed a correlation between two variables, and 90 studies did not include one or more of the four variables.

Results

Table 1a gives the fit results of the several models we fitted to the 99 correlation matrices, using OpenMx (Boker et al., 2011). Model 1 is a saturated model, meaning that a correlation matrix is estimated for each study, without equality restrictions across studies. This model is used as a baseline model, to obtain fit indices for Models 2 to 4. Model 2 is a model in which we restricted all covariances to be equal, without restrictions on the variances across studies (equal covariances). Model 3 is a model in which we restricted all covariances and variances to be equal across studies (equal variances and covariances). Model 4 is a model in which we additionally restricted all variances to be unity (equal variances across studies).

All three models at Stage 1 had significant chi-square values, indicating that the models do not fit

Table 1 Fit results of models at Stage 1 and Stage 2 using two approaches (N = 129,184)

a) Fit results using the OC approach

Model	-2 log likelihood	df	χ^2	RMSEA + 95% CI
STAGE 1				
1. Saturated	271981.9	0	0	-
2. Equal covariances (D_g = free)	277400.2	214	5418.34	.0137 [.0133 ; .0141]
3. Equal variances and covariances ($D_g = D$)	277957.2	474	5975.35	.0095 [.0092 ; .0097]
4. Variances equal to unity ($D_g = I$)	277957.4	478	5975.53	.0094 [.0092 ; .0097]
STAGE 2				
5. Mediation model (based on Model 2)	144.73	2	144.73	.0235 [.0197 ; .0274]

b) Fit results using the OV approach

Model	-2 log likelihood	df	χ^2	RMSEA + 95% CI
STAGE 1				
1. Saturated	249821.0	0	0	-
2. Equal covariances (D_g = free)	254872.7	193	5051.64	.0140 [.0136 ; .0144]
3. Equal variances and covariances ($D_g = D$)	255420.2	434	5599.22	.0096 [.0093 ; .0099]
4. Variances equal to unity ($D_g = I$)	255420.3	438	5599.28	.0096 [.0093 ; .0098]
STAGE 2				
5. Mediation model (based on Model 2)	178.48	2	78.48	.0261 [.0223 ; .0301]

Table 2 Pooled correlation matrix across 99 studies, N = 129,184

	1.	2.	3.	4.
1. Positive teacher-student relations	1			
2. Negative teacher-student relations	-.373	1		
3. Student engagement	.385	-.345	1	
4. Student achievement	.157	-.152	.284	1

the data exactly. The RMSEAs were all below .05, indicating close approximate fit (Browne & Cudeck, 1992). As we do not have any hypothesis on the variances being equal across studies, we take the result of Model 2 as the estimate of the population correlation matrix. This correlation matrix is given in Table 2, and is used as the input for the Stage 2 analysis.

The Stage 2 model is based on social-motivational theory (Connell & Wellborn, 1991), in which it has been hypothesised that student engagement acts as a mediator in the association between teacher-student relations and student achievement. Empirical studies have provided some support for the mediating role of engagement (e.g., Hughes, Luo, Kwok, & Loyd, 2008). Therefore, the path model we fitted was a mediation model, in which the influence of positive and negative teacher-child relationships on student achievement was mediated by student engagement. This model fitted the population correlation matrix from Stage 1 closely according to the RMSEA. Figure 1 provides a graphical representation of the model, with standardised parameter estimates and 95% confidence intervals. Positive teacher-student relations had a positive effect on Student engagement (β = .296, p < .05). Negative teacher-student relations had a negative influence of about the same size (β = -.255, p < .05). Student engagement had a medium sized positive effect on Student achievement (β = .322, p < .05). The indirect effect of Positive teacher-student relations on Student achievement via Student engagement was small and positive (β = .095, p < .05). Negative teacher-student relations had a similar small-sized negative indirect effect on Student achievement (β = -.082, p < .05). The model explained 20.8 % of the variance in Student engagement, and 10.0% of the variance in Student achievement.

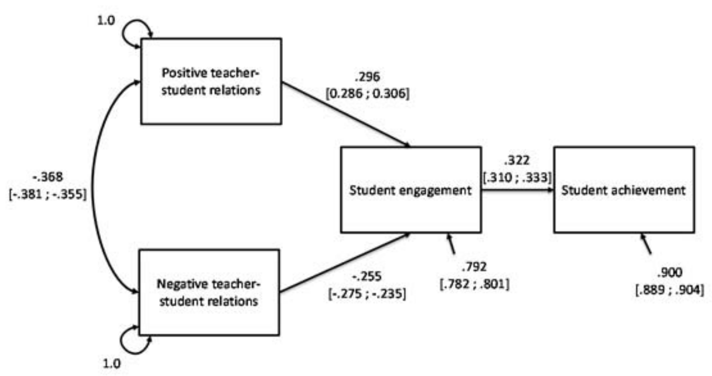

Figure 1 Path model with parameter estimates and their 95% confidence intervals

Results when deleting variables with missing correlations

We compared our results with the results obtained with the OV approach. This involved the deletion of a variable in 20 of the 99 studies. As can be seen in Table 1b, this approach leads to a loss of degrees of freedom and slightly worse model fit. However, the models still fit closely according to the RMSEA. Some parameter estimates are different from the previous analysis, and the likelihood based confidence intervals are somewhat wider. Figure 2 shows a graphical comparison of the parameter estimates and confidence intervals from the two analyses. Each graph pictures the parameter estimate (the dot) with its 95% confidence interval (the line) for the analysis with the OC approach (upper part) and for the analysis with the OV approach (lower part).

Simulation study

In order to investigate the effect of accounting for the missing correlations (OC approach), compared with deleting variables associated with missing correlations (OV approach), we performed an analysis of simulated data. We generated complete data for 100 studies. The pooled correlation matrix was estimated based on the full data, and based on data with missing correlations using the two approaches. The data were simulated under extreme conditions, so that differences between the two analysis methods became more apparent than in the illustration. Our expectation is that the OC approach leads to models with more power, better parameter estimates and smaller confidence intervals than the OV approach.

Data generation

We chose values of the population correlations based on the data from Roorda et al. (see Table 2). For each of the 500 replications, complete raw data were

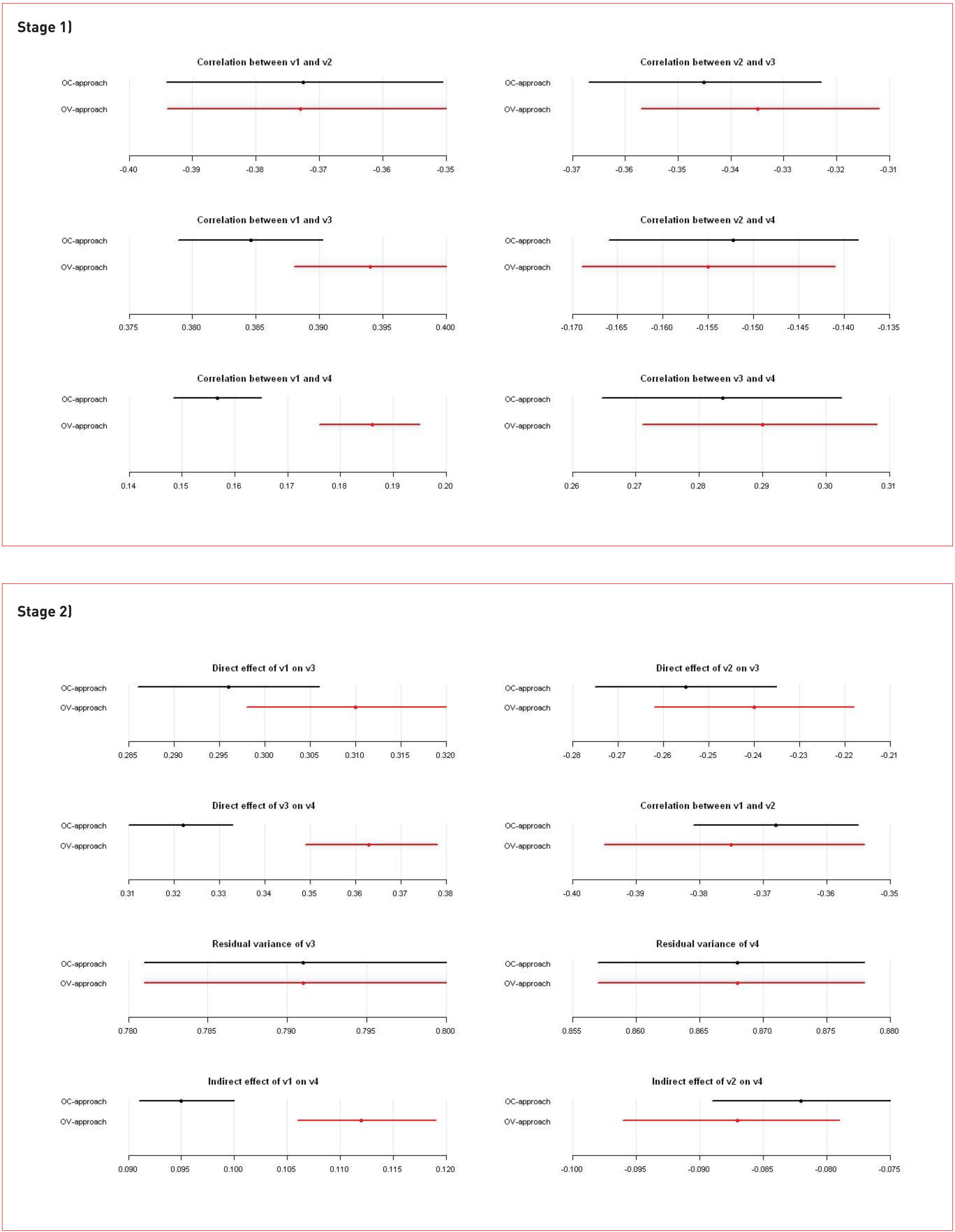

Figure 2 Parameter estimates (dots) with 95% confidence intervals (lines) of all parameters for the analysis with the OC approach (upper dot + line) and the OV approach (lower dot + line)

Note: v1 = Positive teacher-student relations, v2 = Negative teacher-student relations, v3 = Student engagement, v4 = Student achievement

Table 3 Estimated average pooled correlations and bias percentages with a) complete data, b) the OC approach and c) the OV approach

a) Complete data		*Estimated correlations*				*Estimation bias (%)*	
	Variable 1	Variable 2	Variable 3	Variable 4	Variable 1	Variable 2	Variable 3
Variable 1	1						
Variable 2	-.371	1			- 0.54%		
Variable 3	.356	-.343	1		- 0.56%	- 0.58%	
Variable 4	.156	-.151	.283	1	- 0.64%	- 0.66%	- 0.35%
b) OC approach		*Estimated correlations*				*Estimation bias (%)*	
	Variable 1	Variable 2	Variable 3	Variable 4	Variable 1	Variable 2	Variable 3
Variable 1	1						
Variable 2	-.373	1			0.00%		
Variable 3	.359	-.349	1		0.28%	1.16%	
Variable 4	.159	-.153	.287	1	1.27%	0.66%	1.06%
c) OV approach		*Estimated correlations*				*Estimation bias (%)*	
	Variable 1	Variable 2	Variable 3	Variable 4	Variable 1	Variable 2	Variable 3
Variable 1	1						
Variable 2	-.371	1			-0.54 %		
Variable 3	.355	-.345	1		-0.84 %	0.00 %	
Variable 4	.156	-.151	.283	1	-0.64 %	-0.66 %	-0.35 %

drawn from the multivariate normal distribution, with means equal to zero, and variance covariance matrix equal to the population correlations. We chose 100 as the number of studies, with each study having a sample size of 100. So, each dataset contained 100 x 100 = 10000 scores on 4 variables. For each study, the correlation matrix was calculated and included in the meta-analysis.

Results with complete data

Fitting the model from Equation 1 (with $D_g = I$) to the complete data in 500 samples led to the average correlations in Table 3a. Percentages of estimation bias in all parameters are calculated as 100 × (mean estimated value – population value) / population value. According to Muthén, Kaplan and Hollis (1987), estimation bias less than 10% can be considered negligible. With complete data, estimation bias was below 1% for all parameters. The model had 994 degrees of freedom, the average of all chi-square values was 294.48 (*SD* = 41.28).

Results with missing correlations

The correlation between variables 1 and 3 was deleted randomly for 80% of the studies. Also for 80% of the studies we randomly deleted the correlation between variables 2 and 3. In this way, about 64% of the studies missed both correlations, and about 32% studies missed one of them, while only about 4% of the studies had complete data.

Using the OC approach, we obtained the pooled correlation matrix shown in Table 3b. The model had 834 degrees of freedom, the average chi-square value was 119.29 (*SD* = 39.60). The estimated correlations based on the data with missing correlations were close to the population correlations. The largest difference was found for the correlation between variables 1 and 4, which deviated 1.27% from the population correlation. The results in Table 3c were obtained by removing one variable for each missing correlation (the OV approach). This model had 618 degrees of freedom, the average chi-square value was 62.85 (*SD* = 29.98). The results do not differ very much from the results in Table 3b. The largest difference was found for the correlation between variables 1 and 3, which deviated -0.84 % from the population value.

The bias in parameter estimates is not very different across the three models (complete data vs. OC approach vs. OV approach). However, a structural

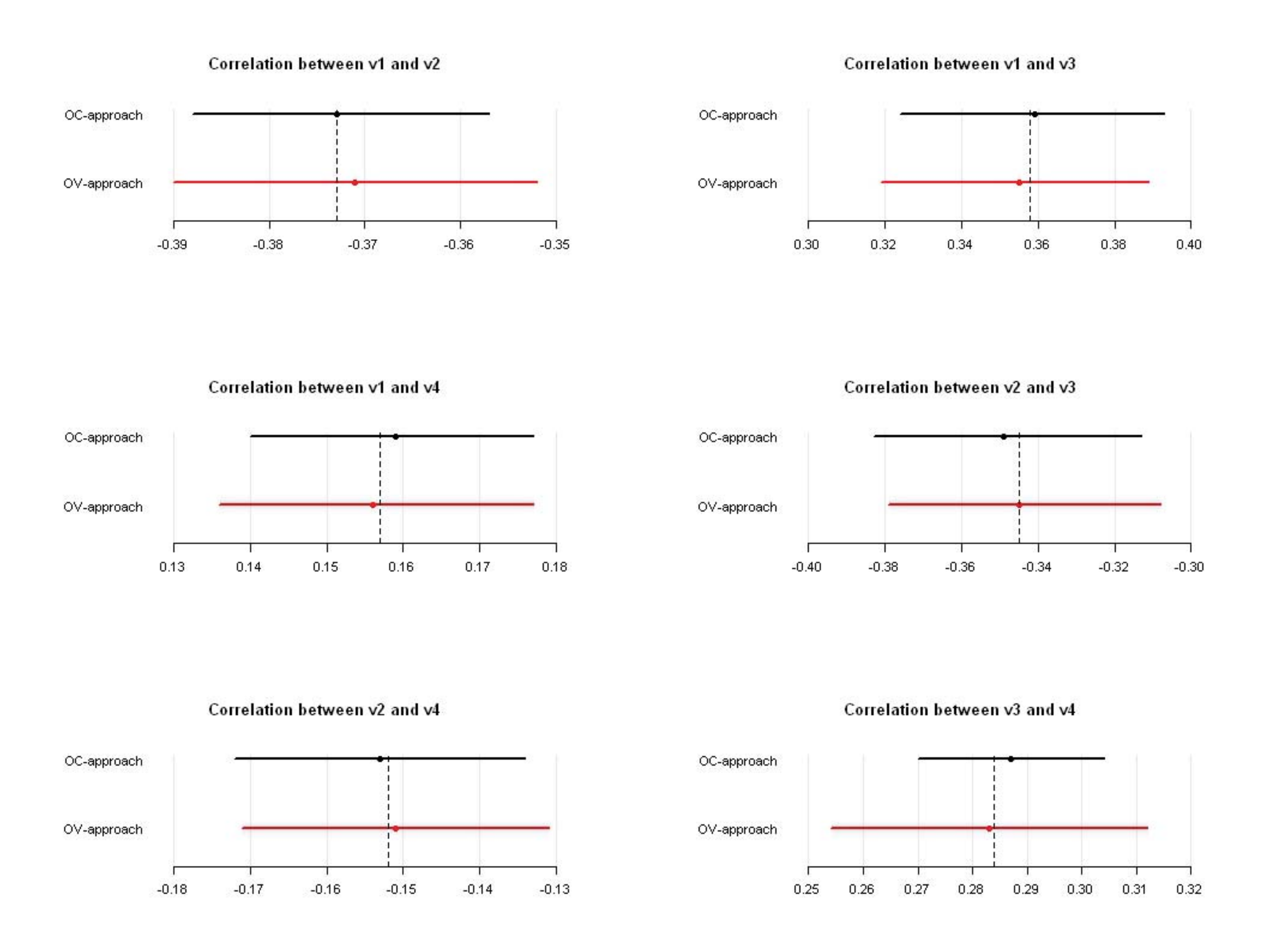

Figure 3. Average parameter estimates (dots) with 95% confidence intervals (lines) of all correlations with the OC approach (upper dot + line) and the OV approach (lower dot + line). The dotted vertical line denotes the population value of the parameter

difference can be seen in width of the likelihood based confidence intervals. These are smaller with the use of the OC approach, compared with the OV approach. Figure 3 gives a graphical representation of the parameter estimates and the associated 95% likelihood based confidence intervals. The upper dot and line denote the average parameter estimate and confidence interval of the analysis with the OC approach, while the lower dot and line denote the average parameter estimate and confidence interval of the analysis with the OV approach. The dotted vertical line shows the population value of the parameter. As expected, omitting information leads to larger confidence intervals. The difference is most clearly seen in the correlations between variables 3 and 4 in the lower right corner of the figure. It is not surprising that the difference is so apparent for the correlation between variables 3 and 4. As we deleted the correlation between variables 1 and 3 and between variables 2 and 3, both methods use equal amounts of information about these correlations. However, where our method still uses information about the correlation between variables 3 and 4, this information is often deleted in the other approach, leading to less precise parameter estimates.

Discussion

In this paper we have proposed a method to incorporate missing correlations in MASEM. The method was demonstrated with an example from teacher-child interactions. Using simulated data, the method was compared with the current practice, which is to delete one of the variables that is associated with the missing correlation. As the OC approach uses more information than the OV approach we expected that our method would lead to better parameter estimates and smaller confidence intervals. Results from the very small simulation study indicated that the OC approach leads to models with larger degrees of freedom and smaller confidence intervals. The parameter estimates were close to the population values for all methods, indicating that deleting one or two variables in 80% of the studies did not really influence parameter recovery.

A possible explanation of the similar results with respect to parameter bias between the methods is that in our study the missingness of the correlation coefficients were introduced randomly. Maximum likelihood estimation with data missing at random is known to lead to unbiased parameter estimates (e.g., Enders & Bandalos, 2001; Newman, 2003). In true meta-analysis, the missing correlations may not be missing at random, and it would be interesting to investigate the effect of not-random missingness. For example, if the correlations between variables 3 and 1 and variables 3 and 2 are mainly missing in studies where the correlation between variable 4 and 3 is high, then if we delete variable 3 in the OV approach, the information about the correlation between variable 4 and 3 is lost, and the parameter will be underestimated. Using the OC approach, all remaining information would be used and the parameter estimate is expected to be closer to the true value.

References

Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., et al. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76,* 306-317.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research, 21,* 230-258.

Cheung, M. W. L. (2011). 'metaSEM: Meta-analysis: A Structural Equation Modeling Approach', R package version 0.7-0, http://courses.nus.edu.sg/course/psycwlm/internet/metaSEM/ (accessed 12 March 2012).

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: A two-stage approach. *Psychological Methods, 10,* 40-64.

Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. In M. Gunnar & A. Sroufe (Eds.), *The Minnesota symposium on child development: Self-processes and development* (Vol. 23, pp. 43-77). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin, 105*, 317-327.

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.

Hughes, J. N., Luo, W., Kwok, O.-M., & Loyd, L. K. (2008). Teacher-student support, effortful engagement, and achievement: A 3-year longitudinal study. *Journal of Educational Psychology, 100*, 1-14.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431-462.

Neale, M. C., & Miller, M. B. (1997). The use of likelihood based confidence intervals in genetic models. *Behavior Genetics, 27*, 113-120.

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods, 6,* 328-362.

Pianta, R. C. (1999). *Enhancing relationships between children and teachers*. Washington, DC: American Psychological Association.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. [Computer software]. Vienna, Austria.

Roorda, L. R., Koomen, H. M. Y., Spilt, J. L. & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Research, 81*, 493-529.

Steiger, J. H., & Lind, J. C. (1980). Statistically based tests for the number of common factors. *Paper presented at the annual meeting of the Psychometric Society,* Iowa City, IA.

SUZANNE JAK
PhD candidate in Methods and Statistics, Research Institute of Child Development and Education, University of Amsterdam. Her research interests include structural equation modelling, measurement bias, multilevel structural equation modelling and meta-analytical structural equation modelling.

FRANS J. OORT
Full professor of Methods and Statistics and Director of the Research Institute of Child Development and Education, University of Amsterdam. His research interests include measurement and measurement bias, and statistical modelling, especially non-standard applications of structural equation modelling.

DEBORA L. ROORDA
Postdoc research fellow at the School Psychology and Child and Adolescent Development Department at the University of Leuven, Belgium. Her research interests focus on teacher-student interactions and their influence on children's emotional well-being, school engagement, and achievement.

HELMA M. Y. KOOMEN
An assistant professor in special education at the Research Institute of Child Development and Education at the University of Amsterdam. Her research interests address student-teacher interaction and teachers' pedagogical practices in relation to emotional well-being, social adjustment, task behaviours, and achievement of children with and without special need.