

Meta-Analytic Structural Equation Modeling: A Two-Stage Approach

Mike W.-L. Cheung
University of Hong Kong

Wai Chan
Chinese University of Hong Kong

To synthesize studies that use structural equation modeling (SEM), researchers usually use Pearson correlations (univariate r), Fisher z scores (univariate z), or generalized least squares (GLS) to combine the correlation matrices. The pooled correlation matrix is then analyzed by the use of SEM. Questionable inferences may occur for these ad hoc procedures. A 2-stage structural equation modeling (TSSEM) method is proposed to incorporate meta-analytic techniques and SEM into a unified framework. Simulation results reveal that the univariate- r , univariate- z , and TSSEM methods perform well in testing the homogeneity of correlation matrices and estimating the pooled correlation matrix. When fitting SEM, only TSSEM works well. The GLS method performed poorly in small to medium samples.

Structural equation modeling (SEM) is a frequently used multivariate technique for testing hypothetical models in behavioral sciences. A simple review of the literature shows its increase in popularity as a research tool (Hershberger, 2003; MacCallum & Austin, 2000; Tremblay & Gardner, 1996). However, the increase of research findings based on SEM does not necessarily imply that consistent conclusions can be drawn from these studies. Different researchers may propose redundant models that are supported by their own data, and it is difficult to compare these models systematically. Moreover, it has been found that researchers are often reluctant to consider using alternative models (MacCallum & Austin, 2000). This confirmation bias—the prejudice in favor of the model being evaluated—hinders the development of research progress (Greenwald, Pratkanis, Leippe, & Baumgardner, 1986). Hence, conducting more empirical research does not necessarily decrease the uncertainty of a

particular topic if the findings from that research are inconsistent (National Research Council, 1992; Rosenthal & DiMatteo, 2001).

Meta-analysis (MA), a term coined by Glass (1976), is “the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings” (p. 3). It is widely used as a statistical technique to synthesize research findings in social sciences (e.g., Rosenthal & DiMatteo, 2001; Schmidt, 1992). Based on the fruitful results of MA obtained in many research domains, it can be beneficial for researchers to use meta-analytic techniques and SEM in testing causal models (e.g., Becker & Schram, 1994; Miller & Pollock, 1994; Shadish, 1996; Viswesvaran & Ones, 1995).

Meta-analytic structural equation modeling (MASEM), involving the techniques of synthesizing correlation matrices and fitting SEM, is usually done by applying meta-analytic techniques on a series of correlation matrices to create a pooled correlation matrix, which can then be analyzed using SEM (e.g., Viswesvaran & Ones, 1995). However, MA and SEM were developed under different research traditions. The statistical theories of MA and SEM are based on the distributions of correlations and covariance matrices, respectively. Hence, there is no guarantee that the inferences based on combining these two approaches will be correct. Moreover, empirical studies on the validity of these procedures are rare (cf. M. W. L. Cheung, 2002; S. F. Cheung, 2000; Hafdahl, 2001).

Objectives of the Current Study

The main objective of the current study is to propose a two-stage structural equation modeling (TSSEM) method to integrate meta-analytic techniques and SEM into a unified

Mike W.-L. Cheung, Department of Psychology, University of Hong Kong; Wai Chan, Department of Psychology, Chinese University of Hong Kong.

Additional materials are on the Web at <http://dx.doi.org/10.1037/1082-989X.10.1.40.supp>.

Portions of this article were presented at the 25th International Congress of Applied Psychology, Singapore, July 2002. This work was partially supported by the Seed Funding for Basic Research from the University of Hong Kong.

We thank N. Balakrishnan, Ben Chan, Darius Chan, Shu-Fai Cheung, Marissa Dean, Arthur Woodward, and Byerly Woodward for their comments on versions of this article and Kevin Au for sharing with us information about International Social Survey Program data set.

Correspondence concerning this article should be addressed to Mike W.-L. Cheung, Department of Psychology, University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: mikewlcheung@hku.hk

framework.¹ Specifically, we propose to use the technique of multiple-group analysis in SEM to synthesize correlation matrices at the first stage and to fit hypothesized structural models based on the pooled correlation matrix at the second stage. The empirical performance of TSSEM is compared with several widely used meta-analytic procedures, such as the univariate methods that combine Pearson correlations (univariate r ; Hunter & Schmidt, 1990) or Fisher z scores (univariate z ; Hedges & Olkin, 1985), and the multivariate methods, such as generalized least squares (GLS; Becker, 1992, 1995). Although there are several studies combining MA and path analysis or exploratory factor analysis (e.g., Becker, 1992, 1995; S. F. Cheung, 2000; Hafdahl, 2001), to the best of our knowledge this is the first attempt to incorporate meta-analytic techniques under the general framework of SEM. We hope this research will raise the interests of researchers in the methodological development of MASEM.

Common Approaches to Meta-Analytic SEM

Researchers typically conduct MASEM by means of a two-stage procedure (Viswesvaran & Ones, 1995). In the first stage, correlation matrices are tested for homogeneity. If they are not significantly different, they are combined to form a pooled estimate. If they are heterogeneous, potential moderators may be used to explain the variation across studies or a random-effects model may be used to average the correlations. In the second stage, the pooled correlation matrix is treated as the observed correlation matrix and used in SEM.

Univariate Methods for Meta-Analytic SEM

The most frequently used approach refers to a class of univariate methods in which the elements of a correlation matrix are treated as independent within studies and they are pooled separately across studies (e.g., Brown & Peterson, 1993; Schmidt, Hunter, & Outerbridge, 1986).

Stage 1: Testing Homogeneity of Correlation Matrices and Estimating the Pooled Correlation Matrix

In estimating the pooled correlation coefficient in the bivariate case, Hunter and Schmidt (1990) proposed to compute the pooled correlation coefficients by weighting by sample size. Because the sampling distribution of the correlation coefficient becomes skewed when the population correlation deviates from zero, Hedges and Olkin (1985) proposed to use Fisher's z -transformed correlation coefficients. To avoid confusion between the original methods proposed by Hunter and Schmidt (1990) and Hedges and Olkin (1985) in the bivariate case and the current approach,

we use the terms univariate- r and univariate- z methods to denote the applications of Hunter and Schmidt's and Hedges and Olkin's methods to MASEM, respectively. Moreover, the univariate- r method applied in our research is also a simplified version of the method proposed by Hunter and Schmidt because it does not involve issues such as correction for unreliability or range restriction.

Because the homogeneity tests of univariate- r and univariate- z values were developed for bivariate correlations only, modifications should be made when testing the equality of elements across correlation matrices in MASEM.² Researchers have no widely agreed-on approach for testing a composite hypothesis about correlation matrices using univariate- r and univariate- z methods. S. F. Cheung (2000) proposed a Bonferroni-adjusted at-least-one (BA1) approach. The idea is that the composite hypothesis of the homogeneity of correlation matrices across studies implies the homogeneity of all bivariate correlations. The hypothesis of homogeneity of correlation matrices will be rejected if at least one of the correlation coefficients is heterogeneous across studies. To control for the overall Type I error in conducting multiple tests, the decision rule is to reject the hypothesis of homogeneity of all correlation matrices if at least one of the observed probability values (p_{ij}) obtained by testing the individual correlation coefficients ρ_{ij} between the i th and j th variables is smaller than the significance level adjusted for multiple comparisons. In other words,

$$\min_{i \neq j}(p_{ij}) < \frac{\alpha}{p(p-1)/2}. \quad (1)$$

Here, $\min(\cdot)$ is the minimum value of all the elements, p_{ij} is the p value for testing the individual $H_0: \rho_{ij}^{(1)} = \rho_{ij}^{(2)} = \dots = \rho_{ij}^{(K)}$ across K studies, α is the significance level, and p is the number of variables. For instance, if there are four variables, the hypothesis of homogeneity of all the correlation matrices with six correlations will be rejected if any one of the six observed p values from testing the homogeneity of individual correlations is smaller than $.05/6$ or $.00833$. The BA1 approach will be used as the criterion for hypothesis testing for the univariate- r and univariate- z methods in this study because it was found that the rejection rate of the BA1

¹ Although other MASEM procedures discussed here also entail two stages, they use different methods in different stages, namely meta-analytic techniques for synthesizing correlation matrices in Stage 1 and SEM for fitting proposed models in Stage 2. We use the label of TSSEM, highlighting the fact that our procedure uses SEM as the sole statistical framework for both stages.

² Although Hunter and Schmidt (1990) did not recommend homogeneity tests and commented that they have "all the flaws of any significance test" (p. 428), researchers using the univariate- r method sometimes still use these tests in MASEM (e.g., Hom et al., 1992).

approach to testing the homogeneity of all correlation matrices was close to the preset Type I error α (S. F. Cheung, 2000).

One problem in synthesizing correlation matrices is that studies may involve different numbers of variables, because different researchers conduct research independently. There are two common methods to handle this issue (Viswesvaran & Ones, 1995). The first method is to include only studies that contain all the variables in the model, that is, to use listwise deletion (e.g., Hom, Caranikas-Walker, Prussia, & Griffeth, 1992). The second method, which is the dominant method for applied researchers, is to estimate the elements of the pooled correlation matrix based on different numbers of studies, that is, to use pairwise deletion (e.g., Brown & Peterson, 1993; Premack & Hunter, 1988). The main advantage of pairwise deletion is that it includes all the studies that are available.

Stage 2: Fitting SEMs

After obtaining a pooled correlation matrix at the first stage, researchers often use the pooled correlation matrix as the observed covariance matrix in fitting SEM in Stage 2. Software like LISREL (Jöreskog & Sörbom, 1996) and EQS (Bentler, 1995) can be used to analyze path models or models with latent variables (e.g., Brown & Peterson, 1993; Brown & Stayman, 1992; Premack & Hunter, 1988).

When fitting the model in Stage 2, four statistical difficulties can occur. The first is in deciding on an appropriate sample size for the model fitting in SEM. Because the pooled correlation matrix is usually formed by averaging across different studies based on pairwise deletion, researchers have to decide on the appropriate sample size for the analysis in SEM. Researchers have used a variety of sample sizes such as the arithmetic mean (Carson, Carson, & Roe, 1993; Premack & Hunter, 1988; Verhaeghen & Salthouse, 1997), the harmonic mean (Colquitt, LePine, & Noe, 2000; Conway, 1999), the median (Brown & Peterson, 1993), or the total (Hunter, 1983; Tett & Meyer, 1993) of the sample sizes based on the synthesized correlation coefficients.³ The problem is that all of these ideas are simply ad hoc solutions because they are not based on any statistical theory. Because the Type I error of the chi-square test statistics, the goodness-of-fit indices, the statistical power, and the standard errors of parameter estimates are all dependent on the sample size used, using different sample sizes can result in different inferences.

The second difficulty is the possibility of using a non-positive definite matrix as the input matrix in SEM. Because each study may contain a different set of variables, the pooled correlation matrix from pairwise deletion may be nonpositive definite (Marsh, 1998; Wothke, 1993, 2000). In such cases, SEM is no longer appropriate. Moreover, even though the pooled correlation matrix based on pairwise

deletion may be positive definite, its statistical properties are still questionable in SEM because different elements of the pooled correlation matrix are probably based on different samples (Wothke, 2000).

The third difficulty is in ignoring the sampling variation across studies. After pooling the correlation matrices, researchers often use the pooled correlation matrix as the observed correlation matrix without considering the sampling variation across studies (e.g., Colquitt et al., 2000; Tett & Meyer, 1993; Verhaeghen & Salthouse, 1997). There are sampling variations in individual correlation matrices even when they share the same population correlation matrix. This issue is also related to the first difficulty of finding an appropriate sample size for the SEM-fitting model.⁴ Univariate approaches adequately incorporate sampling variation into the pooled correlation estimates by weighting studies based on their precision, but they fail to incorporate these weightings into SEM by relying on an ad hoc choice of sample size. Some estimated mean correlation coefficients may contain more sampling variation, whereas others may have less. However, the sampling variation associated with the pooled correlation matrix is not reflected when fitting SEM under the univariate approaches in which their standard errors are ignored. Moreover, the covariation among the correlations is totally ignored in the univariate approaches despite the fact that the correlations are indeed correlated to a certain extent (Olkin & Siotani, 1976).

The fourth difficulty is analyzing a correlation matrix instead of a covariance matrix. It is generally incorrect to analyze the correlation matrix in SEM, although many published articles using MASEM treat the pooled correlation matrix as a covariance matrix. Many researchers have warned about the problems of analyzing the correlation matrix instead of the covariance matrix in primary research applications of SEM. Specifically, the chi-square statistics and the standard errors of parameter estimates may be incorrect (Cudeck, 1989; Jöreskog & Sörbom, 1996).

The first two difficulties may be encountered only when pairwise deletion is used in handling missing correlations, whereas the third and fourth difficulties may occur regardless of whether pairwise or listwise deletion is used. Although listwise deletion encounters less technical problems than pairwise deletion, listwise deletion is less popular in

³ The sample size used as input is calculated from the sample sizes cumulated for individual correlations. For instance, suppose that \bar{r}_{21} , \bar{r}_{31} , and \bar{r}_{32} are the pooled correlations based on sample sizes of 100, 500, and 1,000, respectively, across studies. Then the arithmetic mean, the harmonic mean, and the median are calculated based on 100, 500, and 1,000. The total sample size is simply the sum of the sample sizes involved in these studies. Note that the total sample size may be larger than 1,000 because some studies that report r_{21} and r_{31} may not include r_{32} .

⁴ We thank an anonymous reviewer for this suggestion.

MASEM because most studies would be deleted as a result of missing correlations. Thus, pairwise deletion is assumed for the univariate methods that are used in this study. Although univariate methods suffer from these technical difficulties and are generally not recommended (Becker, 2000; Becker & Schram, 1994), they are still the ones that are primarily used by applied researchers.

GLS for Meta-Analytic SEM

GLS is one of the multivariate approaches to estimating the pooled correlation matrix and the asymptotic covariance matrix of parameter estimates from independent studies at the first stage (Becker, 1992, 1995, 2000; Becker & Schram, 1994; Hedges & Olkin, 1985). The pooled correlation matrix can be used to fit path models at the second stage (Becker, 1992, 1995, 2000; Becker & Schram, 1994).

Stage 1: Testing Homogeneity of Correlation Matrices and Estimating Pooled Correlation Matrix

Let $P^{(g)}$ and $R^{(g)}$ be the $p^{(g)} \times p^{(g)}$ population and sample correlation matrices and $p^{(g)}$ be the number of variables observed in the g th study. For ease of introduction, it is assumed that all pairwise correlations are reported given the $p^{(g)}$ variables in the g th study, which is a bit more general than the cases presented by Becker (1992) and Hedges and Olkin (1985). Nevertheless, the method is capable of handling missing correlations (see Becker & Schram, 1994).

We then let $\rho^{(g)}$ and $\mathbf{r}^{(g)}$ be the $\frac{p^{(g)}(p^{(g)} - 1)}{2} \times 1$ vectors obtained by stringing out the lower off-diagonal elements in $P^{(g)}$ and $R^{(g)}$, respectively. Olkin and Siotani (1976) showed that $\mathbf{r}^{(g)}$ is approximately distributed with a multivariate normal distribution, $N(\rho^{(g)}, \Omega(\rho^{(g)}))$, in large samples where $\Omega(\rho^{(g)})$ is the $\frac{p^{(g)}(p^{(g)} - 1)}{2} \times \frac{p^{(g)}(p^{(g)} - 1)}{2}$ covariance matrix of $\mathbf{r}^{(g)}$ depending on the parameters of $\rho^{(g)}$. Because the population values of $\rho^{(g)}$ are seldom known, their sample estimates $\mathbf{r}^{(g)}$ are often substituted for the calculation of $\Omega(\rho^{(g)})$, that is, $\Omega(\mathbf{r}^{(g)})$. Notice that when missing correlations are observed, $\rho^{(g)}$ and $\mathbf{r}^{(g)}$ may have different dimensions across studies. Let ρ be a $\frac{p(p-1)}{2} \times 1$ complete correlation vector that contains all possible population correlations among p variables, where $p \geq p^{(g)}$ for all g . Then for the g th study, define a $\frac{p^{(g)}(p^{(g)} - 1)}{2} \times \frac{p(p-1)}{2}$ selection matrix, $G^{(g)}$, with 0s and 1s that select the appropriate correlation coefficients in the g th study. That is, $\rho^{(g)} = G^{(g)}\rho$.

To illustrate with an example, suppose that there are four variables with six correlation coefficients ($\rho_{21}, \rho_{31}, \rho_{32}, \rho_{41}, \rho_{42}, \rho_{43}$) in three studies. The first study is complete, and the

other two studies have incomplete correlation matrices. For instance, let $\rho^{(1)} = \rho = (\rho_{21}, \rho_{31}, \rho_{32}, \rho_{41}, \rho_{42}, \rho_{43})^T$, $\rho^{(2)} = (\rho_{21}, \rho_{31}, \rho_{32})^T$, $\rho^{(3)} = (\rho_{21}, \rho_{41}, \rho_{42})^T$. Then the reduced correlation vectors for the g th study can be selected by the appropriate selection matrices:

$$G^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$G^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix},$$

$$\text{and } G^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}.$$

Now let $G = [G^{(1)T} | G^{(2)T} | \dots | G^{(K)T}]^T$ be a matrix obtained by stacking the selection matrices from the K studies. By assuming a linear combination of the true correlation vector and sampling error, the observed correlation vector can be expressed as

$$\mathbf{r} = G\rho + \mathbf{e} \quad (2)$$

where $\mathbf{r} = [\mathbf{r}^{(1)T} | \mathbf{r}^{(2)T} | \dots | \mathbf{r}^{(K)T}]^T$ and \mathbf{e} is a $\left(\sum_{g=1}^K \frac{p^{(g)}(p^{(g)} - 1)}{2}\right) \times 1$ vector of random errors with $E(\mathbf{e}) = 0$, it can be shown that the value of ρ estimated using GLS is

$$\hat{\rho} = (G^T \Omega^{-1} G)^{-1} G^T \Omega^{-1} \mathbf{r} \quad (3)$$

where $\Omega = \text{Diag}[\Omega^{(1)}, \Omega^{(2)}, \dots, \Omega^{(K)}]$ and $\Omega^{(g)} = \Omega(\mathbf{r}^{(g)})$ (Becker, 1992). The estimated asymptotic covariance matrix of $\hat{\rho}$ is

$$\hat{V} = (G^T \Omega^{-1} G)^{-1}. \quad (4)$$

To test the homogeneity of all correlation matrices across K studies, the test statistic

$$Q_{\text{GLS}} = \mathbf{r}^T [\Omega^{-1} - \Omega^{-1} G (G^T \Omega^{-1} G)^{-1} G^T \Omega^{-1}] \mathbf{r} \quad (5)$$

is approximately distributed as a chi-square with $\left(\sum_{g=1}^K \frac{p^{(g)}(p^{(g)} - 1)}{2}\right) - \frac{p(p-1)}{2}$ degrees of freedom in large samples (Becker, 1992; Hedges & Olkin, 1985).

Stage 2: Fitting SEMs

Becker (1992) proposed to use $\hat{\rho}$ and its asymptotic covariance matrix \hat{V} to fit linear models. In her example, regression models were fitted and confidence intervals were

obtained for the fitted models. Because \hat{V} is available, it is speculated that path models, confirmatory factor analytic (CFA) models, and SEM can also be fitted properly with the asymptotically distribution-free (ADF) method as the estimation method in SEM (see later discussion on TSSEM).

Becker and Fahrbach (1994) suggested that the performance of GLS using Fisher z scores was better than the GLS on the raw correlation coefficients in estimating the pooled correlation matrix. Because this study is focused not only on estimating a pooled correlation matrix but also on fitting SEM in Stage 2, the asymptotic covariance matrix of Fisher z scores cannot be readily used as the weight matrix for the analysis of SEM. In view of this, we do not include the GLS method for summarizing the Fisher z scores here.⁵

Although there are other modifications on the conventional GLS method (e.g., Becker & Fahrbach, 1994; S. F. Cheung, 2000), real applications with the modified GLS methods are rare. The original GLS method is still used by applied researchers as an alternative to the univariate methods (e.g., Geyskens, Steenkamp, & Kumar, 1998; Smith, Avis, & Assmann, 1999). To limit the scope of the current study, we, therefore, only study the conventional GLS method.

A Two-Stage Approach for Meta-Analytic SEM

In SEM terminology, we let \mathbf{x} be a $p \times 1$ random observed variable and μ and Σ be its $p \times 1$ mean vector and $p \times p$ population covariance matrix, respectively. It is assumed that the population covariance matrix Σ is a function of a $q \times 1$ vector of unknown parameters θ , that is, $\Sigma = \Sigma(\theta)$, where $\Sigma(\theta)$ can be a model of path analysis, CFA, or SEM (e.g., Bollen, 1989). The typical discrepancy function to be minimized to obtain the estimated parameter $\hat{\theta}$ is

$$F(\theta) = (s - \sigma(\theta))^T W^{-1} (s - \sigma(\theta)), \quad (6)$$

where s and $\sigma(\theta)$ are the $p^* \times 1$ vectors of $p^* = p(p + 1)/2$ elements obtained by stringing out the lower triangular elements, including the diagonals of the sample and implied covariance matrices S and $\Sigma(\theta)$, respectively. W is a $p^* \times p^*$ weight matrix, which is typically considered to be a fixed, possibly estimated from the sample, positive definite matrix (Browne, 1984). With any consistent weight matrix W , the test statistic T follows a chi-square distribution with $(p^* - q)$ degrees of freedom in large samples, that is,

$$T = (n - 1) F_{\min}(\theta) \sim \chi^2(p^* - q), \quad (7)$$

where $F_{\min}(\theta)$ is the minimum of $F(\theta)$ in Equation 6, n is the sample size, and q is the number of free parameters estimated. If data are assumed to be multivariate normally distributed, maximum likelihood (ML) and GLS with their test statistics T_{ML} and T_{GLS} are usually used (see Bollen,

1989, for the discrepancy functions for these estimation methods).

This estimation theory can be extended to the multiple-group analysis. Assume that there are K studies and each study has its own population covariance matrix, say, $\Sigma^{(1)}, \Sigma^{(2)}, \dots, \Sigma^{(K)}$. The sample sizes, distributional assumptions, and dimensions of $\Sigma(\theta)$ across the K studies can be different (Bentler, Lee, & Weng, 1987). Several researchers have proposed using multiple-group SEM to analyze SEM with missing data (e.g., Allison, 1987; Duncan & Duncan, 1994; B. Muthén, Kaplan, & Hollis, 1987). Simulation and empirical studies show that the multiple-group SEM is unbiased and efficient in handling missing completely at random (MCAR) and missing at random (MAR) data (e.g., Graham, Hofer, & MacKinnon, 1996; B. Muthén et al., 1987; Wothke, 2000). The general idea is to partition the data into a data set that contains both complete data and several data sets with different missing data patterns. By hypothesizing that the same model holds across the complete and incomplete data sets, the whole model can be estimated by applying appropriate equality constraints among different samples.

Let us illustrate with an example. Assume that a two-factor CFA model with two indicators per factor holds in two studies. All variables are present in the first study while only x_1 and x_3 are present in the second study. The CFA models in Studies 1 and 2 are

$$\begin{aligned} \Phi^{(1)} &= \begin{bmatrix} \phi_{11} & & & & \\ \phi_{21} & \phi_{22} & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix}, \Lambda^{(1)} = \begin{bmatrix} 1 & \lambda_{21} & 0 & 0 \\ 0 & 0 & 1 & \lambda_{42} \end{bmatrix}^T, \Psi^{(1)} \\ &= \text{Diag}[\psi_{11} \quad \psi_{22} \quad \psi_{33} \quad \psi_{44}] \text{ and} \\ \Phi^{(2)} &= \begin{bmatrix} \phi_{11} & & & & \\ \phi_{21} & \phi_{22} & & & \\ & & & & \\ & & & & \\ & & & & \end{bmatrix}, \Lambda^{(2)} = \begin{bmatrix} 1 & \text{---} & 0 & 0 \\ 0 & 0 & 1 & \text{---} \end{bmatrix}^T \text{ and } \Psi^{(2)} \\ &= \text{Diag}[\psi_{11} \quad \text{---} \quad \psi_{33} \quad \text{---}], \end{aligned}$$

where Λ , Φ , and Ψ are the factor loadings, factor covariance, and error variance matrices, respectively, and “—” represents the missing parameters. The model in Study 1 is identified with 1 degree of freedom, whereas the model in Study 2 is not identified by itself because of missing values.

⁵ When Fisher z scores are used in the GLS approach, the pooled Fisher z score and its asymptotic covariance matrix can be obtained. The asymptotic covariance matrix of Fisher z scores cannot be transformed into the asymptotic covariance matrix of correlation matrix easily. If the asymptotic covariance matrix is not used as the weight matrix in fitting SEM, we speculate that the performance of using Fisher z scores with the GLS approach would be similar to the univariate methods. An alternative approach is to model Fisher z scores with their asymptotic covariance matrix directly (Fouladi, 2000); however, the problem is that parameter estimates are not interpretable easily because they are then based on Fisher z scores rather than a correlation matrix.

By applying two-group SEM for these models, we can set up several between-group constraints, namely $\phi_{11}^{(1)} = \phi_{11}^{(2)}$, $\phi_{21}^{(1)} = \phi_{21}^{(2)}$, $\phi_{22}^{(1)} = \phi_{22}^{(2)}$, $\psi_{11}^{(1)} = \psi_{11}^{(2)}$, and $\psi_{33}^{(1)} = \psi_{33}^{(2)}$. The combined model is then identified with 4 degrees of freedom.

This approach seems appealing to researchers conducting MASEM because hypothesized models are tested directly and there is no need to estimate the pooled correlation matrix. The major drawback of this approach is that the homogeneity of correlation and covariance matrices is assumed in the proposed model. If the proposed model does not fit the data, it is not clear whether the misfit is caused by the heterogeneity of the correlation matrices or the misspecification of the structural model (see Hafdahl, 2001, on the discussion of benefits and limitations of this approach in conducting MA for exploratory factor analysis). Therefore, we modify this approach to pool and test the homogeneity of the correlation matrices at the first stage of our TSSEM approach. The hypothesized model is fitted at the second stage only when the correlation matrices are homogeneous.

Stage 1: Testing Homogeneity of Correlation Matrices and Estimating Pooled Correlation Matrix

Following the previous notation, let $\mathbf{x}^{(g)}$ be a $p^{(g)} \times 1$ observed vector and $\Sigma^{(g)}$ be its $p^{(g)} \times p^{(g)}$ population covariance matrix in the g th study, where $p^{(g)}$ is the number of observed variables. Note that $p^{(g)}$ equals p only when all the variables are present in the study. Then with a $p^{(g)} \times p$ selection matrix $M^{(g)}$ with the appropriate 0s and 1s as the elements, the observed incomplete data can be related to the complete data directly,

$$\mathbf{x}^{(g)} = M^{(g)}\mathbf{x}. \quad (8)$$

Note that the dimensions of M are different from those of G in the GLS method. The selection matrix G in GLS method is used to select correlation coefficients from a complete set of correlations, whereas the selection matrix M in the TSSEM method is used to select observed variables from a complete set of variables. We define missing data in terms of missing variables because variables are the usual unit for the analysis in SEM. If a variable is missing, all correlation coefficients associated with that variable are also missing. Missing correlation coefficients (e.g., r_{12} is missing only while other correlation coefficients associated with x_1 are present), rather than missing variables, are observed in MASEM sometimes. It is not easy to handle missing correlation coefficients in SEM.

The covariance matrix can be decomposed into the matrices of standard deviations and correlations,

$$\Sigma^{(g)} = D^{(g)}P^{(g)}D^{(g)T} \text{ and } \text{Diag}[P^{(g)}] = I, \quad (9)$$

where $D^{(g)}$ is the $p^{(g)} \times p^{(g)}$ diagonal matrix of standard

deviations and the $P^{(g)}$ is the $p^{(g)} \times p^{(g)}$ correlation matrix in the g th study, respectively (e.g., Bentler, 1995; Bentler & Lee, 1983; Jöreskog & Sörbom, 1996; Krane & McDonald, 1978).

When implementing this model using SEM software, it is easier to consider the model in Equation 9 from the perspective of the CFA model,

$$\Sigma(\theta)^{(g)} = \Lambda^{(g)}\Phi^{(g)}\Lambda^{(g)T} + \Psi^{(g)}, \quad (10)$$

where Λ , Φ , and Ψ are the factor loadings, factor covariance, and error variance matrices, respectively. Then it is clear that Equation 9 is a special case of the CFA model by applying the following constraints,

$$\Lambda^{(g)} \text{ is a } p^{(g)} \times p^{(g)} \text{ diagonal matrix,}$$

$$\Phi^{(g)} \text{ is a } p^{(g)} \times p^{(g)} \text{ standardized matrix,}$$

$$\text{i.e., } \text{diag}[\Phi^{(g)}] = I \text{ and}$$

$$\Psi^{(g)} \text{ is a } p^{(g)} \times p^{(g)} \text{ zero matrix.} \quad (11)$$

Therefore, the standardized factor correlation matrix $\Phi^{(g)}$ and the diagonal factor loading matrix $\Lambda^{(g)}$ represent the correlation matrix $P^{(g)}$ and standard deviation matrix $D^{(g)}$ in Equation 9, respectively (see M. W. L. Cheung & Chan, 2004; Raykov, 2001, for a similar model parameterization for testing correlation and multivariate hypotheses using SEM). To test the homogeneity of correlation matrices, researchers can set all the factor correlation matrices $\Phi^{(g)}$ equal. The estimate $\hat{\Phi}$ under these constraints is the estimate of the pooled correlation matrix P .

Because correlation matrices are usually used in MASEM, the elements of $D^{(g)}$ are allowed to be free. The estimated $D^{(g)}$ conveys no particular meaning. However, it is necessary to make the distribution theory of covariance matrix in SEM applicable to the correlation matrix (see Bentler & Lee, 1983; Krane & McDonald, 1978). If the measurement scales or the variances of the variables are comparable and interesting to the researchers, equality constraints on the standard deviations $D^{(1)} = D^{(2)} = \dots = D^{(K)}$ can also be applied. Then the hypothesis being tested is the equality of covariance matrices across studies.

By assuming that all the correlation matrices are homogeneous, we can use the multiple-group SEM techniques to obtain the pooled estimate of P . Let $h(\rho)$ be the set of between-group constraints on the equality of correlation coefficients present across the K studies. To test the homogeneity of all the correlation matrices, the number of constraints imposed is

$$\sum_{g=1}^K \frac{p^{(g)}(p^{(g)} - 1)}{2} - \frac{p(p - 1)}{2}.$$

By comparing the model with constraints on the equality of correlation matrices against the model without constraints (with 0 degrees of freedom), a chi-square difference test can be used to evaluate the appropriateness of these equality constraints. The test statistic is asymptotically distributed as a chi-square with

$$\sum_{g=1}^K \frac{p^{(g)}(p^{(g)} - 1)}{2} - \frac{p(p - 1)}{2}$$

degrees of freedom. Moreover, goodness-of-fit indices can also be used to evaluate the model fit. Because of the diagonal constraints on $P^{(g)}$, the estimate \hat{P} (or $\hat{\Phi}$ in SEM notation) is the pooled correlation matrix and the asymptotic covariance matrix of parameter estimates \hat{V} is the asymptotic covariance matrix of the pooled correlation matrix. Because the number of constrained parameters is the same as that in the GLS approach, the degrees of freedom of this test are exactly the same as those of the GLS test in Equation 5.

Let us consider the previous GLS example. Recall that the first study includes all variables while the other two studies are incomplete:

$$\mathbf{x}^{(1)} = [x_1 \ x_2 \ x_3 \ x_4]^T, \mathbf{x}^{(2)} = [x_1 \ x_2 \ x_3]^T, \\ \mathbf{x}^{(3)} = [x_1 \ x_2 \ x_4]^T, \text{ and}$$

$$P^{(1)} = \begin{bmatrix} 1 & & & \\ \rho_{21} & 1 & & \\ \rho_{31} & \rho_{32} & 1 & \\ \rho_{41} & \rho_{42} & \rho_{43} & 1 \end{bmatrix},$$

$$P^{(2)} = \begin{bmatrix} 1 & & \\ \rho_{21} & 1 & \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix} \text{ and } P^{(3)} = \begin{bmatrix} 1 & & \\ \rho_{21} & 1 & \\ \rho_{41} & \rho_{42} & 1 \end{bmatrix}.$$

$$\text{Then } M^{(1)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$M^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \text{ and } M^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

are the appropriate selection matrices. And

$$D^{(1)} = \text{Diag}[s_1 \ s_2 \ s_3 \ s_4], D^{(2)} = \text{Diag}[s_1 \ s_2 \ s_3] \\ \text{and } D^{(3)} = \text{Diag}[s_1 \ s_2 \ s_4]$$

are the matrices of standard deviations where s_i is the standard deviation for the i th variable. The degrees of freedom for testing homogeneity are

$$\left(\frac{4 \times 3}{2} + \frac{3 \times 2}{2} + \frac{3 \times 2}{2} \right) - \frac{4 \times 3}{2} \text{ or } 6.$$

When the hypothesis of homogeneity of the correlation matrices is not rejected, we can use the estimate of the pooled correlation matrix P to fit SEM in Stage 2. However, if the studies are heterogeneous, they cannot be aggregated legitimately. Researchers may classify the studies into homogeneous subgroups according to potential categorical moderators (e.g., Brown & Peterson, 1993; Brown & Stayman, 1992; Hom et al., 1992; Tett & Meyer, 1993; Verhaeghen & Salthouse, 1997). Alternatively, cluster analysis may be used to classify studies into relatively homogeneous subgroups (M. W. L. Cheung & Chan, in press). The moderators can then be used to explain the variation among the correlation matrices (Marín-Martínez & Sánchez-Meca, 1998; Viswesvaran & Sanchez, 1998). Categorical moderators are used generally in MASEM for two reasons. First, data for continuous moderators are usually not available from primary studies. Second, the aim of MASEM is to use the pooled correlation matrix and fit SEM. Using continuous moderators to explain variation among correlation coefficients is not typically the main concern of MASEM researchers. Based on these arguments, fixed-effects models are reasonably assumed in this study.

Stage 2: Fitting SEMs

After the analysis from Stage 1, we have the estimate of the pooled $p \times p$ correlation matrix P with its

$$\frac{p(p - 1)}{2} \times \frac{p(p - 1)}{2}$$

asymptotic covariance matrix of parameter estimates \hat{V} and the total sample size N , which equals the sum of all sample sizes, that is,

$$N = \sum_{g=1}^K n^{(g)}.$$

Because the diagonal elements in the pooled correlation matrix are fixed at 1.0, the discrepancy function given in Equation 6 cannot be applied directly because it is developed for analyzing covariance structures. Nevertheless, it can be easily modified for analyzing the correlation structure with

$$F(\gamma) = (\mathbf{r}^* - \rho(\gamma))^T \hat{V}^{-1} (\mathbf{r}^* - \rho(\gamma)), \quad (12)$$

where \mathbf{r}^* and $\rho(\gamma)$ are the $p' \times 1$ vectors of $p' = p(p - 1)/2$ elements obtained by stringing out the lower triangular elements, excluding the diagonals in the sample and the implied correlation matrices R and $P(\gamma)$, respectively. \hat{V} is the $p' \times p'$ weight matrix estimated from the first stage and γ is a structural parameter vector (Jöreskog, Sörbom, Du Toit, & Du Toit, 1999). The weight matrix can be any

consistent weight matrix in Equation 12, but we propose to use the asymptotic covariance matrix of the pooled correlation matrix \hat{V} from Stage 1 as the weight matrix in Stage 2 with the ADF estimation method in the TSSEM method. The fitting procedure can be implemented with SEM software such as LISREL (Jöreskog & Sörbom, 1996) and Mx (Neale, Boker, Xie, & Mase, 2003).⁶ The test statistic T_{ADF} and other goodness-of-fit indices can be used to test whether the proposed model fits the data satisfactorily.

Summaries of Steps in Conducting TSSEM

To summarize, here are the necessary steps in conducting TSSEM:

1. Data requirements: Correlation or covariance matrices with their corresponding sample sizes are needed for each study. Missing variables are allowed.
2. Stage 1 of TSSEM: The multiple-group CFA approach can be used to test the homogeneity of correlation matrices across studies. The pooled correlation matrix \hat{P} and its asymptotic covariance matrix \hat{V} can be obtained at this stage of analysis. If the homogeneity of the correlation matrices is not rejected, we can proceed to Stage 2 of TSSEM. If the homogeneity hypothesis is rejected, potential categorical moderators can be used to classify the studies into homogeneous subgroups.
3. Stage 2 of TSSEM: The pooled correlation matrix \hat{P} and its asymptotic covariance matrix \hat{V} are used as the input to the ADF estimation method as specified in Equation 12. The total sample size (N) of all studies is used as the sample size for fitting the model.

There are several advantages to using the TSSEM rather than the univariate methods. First, the TSSEM uses the total sample size and it is free from the ambiguity of choosing among different sample size values that have been proposed for the univariate approaches. In the TSSEM method, the total sample size is used and correct inferences, in terms of controlling Type I error, power, and standard errors of parameter estimates, can also be drawn. The issue of sample size choice is discussed later in the General Discussion and Future Directions section.

Second, the TSSEM method provides us with information about sampling variation in the pooled correlations, which is reflected in the weight matrix \hat{V} in Equation 12. If there is much (little) sampling variation in the pooled correlations, the asymptotic variances of the pooled correlations that are the diagonal elements of \hat{V} will also be larger (smaller). Because the weight matrix \hat{V} in Equation 12 is

inverted, less (more) weight will be given to elements with large (small) amounts of sampling variation. This is consistent with common meta-analytic techniques in which less (more) weight is given to studies with greater (smaller) sampling variances. Moreover, the covariances among the elements of the estimated pooled correlation matrix are also accounted for in the weight matrix \hat{V} .

One may argue that the multivariate GLS approach with its asymptotic covariance matrix as the weight matrix may also share similar advantages with the TSSEM. In our opinion, the major difference between these two methods lies in the estimation of \hat{P} and \hat{V} in Stage 1 because both methods are subject to the same Stage 2 analysis. If both methods give similar estimates of \hat{P} and \hat{V} , then their empirical performance for model fitting in Stage 2 should be similar.

Simulation Studies

Purpose of the Studies

Three simulation studies were conducted to compare the empirical performance of the four different approaches, namely univariate r , univariate z , GLS, and TSSEM. Study 1 investigated the empirical performance of each approach at Stages 1 and 2 under the homogeneous case. Specifically, Type I errors, parameter estimates, and their standard errors were examined when all the correlation matrices were sampled from the same population.

In Study 2, two different correlation matrices were used to simulate studies for heterogeneous cases. Because the current study focused on the fixed-effects model rather than the random-effects model, fixed parameters rather than random parameters were used. Although Study 2 uses two fixed population matrices for data generation, it represents a fixed-effects case with between-groups differences because the two correlation matrices are not equal. The homogeneity hypothesis posits that the simplest fixed-effect model, with only one common population matrix, applies to all studies.⁷ With a significance level of .05, Study 2 investigated the statistical power for rejecting the homogeneity hypothesis correctly at Stage 1.

Finally, Study 3 was used to study statistical power for rejecting misspecified models in Stage 2. A model is said to be misspecified when one of three conditions occurs: (a)

⁶ The ADF estimation method is called generally weighted least squares method in LISREL and asymptotic weighted least squares in Mx. Both LISREL and Mx expect that the asymptotic covariance matrix estimated from Stage 1 is multiplied by the sample size before being used as the weight matrix in the Stage 2 analysis (Neale et al., 2003, p. 41). EQS 5.7b (Bentler, 1995) and Mplus 3 (L. K. Muthén & Muthén, 2004) do not allow this approach.

⁷ We thank the associate editor for this suggestion.

One or more free parameters are estimated when their population values are zero, (b) one or more parameters are fixed to zero when their population values are nonzero, or (c) both conditions happen at the same time (Hu & Bentler, 1998). Because the population values of the misspecified paths of the first type of misspecified models are zeros, they are used to evaluate the accuracy of the z test of parameter estimates (Kaplan, 1989). The model fit with the second and third conditions is usually very poor, and model fit tests can be used to evaluate the statistical power for rejecting misspecified models. Misspecified models with the first and third conditions were used in Study 3.

Test Statistics Used in the Simulation Studies

For the univariate- r and univariate- z methods, BAI (Equation 1) was used as the criterion for rejecting the homogeneity of all correlation matrices in Stage 1. When fitting SEM with the univariate approaches, the ML estimation method with the arithmetic mean as the sample size was used in fitting SEM at Stage 2. The choice of sample size does not affect the parameter estimates in fitting SEM; however, it will affect the chi-square test statistics, some goodness-of-fit indices, and the standard errors of parameter estimates. This issue is discussed later. For the GLS approach, Q_{GLS} (Equation 5) was used to test the homogeneity of all correlation matrices in Stage 1 and the ADF method was used as the estimation method in Stage 2. For the TSSEM approach, the ML and the ADF estimation methods were used in Stages 1 and 2, respectively. The total sample size was used in fitting SEM for the GLS and TSSEM approaches in Stage 2.

For ease of manipulation, CFA models were used in these studies. In discussing the design and implementation of Monte Carlo experiments in SEM, Paxton, Curran, Bollen, Kirby, and Chen (2001) suggested that 500 replications are usually large enough to provide accurate statistical estimates. To provide more accurate and stable estimates, 1,000 replications were used in each condition.

Assessment of the Empirical Performance

Overall Model

Several criteria were used to evaluate the empirical performance of the different methods. The overall rejection percentage (RP) was used to assess the empirical rejection rates of the different methods.⁸ Because the artificial data are generated from known structures, the chi-square test statistics in Study 1 are expected to follow the chi-square distribution with the corresponding degrees of freedom. We compare the empirical means and standard deviations of these test statistics against their theoretical means, which

equal their degrees of freedom, and standard deviations, which equal $\sqrt{2df}$ (e.g., Freund, 1992).

Individual Parameter Estimates

The relative percentage bias of each parameter estimate was computed, which is defined as

$$B(\hat{\theta}) = \frac{\bar{\hat{\theta}} - \theta}{\theta} \times 100\%, \quad (13)$$

where θ is the population value of the parameter and $\bar{\hat{\theta}}$ is the mean of the estimates of the parameters across the 1,000 replications. Good estimation methods should have relative bias less than 5% (Hoogland & Boomsma, 1998).

The relative percentage bias of the standard error of each parameter estimate, which is defined as

$$B(\overline{SE}(\hat{\theta})) = \frac{\overline{SE}(\hat{\theta}) - SD(\hat{\theta})}{SD(\hat{\theta})} \times 100\%, \quad (14)$$

where $\overline{SE}(\hat{\theta})$ is the mean of the estimated standard errors and $SD(\hat{\theta})$ is the empirical standard deviation of the parameter estimates across 1,000 replications, was used to assess the accuracy of the standard error estimates in fitting SEM (Hoogland & Boomsma, 1998). The relative percentage bias of the standard errors of parameter estimates indicates whether the standard errors of parameter estimates accurately reflect the sampling variation in the estimates of the parameter. Good estimation methods should have relative bias less than 10% in their standard errors (Hoogland & Boomsma, 1998).

Study 1: Analysis of Stages 1 and 2 in the Homogeneous Case

Method

SAS/IML (SAS Institute, 1995) was used to generate correlation matrices with known data structures. Routines to analyze Stage 1 with the univariate- r , univariate- z , and GLS approaches were also coded in SAS/IML (M. W. L. Cheung, 2003), whereas LISREL 8.30 (Jöreskog et al., 1999) was used to analyze Stages 1 and 2 of the TSSEM and Stage 2 of the other approaches.

⁸ Confidence intervals for RP with 1,000 replications can be approximated by the normal distribution. The 95% and 99% acceptance regions of the RP for $\alpha = .05$ are $5.0 \pm 1.960 \times \sqrt{1000(0.05)(0.95)/10} = [3.65, 6.35]$ and $5.0 \pm 2.576 \times \sqrt{1000(0.05)(0.95)/10} = [3.22, 6.78]$, whereas the 95% and 99% acceptance regions for $\alpha = .01$ are $1.0 \pm 1.960 \times \sqrt{1000(0.01)(0.99)/10} = [0.38, 1.62]$ and $1.0 \pm 2.576 \times \sqrt{1000(0.01)(0.99)/10} = [0.19, 1.81]$.

By searching PsycINFO, with (meta-analysis OR meta-analytic) AND (structural equation model OR factor analysis OR path analysis) as keywords and other cross-reference searching manually, M. W. L. Cheung (2002) found 15 studies applying MASEM. Although these 15 studies are not exhaustive, they provide some ideas on how practitioners use MASEM in their research. M. W. L. Cheung (2002) found that the median numbers of variables and studies were 6 and 26, respectively.

A six-variable CFA model with three indicators per factor was used in this study (see Figure 1 for the model). Because there are no widely accepted models and parameters for simulation studies in MASEM (cf. S. F. Cheung, 2000; Hafdahl, 2001) and parameter estimates are generally unbiased in the presence of different amounts of measurement error (e.g., DeShon, 1998), we chose the factor loadings and error variances arbitrarily. The factor loadings for the variables were fixed at .80, .70, and .60 for each factor while the factor correlation was .30. The error variances were fixed correspondingly at .36, .51, and .64, so that the reliability coefficients for the variables were .64, .49, and .36. Consequently, the population covariance matrix, which is equivalent to the correlation matrix, was

$$P_1 = \begin{bmatrix} 1.000 & & & & & \\ .560 & 1.00 & & & & \\ .480 & .420 & 1.00 & & & \\ .192 & .168 & .144 & 1.00 & & \\ .168 & .147 & .126 & .560 & 1.00 & \\ .144 & .126 & .108 & .480 & .420 & 1.00 \end{bmatrix}$$

Multivariate normal data were generated based on the population values of P_1 . Correlations were computed from the raw data for the analysis.

Number of Studies (K)

Although the surveyed median number of studies used in MASEM was 26 in M. W. L. Cheung's (2002) study, a simulation study with so many generated studies would be computationally very intensive because of the large number of cross-group constraints involved in Stage 1 and the inversion of the weight matrix with the ADF estimation method in Stage 2 of the TSSEM (Bentler, 1995). This puts much computation demand on a simulation with more than 10 studies. To compromise on computation time and the generalizability of the simulation results, we chose three levels for K in the current study: 5, 10, and 15. These values of K are not atypical in MA, especially in the presence of moderators (Field, 2001).

Sample Size per Study

There were five levels for n : 50, 100, 200, 500, and 1,000. The condition of $n = 50$ has been used as the small-sample condition in CFA (e.g., Marsh & Hau, 1999), whereas $n = 1,000$ is generally considered large enough for most SEM applications (e.g., Hoogland & Boomsma, 1998). Moreover, for ease of manipulation, all the studies within a simulated MA had equal sample

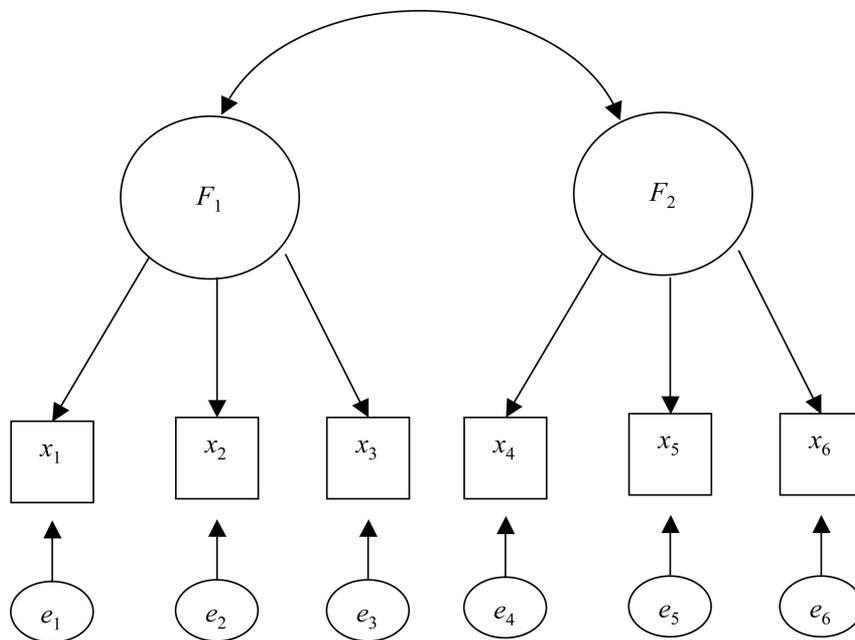


Figure 1. Confirmatory factor analytic model used in the simulation studies.

sizes. Although the sample sizes were equal in individual studies, the resultant sample sizes for individual correlation coefficients were different because of missing variables. Hence, the design was as follows: Number of Studies (3) \times Sample Size per Study (5). There were 15 conditions for each of the 4 methods (univariate r , univariate z , GLS, and TSSEM).

Missing Data

The patterns of the missing variables are shown in Table 1. The missing data patterns were used to create studies with missing correlations. When a variable is missing, all correlations associated with it will also be missing. Because the missing patterns were fixed a priori, the missing patterns were not related to the values of the missing correlations

Table 1
Design of Missing Patterns in Studies 1–3

Study	Observed variable					
No. studies = 5						
1	x_1	x_2	x_3	x_4	x_5	x_6
2		x_2	x_3	x_4	x_5	x_6
3	x_1		x_3	x_4	x_5	
4	x_1	x_2		x_4		x_6
5	x_1	x_2	x_3		x_5	x_6
No. studies = 10						
1	x_1	x_2	x_3	x_4	x_5	x_6
2		x_2	x_3	x_4	x_5	x_6
3	x_1		x_3	x_4	x_5	
4	x_1	x_2		x_4		x_6
5	x_1	x_2	x_3		x_5	x_6
6				x_4	x_5	x_6
7	x_1				x_5	x_6
8	x_1	x_2				x_6
9	x_1	x_2	x_3			
10		x_2	x_3	x_4	x_5	
No. studies = 15						
1	x_1	x_2	x_3	x_4	x_5	x_6
2		x_2	x_3	x_4	x_5	x_6
3	x_1		x_3	x_4	x_5	
4	x_1	x_2		x_4		x_6
5	x_1	x_2	x_3		x_5	x_6
6				x_4	x_5	x_6
7	x_1				x_5	x_6
8	x_1	x_2				x_6
9	x_1	x_2	x_3			
10			x_3	x_4	x_5	
11					x_5	x_6
12	x_1					x_6
13	x_1	x_2				
14		x_2	x_3			
15				x_4	x_5	

and the values of the other correlations. The missing data mechanism is, therefore, MCAR (Graham et al., 1996; Kaplan, 1995). Fewer variables were allowed to be missing with smaller numbers of studies than for larger simulated MAs. This appears to represent real applications: When researchers try to include more and more studies in MASEM, it is likely they will include studies with more missing variables.

Results and Discussion

Results of Stage 1: Testing Homogeneity of Correlation Matrices

Rejection percentages. The expected and observed RPs are shown in Table 2. The RP of the univariate- r , univariate- z , and TSSEM methods performed very well in most conditions. The percentages of significant tests for the univariate- r and univariate- z methods and the TSSEM methods were close to their nominal Type I error rates. On the other hand, the GLS method overrejected the true model frequently except in large samples and for large numbers of studies. These results are in line with the general findings of Becker and Fahrback (1994), S. F. Cheung (2000), and Hafdahl (2001).

Chi-square test statistics. Descriptive statistics for chi-square tests from the GLS and TSSEM approaches in Stage 1 are shown in Table 3. The results further provide a clear comparison between the GLS and TSSEM approaches at Stage 1 when the data are homogeneous. The Q_{GLS} of the GLS approach failed to follow the expected chi-square distribution in small to medium sample sizes. The observed means and standard deviations of Q_{GLS} were larger than the expected values. For instance, when $K = 10$ and $n = 100$, the observed mean and standard deviation of Q_{GLS} were 54.6 and 11.9, whereas the expected values were 50 and 10, respectively. The means and standard deviations of Q_{GLS} were both positively biased in all conditions. However, Q_{GLS} seems to converge to the expected means and standard deviations when the sample size is large (e.g., 500 or 1,000). On the contrary, the test statistic T_{ML} in TSSEM closely approximated the expected means and standard deviations in all conditions. The performance of the TSSEM was still very good even with small sample sizes per study (e.g., 50).

Results of Stage 2: Fitting Confirmatory Factor Analytic Models

Nonconvergent and improper solutions. First, all the methods were found to converge successfully and there were no improper solutions. This suggests that pairwise deletion may not lead to serious problems with nonconvergent or improper solutions in fitting SEM.

Table 2
Rejection Percentages of Stage 1 on Study 1
With Homogeneous Case

No. studies	<i>n</i> per study	Univariate <i>r</i> (BA1)	Univariate <i>z</i> (BA1)	GLS (Q_{GLS})	TSSEM (T_{ML})
$\alpha = .01$					
5	50	0.9	0.7	7.7 ^a	1.1
	100	0.5	0.8	3.2 ^a	0.8
	200	1.1	1.1	1.7 ^b	1.2
	500	1.1	1.1	0.8	0.7
	1,000	0.8	0.8	1.2	1.1
10	50	1.0	1.0	10.2 ^a	0.8
	100	0.9	0.9	4.6 ^a	1.5
	200	1.2	1.2	2.3 ^a	1.0
	500	0.9	1.0	2.0 ^a	1.4
	1,000	0.8	0.9	0.4	0.2 ^b
15	50	1.2	1.8 ^b	9.1 ^a	0.5
	100	0.7	1.2	4.0 ^a	0.8
	200	0.9	1.0	2.5 ^a	1.2
	500	0.7	0.9	1.6	1.0
	1,000	1.2	1.1	1.7 ^b	1.0
$\alpha = .05$					
5	50	4.2	4.6	19.1 ^a	5.1
	100	4.0	4.6	10.2 ^a	5.6
	200	5.1	5.0	7.8 ^a	5.3
	500	3.8	3.7	5.0	4.6
	1,000	4.3	4.7	5.5	4.6
10	50	4.6	5.5	23.5 ^a	5.8
	100	4.5	5.0	13.8 ^a	4.7
	200	5.5	5.0	7.1 ^a	4.2
	500	4.7	4.6	6.3	5.2
	1,000	4.3	4.7	5.5	4.6
15	50	4.7	5.3	23.0 ^a	3.3 ^b
	100	4.6	3.9	11.2 ^a	4.5
	200	4.6	4.6	7.9 ^a	4.6
	500	4.6	5.0	6.5 ^b	5.7
	1,000	4.7	4.3	4.8	4.4

Note. BA1 = Bonferroni-adjusted at-least-one approach; GLS = generalized least squares; Q_{GLS} = test statistic in GLS method; TSSEM = two-stage structural equation modeling; T_{ML} = test statistic in TSSEM method with maximum-likelihood estimation.

^a Rejection percentage falls outside the 99% acceptance regions. ^b Rejection percentage falls outside the 95% acceptance regions.

Chi-square test statistics and rejection frequencies. The RPs and their chi-square test statistics (T_{ML} and T_{ADF}) are shown in Tables 4 and 5, respectively. The univariate-*r* and univariate-*z* methods overrejected the true model significantly. When the significance level was .05, the univariate *r* and univariate *z* always overrejected the true model by factors of 3 to 7. The problems were even worse, by factors of 5 to 20, when the significance level was .01. Checking the means and standard deviations of

T_{ML} , the test statistic T_{ML} (from ML estimation under the univariate-*r* and univariate-*z* methods) appears to be positively biased. The range of the bias was from 20% to 80% across conditions. The larger the number of studies, the larger were the biases of the chi-square test statistics. Based on the results of this study, it can be concluded that the test statistics from the univariate approaches to fitting SEM are generally incorrect. Because the arithmetic mean was used as the sample size in the current study, the performance of the test statistics would have been worse if the total sample size had been used.

The GLS approach overrejected the true model, but the effects were less severe than the univariate-*r* and univariate-*z* methods for larger samples. The chi-square test statistics from the GLS approach were also biased, and the biases decreased when the sample size per study increased. The largest percentage of bias was about 60%. Similar to the results at Stage 1, the test statistics of GLS tended to converge to the expected means and standard deviations when the sample size increased, but they were generally too large in small to medium samples (50–500).

The TSSEM also overrejected the true models slightly in small samples (50 or 100). However, the performance of TSSEM was much better than other approaches in all conditions. When the sample size increased, the chi-square of TSSEM approached the expected values rapidly. Generally, the test statistics of the TSSEM method follow the theoretical values closely.

Parameter estimates. The relative percentage biases of parameter estimates are shown in Table 6 for some selected representative parameters to save journal space.⁹ Using bias less than 5% as the criterion for unbiased parameter estimates (Hoogland & Boomsma, 1998), the parameter estimates of the univariate *r* and univariate *z* were generally unbiased. The parameter estimates of the GLS approach were generally biased except when the sample size was large. Some parameter estimates were more than 10% above the expected values.

The parameter estimates of the factor loadings and factor correlations in the TSSEM method were generally unbiased; however, the estimates of error variances were slightly underestimated (these are not reported on Table 6). Jöreskog (1972) and Lee, Poon, and Bentler (1995) also found similar results of slightly underestimated error variances in their studies using the ADF estimation method. However, the underestimation of the error variance was not serious. It was usually less than 5% in our findings. Moreover, error vari-

⁹ Because of space limitations, only representative parameter estimates were reported in Tables 6 and 7. Readers can download the full table at <http://dx.doi.org/10.1037/1082-989X.10.1.40.supp>.

Table 3
*Chi-Square Statistics and Their Standard Deviations of Stage 1 on Study 1
 With Homogeneous Case*

No. studies	n per study	M		SD	
		GLS (Q_{GLS})	TSSEM (T_{ML})	GLS (Q_{GLS})	TSSEM (T_{ML})
5 ^a	50	37.0**	32.5	10.8	7.9
	100	34.8**	32.5	9.1	8.0
	200	33.2**	32.2	8.8	8.3
	500	33.1**	31.6	8.1	7.9
	1,000	32.4	32.1	8.2	8.1
10 ^b	50	59.0**	50.4	13.1	9.8
	100	54.6**	50.4	11.9	10.0
	200	52.3**	50.3	10.6	9.7
	500	50.9*	50.0	10.8	10.4
	1,000	49.4	49.1**	9.3	9.1
15 ^c	50	61.2**	52.2	13.2	9.4
	100	55.9**	51.6	11.9	10.1
	200	54.0**	52.1	11.0	10.2
	500	52.6	51.9	10.6	10.2
	1,000	52.3	51.8	10.2	9.9

Note. GLS = generalized least squares; Q_{GLS} = test statistic in GLS approach; TSSEM = two stage structural equation modeling; T_{ML} = test statistic in TSSEM approach with maximum-likelihood estimation method.

^a Expected means and standard deviations of chi-squares are 32 and 8. ^b Expected means and standard deviations of chi-squares are 50 and 10. ^c Expected means and standard deviations of chi-squares are 52 and 10.2.

* $p < .05$ for testing that the empirical means of Q_{GLS} and T_{ML} equal their theoretical means (i.e., $H_0: \mu = df$). ** $p < .01$ for testing that the empirical means of Q_{GLS} and T_{ML} equal their theoretical means (i.e., $H_0: \mu = df$).

ances are usually less important than other parameter estimates, such as the factor loadings and factor correlations.

Relative percentage bias of standard errors of parameter estimates. Table 7 shows the relative percentage bias of the standard errors. The relative biases of the univariate- r and univariate- z standard errors were relatively large and stable. The standard errors were usually overestimated for the factor loadings and factor correlations but underestimated for the error variances. The relative biases of the GLS method were extremely large in small samples; however, they decreased dramatically when the sample sizes increased. The standard errors were underestimated by GLS for the factor loadings and factor correlations. Put differently, the confidence intervals constructed using GLS estimates are much narrower than true values.

Except for the error variances, TSSEM usually had the smallest relative percentage bias in standard errors, and the bias decreased when the sample size increased. The standard errors of the error variances of TSSEM were generally overestimated. However, this may not be a serious problem because researchers are seldom interested in constructing confidence intervals or conducting hypothesis testing on the error variances, whereas the accuracy of the standard errors of the factor loadings, factor correlations, and path coefficients is usually critical to researchers.

Study 2: Analysis of Stage 1 in the Heterogeneous Case

Method

CFA Models

Two CFA models were used to generate correlation matrices from different population values. The first was the two-factor model used in Study 1 ($\Lambda_1, \Phi_1, \Psi_1,$ and P_1). The second was a one-factor model with $\Lambda_2 = [0.6 \ 0.5 \ 0.4 \ 0.6 \ 0.5 \ 0.4]^T$, $\Phi_2 = [1.00]$ and $\Psi_2 = \text{Diag}[0.64 \ 0.75 \ 0.84 \ 0.64 \ 0.75 \ 0.84]$. The added population correlation matrix was

$$P_2 = \begin{bmatrix} 1.000 & & & & & & \\ .300 & 1.00 & & & & & \\ .240 & .200 & 1.00 & & & & \\ .360 & .300 & .240 & 1.00 & & & \\ .300 & .250 & .200 & .300 & 1.00 & & \\ .240 & .200 & .160 & .240 & .200 & 1.00 & \end{bmatrix}.$$

The number of studies, sample sizes, and missing data pattern were the same as those in Study 1.

Table 4
Rejection Percentages of Stage 2 on Study 1
With Homogeneous Case

No. studies	<i>n</i> per study	Univariate <i>r</i> (T_{ML})	Univariate <i>z</i> (T_{ML})	GLS (T_{ADF})	TSSEM (T_{ADF})
$\alpha = .01$					
5	50	7.1 ^a	9.4 ^a	10.6 ^a	2.0 ^a
	100	6.4 ^a	7.3 ^a	4.7 ^a	1.6
	200	5.7 ^a	6.2 ^a	2.2 ^a	1.2
	500	5.9 ^a	6.3 ^a	2.1 ^a	1.8 ^b
	1,000	6.2 ^a	6.4 ^a	0.8	1.0
10	50	13.9 ^a	16.7 ^a	13.1 ^a	2.9 ^a
	100	13.2 ^a	14.7 ^a	4.8 ^a	1.1
	200	13.7 ^a	14.5 ^a	2.1 ^a	0.9
	500	12.7 ^a	12.8 ^a	1.9 ^a	1.4
	1,000	11.9 ^a	11.9 ^a	1.4	1.3
15	50	17.3 ^a	21.1 ^a	13.4 ^a	2.4 ^a
	100	16.3 ^a	17.6 ^a	4.9 ^a	1.7 ^b
	200	18.1 ^a	18.5 ^a	2.8 ^a	1.5
	500	17.1 ^a	17.7 ^a	1.9 ^a	1.8 ^b
	1,000	19.0 ^a	18.9 ^a	1.1	0.8
$\alpha = .05$					
5	50	18.8 ^a	21.7 ^a	24.6 ^a	7.7 ^a
	100	18.4 ^a	19.8 ^a	14.1 ^a	6.3
	200	16.9 ^a	17.5 ^a	8.8 ^a	5.5
	500	16.7 ^a	16.9 ^a	7.4 ^a	6.1
	1,000	17.1 ^a	17.0 ^a	5.8	5.1
10	50	29.4 ^a	32.5 ^a	27.6 ^a	8.2 ^a
	100	27.5 ^a	29.1 ^a	13.5 ^a	6.0
	200	27.5 ^a	28.2 ^a	7.9 ^a	4.9
	500	26.5 ^a	26.6 ^a	7.1 ^a	6.1
	1,000	27.9 ^a	28.1 ^a	6.4 ^b	5.5
15	50	35.8 ^a	38.8 ^a	28.9 ^a	7.7 ^a
	100	33.9 ^a	35.6 ^a	14.3 ^a	7.5 ^a
	200	35.1 ^a	36.3 ^a	9.8 ^a	4.9
	500	32.8 ^a	33.0 ^a	6.5 ^b	5.9
	1,000	34.7 ^a	34.4 ^a	4.8	4.9

Note. T_{ML} = test statistic at Stage 2 with maximum likelihood estimation method; GLS = generalized least squares; T_{ADF} = test statistic at Stage 2 with asymptotically distribution-free method; TSSEM = two-stage structural equation modeling.

^a Rejection percentage falls outside the 99% acceptance regions. ^b Rejection percentage falls outside the 95% acceptance regions.

Proportions of Heterogeneity

Two levels of heterogeneity were studied: small and large. Small heterogeneity means that 20% of the correlation matrices were drawn from P_2 , whereas 80% of the correlation matrices were drawn from P_1 . Large heterogeneity means that about 50% of the correlation matrices were drawn from P_1 and about 50% from P_2 .

Results and Discussion

Table 8 shows the rejection percentages of the different methods. Because the significance level was controlled as .05, the rejection percentages give us the empirical power of the different methods under different conditions. Generally, the statistical power of these tests increased when the sample size or number of studies increased. When the number of studies was 5, the GLS method had the highest power of rejecting the incorrect null hypothesis. When the numbers of studies were 10 and 15, the TSSEM method had the highest power to reject the incorrect null hypothesis. In fact, the GLS and TSSEM methods had higher statistical power than the univariate-*r* and univariate-*z* methods for rejecting incorrect homogeneity hypotheses. They all had high power to detect heterogeneous correlation matrices when the sample sizes per study were larger than 100 or 200. Because the data are heterogeneous, the estimated parameters convey no special meaning for comparisons. They are, therefore, not reported here.

Based on the results of Studies 1 and 2, the TSSEM approach was found to be the best among all the methods. The TSSEM approach has a good balance between Type I error control and statistical power for rejecting heterogeneity. Although the univariate-*r* and univariate-*z* methods are good at controlling the Type I error, they are less powerful than the GLS and TSSEM methods. The GLS approach has high power for rejecting the null hypothesis of homogeneity when samples are heterogeneous; however, it also renders an inflated Type I error rate at the same time.

Study 3: Analysis of Stage 2 With Misspecified Models

Method

Design

The number of studies, sample sizes, and missing data pattern were the same as those in Study 1. The same CFA model in Study 1 was used to generate the correlation matrices.

Misspecified Models

Two types of misspecification were studied. Originally, x_1 to x_3 and x_4 to x_6 were loaded on F_1 and F_2 , respectively (see Figure 1). The population factor loadings of F_1 to x_4 to x_6 were all zero. In the first misspecified model (Model 1), x_4 was misspecified as loaded on F_1 rather than F_2 . In the second misspecified model (Model 2), the model was over-parameterized with x_4 misspecified as double-loaded on both F_1 and F_2 rather than just F_2 . Because the population value of the factor loading from F_1 to x_4 is zero in Model 2, the estimated factor loading divided by its standard error

Table 5
Chi-Square Statistics and Their Standard Deviations of Stage 2 on Study 1 With Homogeneous Case

No. studies	<i>n</i> per study	<i>M</i>				<i>SD</i>			
		Univariate <i>r</i> (T_{ML})	Univariate <i>z</i> (T_{ML})	GLS (T_{ADF})	TSSEM (T_{ADF})	Univariate <i>r</i> (T_{ML})	Univariate <i>z</i> (T_{ML})	GLS (T_{ADF})	TSSEM (T_{ADF})
5	50	10.8**	11.2**	12.1**	8.7**	5.8	6.0	6.6	4.4
	100	10.9**	11.1**	10.0**	8.4**	5.5	5.6	5.1	4.2
	200	10.5**	10.6**	8.9**	8.1	5.5	5.6	4.6	4.1
	500	10.6**	10.6**	8.5**	8.2	5.5	5.6	4.5	4.3
	1,000	10.6**	10.6**	8.1	8.0	5.3	5.3	4.0	4.0
10	50	13.0**	13.6**	12.7**	8.7**	7.0	7.4	7.4	4.7
	100	12.6**	12.9**	10.0**	8.3*	6.5	6.7	5.1	4.1
	200	12.8**	12.9**	8.9**	8.1	7.2	7.3	4.4	4.0
	500	12.5**	12.5**	8.4*	8.1	6.9	6.9	4.3	4.2
	1,000	12.6**	12.6**	8.2	8.1	6.5	6.5	4.2	4.1
15	50	14.5**	14.8**	12.9**	8.8**	7.7	8.1	7.0	4.5
	100	13.8**	14.1**	10.2**	8.5**	7.1	7.2	5.3	4.2
	200	14.0**	14.2**	9.1**	8.2	7.7	7.8	4.7	4.1
	500	13.7**	13.7**	8.4**	8.1	7.5	7.5	4.4	4.3
	1,000	14.0**	14.0**	8.2	8.1	7.5	7.5	4.0	3.9

Note. Expected means and standard deviations of chi-squares are 8 and 4, respectively. T_{ML} = test statistic at Stage 2 with maximum likelihood estimation method; GLS = generalized least squares; T_{ADF} = test statistic at Stage 2 with asymptotically distribution-free method; TSSEM = two-stage structural equation modeling.

* $p < .05$ for testing that the empirical means of T_{ML} and T_{ADF} equal their theoretical means (i.e., $H_0: \mu = df$). ** $p < .01$ for testing that the empirical means of T_{ML} and T_{ADF} equal their theoretical means (i.e., $H_0: \mu = df$).

should follow a z distribution as a test of $H_0: \lambda_{41} = 0$ if the estimation is good (Kaplan, 1989). Thus, we could also test the accuracy of the z test in Model 2.

Results and Discussion

Model 1: Misspecified x_4 on F_1

The GLS and TSSEM approaches identified more non-convergent and improper solutions than the univariate- r and univariate- z approaches. The maximum percentage of nonconvergent and improper solutions found by the GLS and TSSEM approaches was about 25%, whereas the maximum percentage for the univariate approaches was only 4%. The percentage decreased when the sample size or number of studies increased. These results are consistent with the general findings of nonconvergent and improper solutions in SEM (e.g., Chen, Bollen, Paxton, Curran, & Kirby, 2002). Because the proposed model is misspecified, nonconvergent or improper solutions are indicative of the problem and force researchers to re-specify the models to obtain convergent or proper solutions. In every condition, all models were rejected at the 5% significance level. Hence, all approaches have extremely high statistical power for rejecting the misspecified model.

Model 2: Overmisspecified x_4 as Double-Loaded on F_1 and F_2

All the solutions converged without difficulty, and there was no improper solution. The RPs of the z test on testing $H_0: \lambda_{41} = 0$ are given in Table 9. As shown, only the RPs of the TSSEM method are close to the expected values. The univariate- r , univariate- z , and GLS methods overrejected the null parameter, especially when the number of studies was large. The performance of the GLS method began to converge to the expected RP only when the sample size was large ($n > 200$).

A Real Example: International Social Survey Program

To demonstrate how the proposed procedures could be applied to real data research, a real data set on work-related attitudes was considered (Inter-University Consortium for Political and Social Research, 1989). Persons aged 18 years and older from 11 countries were sampled based on multistage stratified probability sampling. The minimum and maximum sample sizes per country were 319 and 1,047, respectively. The total sample size was 7,155.

Table 6
Relative Percentage Bias of Parameter Estimates in Stage 2 of Study 1 With Homogeneous Case

n	Methods	5 Studies				10 Studies				15 Studies			
		λ_{11}	λ_{21}	λ_{31}	ϕ_{21}	λ_{11}	λ_{21}	λ_{31}	ϕ_{21}	λ_{11}	λ_{21}	λ_{31}	ϕ_{21}
50	Univariate <i>r</i>	-0.16	-0.32	0.25	-0.56	0.08	-0.32	-0.23	0.34	0.21	-0.78	-0.63	0.65
	Univariate <i>z</i>	0.26	0.16	0.84	-0.13	0.56	0.32	0.48	0.68	0.76	-0.15	0.02	0.93
	GLS	3.39	5.38 ^a	8.01 ^a	14.21 ^a	3.15	4.98	7.56 ^a	11.96 ^a	3.54	4.14	6.31 ^a	10.84 ^a
	TSSEM	1.12	1.44	2.23	2.79	0.88	1.15	2.03	2.49	1.14	0.53	1.48	2.11
100	Univariate <i>r</i>	0.24	-0.02	-0.92	0.10	-0.10	-0.49	-0.25	-0.17	0.08	-0.27	0.03	0.01
	Univariate <i>z</i>	0.45	0.23	-0.60	0.29	0.14	-0.19	0.08	0.01	0.34	0.03	0.37	0.13
	GLS	1.75	2.93	3.13	7.15 ^a	1.44	2.16	3.41	5.56 ^a	1.79	1.91	3.42	4.55
	TSSEM	0.70	0.93	0.42	2.15	0.33	0.23	0.86	1.03	0.57	0.31	1.01	1.06
200	Univariate <i>r</i>	0.08	-0.33	0.07	-0.07	0.12	-0.09	-0.42	0.01	-0.06	0.15	-0.18	0.00
	Univariate <i>z</i>	0.19	-0.19	0.22	0.01	0.25	0.07	-0.25	0.08	0.08	0.31	-0.01	0.05
	GLS	0.96	1.09	1.90	3.16	0.90	1.24	1.40	2.46	0.82	1.24	1.45	1.93
	TSSEM	0.39	0.08	0.63	0.94	0.34	0.27	0.11	0.51	0.25	0.39	0.30	0.25
500	Univariate <i>r</i>	-0.04	-0.04	0.02	-0.13	0.09	-0.16	-0.03	0.36	0.03	-0.13	0.06	-0.16
	Univariate <i>z</i>	0.00	0.01	0.08	-0.10	0.14	-0.09	0.04	0.39	0.08	-0.06	0.13	-0.13
	GLS	0.31	0.53	0.73	1.15	0.35	0.41	0.67	1.15	0.38	0.33	0.71	0.69
	TSSEM	0.11	0.13	0.25	0.26	0.14	0.02	0.16	0.30	0.14	0.00	0.24	-0.02
1,000	Univariate <i>r</i>	0.03	-0.09	0.03	0.06	0.07	-0.12	-0.08	-0.10	0.05	-0.12	0.03	-0.27
	Univariate <i>z</i>	0.06	-0.07	0.05	0.08	0.09	-0.09	-0.05	-0.09	0.07	-0.09	0.07	-0.26
	GLS	0.14	0.23	0.44	0.70	0.18	0.15	0.29	0.32	0.20	0.11	0.35	0.20
	TSSEM	0.03	0.04	0.19	0.26	0.07	-0.02	0.04	-0.08	0.08	-0.05	0.13	-0.15

Note. GLS = generalized least squares; TSSEM = two-stage structural equation modeling.

^a Values greater than 5%.

Method

Nine variables were selected for demonstration purposes. They were grouped into three meaningful constructs: job prospects (F_1), including job security (x_1), income (x_2), and advancement opportunity (x_3); job nature (F_2), including interesting job (x_4), independent work (x_5), help other people (x_6), and useful to society (x_7); and time demand (F_3), represented by flexible working hours (x_8) and lots of leisure time (x_9) (see Figure 2 for the proposed model).

Based on the nine variables, 11 correlation matrices were calculated, one for each country. These correlation matrices were treated as 11 studies. Then procedures of the univariate *r*, univariate *z*, GLS, and TSSEM were used to conduct meta-analytic factor analysis. To test the homogeneity of correlation matrices with the univariate-*r* and univariate-*z* approaches, the BA1 approach (Equation 1) was used.¹⁰ Because there were no missing correlations in the data set, all correlations were pooled across the 11 studies with the same cumulated sample size ($N = 7,155$). Thus, the total sample size (7,155), which was the same as the mean sample size across relationships, was used for the univariate-*r* and univariate-*z* approaches.

Results and Discussion

Stage 1: Testing Homogeneity of Correlation Matrices

The adjusted *p* values for the univariate-*r* and univariate-*z* approaches were smaller than .01. For the multivariate methods, the $\chi^2(360, N = 7,155)$ of the GLS approach (Q_{GLS}) was 992, $p < .01$, whereas the chi-square test statistic of the TSSEM approach (T_{ML}) was $\chi^2(360, N = 7,155) = 941, p < .01$. On the basis of these statistical tests, we may conclude that the correlation matrices are not homogeneous. However, significant results may have several alternative explanations (see the discussion on the empirical performance of different fit indexes later). Researchers seldom make decisions based on the chi-square test solely.

One of the advantages of using the TSSEM approach in testing the homogeneity of the correlation matrices is that goodness-of-fit indices other than the chi-square test are provided. Several goodness-of-fit indices in the TSSEM

¹⁰ Because of space limitations, the LISREL code was uploaded to <http://dx.doi.org/10.1037/1082-989X.10.1.40.supp>. A DOS program to facilitate the analysis can be downloaded at <http://www.hku.hk/~mikewlch/>.

Table 7
Relative Percentage Bias of Mean Standard Errors of Parameter Estimates in Stage 2 of Study 1 With Homogeneous Case

n	Methods	5 Studies				10 Studies				15 Studies			
		λ_{11}	λ_{21}	λ_{31}	ϕ_{21}	λ_{11}	λ_{21}	λ_{31}	ϕ_{21}	λ_{11}	λ_{21}	λ_{31}	ϕ_{21}
50	Univariate <i>r</i>	14.0 ^a	18.5 ^a	21.5 ^a	5.4	16.7 ^a	27.0 ^a	23.5 ^a	4.6	14.5 ^a	25.2 ^a	23.2 ^a	0.3
	Univariate <i>z</i>	13.3 ^a	18.5 ^a	21.5 ^a	4.5	15.8 ^a	26.0 ^a	23.1 ^a	3.4	13.8 ^a	24.8 ^a	22.5 ^a	-0.6
	GLS	-21.9 ^a	-22.7 ^a	-26.0 ^a	-29.3 ^a	-22.1 ^a	-23.3 ^a	-23.1 ^a	-28.8 ^a	-21.1 ^a	-21.5 ^a	-21.0 ^a	-26.3 ^a
	TSSEM	-5.1	-10.6 ^a	-4.0	-9.4	-3.2	-4.8	-6.8	-8.2	-5.1	-5.9	-2.6	-5.6
100	Univariate <i>r</i>	13.6 ^a	21.9 ^a	25.1 ^a	14.1 ^a	19.1 ^a	24.1 ^a	29.5 ^a	5.6	17.4 ^a	31.1 ^a	24.4 ^a	2.0
	Univariate <i>z</i>	13.5 ^a	22.0 ^a	24.6 ^a	13.5 ^a	18.7 ^a	24.5 ^a	29.6 ^a	5.3	16.9 ^a	31.3 ^a	23.8 ^a	1.6
	GLS	-11.6 ^a	-12.8 ^a	-12.7 ^a	-13.5 ^a	-5.9	-12.6 ^a	-8.1	-14.6 ^a	-12.4 ^a	-7.9	-11.6 ^a	-14.7 ^a
	TSSEM	-3.6	-1.7	-1.1	-1.0	1.3	-4.6	1.9	-3.0	-5.0	-0.6	-2.3	-4.3
200	Univariate <i>r</i>	16.4 ^a	20.3 ^a	15.6 ^a	13.9 ^a	15.7 ^a	26.9 ^a	23.4 ^a	9.6	24.0 ^a	32.2 ^a	26.1 ^a	7.1
	Univariate <i>z</i>	16.4 ^a	20.3 ^a	15.9 ^a	13.9 ^a	15.7 ^a	26.6 ^a	23.0 ^a	9.3	24.0 ^a	32.2 ^a	26.1 ^a	7.1
	GLS	-7.4	-7.0	-11.4 ^a	-8.2	-5.6	-5.9	-8.0	-5.6	-4.0	-5.0	-6.4	-3.7
	TSSEM	-3.7	-2.4	-7.6	-2.4	-2.1	-1.5	-3.2	-0.6	1.5	0.4	-0.7	1.2
500	Univariate <i>r</i>	17.7 ^a	25.7 ^a	21.9 ^a	10.8 ^a	23.4 ^a	32.4 ^a	24.7 ^a	8.7	26.5 ^a	29.9 ^a	23.9 ^a	1.6
	Univariate <i>z</i>	17.7 ^a	25.7 ^a	21.9 ^a	10.8 ^a	23.4 ^a	32.4 ^a	24.7 ^a	8.7	26.5 ^a	29.3 ^a	23.9 ^a	1.6
	GLS	-3.7	-1.0	-3.7	-5.8	0.0	-1.8	-4.0	-3.9	0.0	-5.3	-4.0	-6.0
	TSSEM	-1.8	1.5	-1.4	-2.6	1.7	0.0	-1.7	-1.8	2.4	-3.6	-2.3	-3.5
1,000	Univariate <i>r</i>	19.1 ^a	21.0 ^a	25.2 ^a	10.1 ^a	21.5 ^a	30.7 ^a	27.0 ^a	11.5 ^a	21.8 ^a	35.6 ^a	29.5 ^a	6.4
	Univariate <i>z</i>	19.1 ^a	21.0 ^a	25.2 ^a	10.1 ^a	21.5 ^a	30.7 ^a	27.0 ^a	11.5 ^a	21.8 ^a	33.6 ^a	29.5 ^a	6.4
	GLS	-2.0	-0.7	0.0	-4.6	-3.1	-3.3	-2.4	0.0	-0.8	0.0	0.0	-0.6
	TSSEM	-1.3	0.0	1.4	-3.6	-1.6	-2.5	-1.6	1.3	-0.8	1.8	1.7	0.6

Note. GLS = generalized least squares; TSSEM = two-stage structural equation modeling.
^a Values greater than 10%.

approach suggest that the model fit on the equality of correlation matrices is reasonable, for instance, the comparative fit index (CFI) = 0.94, nonnormed fit index (NNFI) = 0.94, and root mean square error of approximation (RMSEA) = 0.05. In order words, the correlation matrices are homogeneous based on conventional SEM rules of thumb or at least marginally homogeneous from the cutoff criteria suggested by Hu and Bentler (1999). Hence, we may proceed to the second stage to fit the CFA model.

Stage 2: Fitting CFA Model

After obtaining a pooled correlation matrix, it was used as input to fit the CFA model. The fit indices of these approaches are shown in Table 10. Based on these fit indices, the results suggest that the fit of the proposed model could be improved. For the purpose of demonstration, we compare these results at this stage. When compared with the TSSEM method, the chi-square test statistics of the univariate-*z* and univariate-*r* methods are 27% and 26% larger, respectively, than the chi-square test statistic of TSSEM, whereas the chi-square test statistics of the TSSEM and GLS methods are comparable. The results of this real example are consistent with the results of the simulation studies with large samples (e.g., *n* > 500). The chi-square test statistics for fitting SEM are largest in the univariate-*z* and univariate-*r*

methods followed by those of the GLS method, and they are smallest for the TSSEM method. The goodness-of-fit indices for the multivariate methods usually indicated better fit than did the univariate methods. However, before concluding that the results in the simulation studies can be generalized to real data analyses, more empirical studies should be conducted. The parameter estimates and their standard errors from the TSSEM method are also shown on Figure 2.

Because the variances and covariances of the variables were available, a reviewer suggested that we estimate a pooled covariance matrix with the multiple-group SEM approach and to fit a model using the ML estimation method. The fit indices for testing the homogeneity of the covariance matrices in Stage 1 were $\chi^2(450, N = 7,155) = 2,510, p < .01, CFI = 0.80, NNFI = 0.82,$ and $RMSEA = 0.09,$ whereas the fit indices for fitting the proposed model using ML estimation method in Stage 2 were $\chi^2(24, N = 7,155) = 1,599, p < .01, CFI = 0.83, NNFI = 0.75,$ $RMSEA = 0.10,$ and standardized root mean square residual (SRMR) = 0.06. These results indicated that the TSSEM approach fitted the data better than the approach using the covariance matrices. The reason may be that homogeneity of covariance matrices is generally more restrictive than homogeneity of correlation matrices. In real applications of MASEM, the variances of the variables may

Table 8
Rejection Percentages of Stage 1 on Study 2 in the Heterogeneous Case

No. studies	<i>n</i> per study	Univariate <i>r</i> (BA1)	Univariate <i>z</i> (BA1)	GLS (Q_{GLS})	TSSEM (T_{ML})
Small heterogeneity condition					
5	50	23.9	21.7	60.6	26.2
	100	55.7	84.8	64.6	57.7
	200	93.1	91.8	99.7	98.1
	500	100.0	100.0	100.0	100.0
10	50	36.1	30.0	84.0	95.3
	100	72.8	68.0	97.4	100.0
	200	99.0	98.7	100.0	100.0
	500	100.0	100.0	100.0	100.0
15	50	46.3	38.3	93.7	99.5
	100	83.8	88.2	99.7	100.0
	200	100.0	99.9	100.0	100.0
	500	100.0	100.0	100.0	100.0
Large heterogeneity condition					
5	50	28.1	31.4	75.5	49.9
	100	65.1	95.5	91.3	64.5
	200	96.8	97.5	99.9	99.9
	500	100.0	100.0	100.0	100.0
10	50	49.5	51.4	97.0	96.7
	100	93.6	93.6	100.0	100.0
	200	100.0	100.0	100.0	100.0
	500	100.0	100.0	100.0	100.0
15	50	61.4	60.4	99.2	99.5
	100	96.4	96.3	100.0	100.0
	200	100.0	100.0	100.0	100.0
	500	100.0	100.0	100.0	100.0

Note. BA1 = Bonferroni-adjusted at-least-one approach; GLS = generalized least squares; Q_{GLS} = test statistics in GLS method; TSSEM = two-stage structural equation modeling; T_{ML} = test statistics in TSSEM method with maximum likelihood estimation. The significance level was .05.

be neither available nor comparable across studies. Using correlation matrices may be a better choice in conducting MASEM.

Moreover, several multiple-group CFA models with different degrees of restriction were also fitted. The first model was the form-invariant model in which the factor loadings vary across studies. The second model was the factorial-invariant model, in which the factor loadings were invariant, and the third model was the full-invariant model, in which the factor loadings, factor covariances, and error variances were all invariant. The fit indices for the form-invariant model were $\chi^2(264, N = 7,155) = 2,057, p < .01, CFI = 0.82, NNFI = 0.74,$ and $RMSEA = 0.11,$ whereas for the factorial-invariant model and the full-invariant model they were $\chi^2(354, N = 7,155) =$

$2,449, p < .01, CFI = 0.79, NNFI = 0.76,$ and $RMSEA = 0.10$ and $\chi^2(474, N = 7,155) = 4,107, p < .01, CFI = 0.64, NNFI = 0.70,$ and $RMSEA = 0.12,$ respectively. The fit indices of the form-invariant model were only marginal, indicating that the proposed model did not fit equally well in all studies. It is of interest to observe that the fit indices of the TSSEM approach were even better than those of the form-invariant model. Because this is based on a real example, simulation studies are

Table 9
Rejection Percentages of Standard Z Scores With Misspecified Parameter at Stage 2 From Study 3

No. studies	<i>n</i> per study	Univariate <i>r</i> (T_{ML})	Univariate <i>z</i> (T_{ML})	GLS (T_{ADF})	TSSEM (T_{ADF})
$\alpha = .01$					
5	50	1.8 ^b	1.6	2.9 ^a	1.4
	100	0.8	0.7	2.0 ^a	0.7
	200	1.6	1.6	1.2	1.0
	500	0.8	0.6	0.5	0.6
	1,000	1.2	1.2	1.1	1.1
10	50	1.9 ^a	2.2 ^a	3.1 ^a	1.2
	100	1.8 ^b	1.8 ^b	1.7 ^b	0.6
	200	1.5	1.5	1.2	1.0
	500	1.3	1.3	1.2	1.3
	1,000	2.4 ^a	2.4 ^a	1.4	1.3
15	50	2.4 ^a	2.6 ^a	3.4 ^a	1.0
	100	2.4 ^a	2.3 ^a	1.9 ^a	1.3
	200	1.9 ^a	1.9 ^a	1.4	0.9
	500	1.8 ^b	1.9 ^a	1.0	1.0
	1,000	1.7 ^b	1.7 ^b	1.0	1.0
$\alpha = .05$					
5	50	6.8 ^a	6.5 ^b	9.8 ^a	5.7
	100	7.3 ^a	7.2 ^a	7.6 ^a	5.3
	200	5.6	5.6	6.5	5.6
	500	4.7	4.7	5.1	4.9
	1,000	5.7	5.7	4.6	4.3
10	50	8.0 ^a	8.4 ^a	7.4 ^a	4.7
	100	6.5 ^b	6.3	7.0 ^a	4.7
	200	5.2	5.3	5.1	4.9
	500	6.2	6.1	5.5	4.5
	1,000	7.3 ^a	7.3 ^a	5.5	5.3
15	50	9.2 ^a	9.6 ^a	10.0 ^a	6.3
	100	6.4 ^b	6.6 ^b	6.3	4.7
	200	8.0 ^a	8.1 ^a	7.0 ^a	5.9
	500	7.4 ^a	7.6 ^a	6.0	5.3
	1,000	6.4 ^b	6.3	4.8	4.7

Note. T_{ML} = test statistics at Stage 2 with maximum likelihood estimation method; GLS = generalized least squares; T_{ADF} = test statistics at Stage 2 with asymptotically distribution-free method; TSSEM = two-stage structural equation modeling.

^a Rejection percentage falls outside the 99% acceptance regions. ^b Rejection percentage falls outside the 95% acceptance regions.

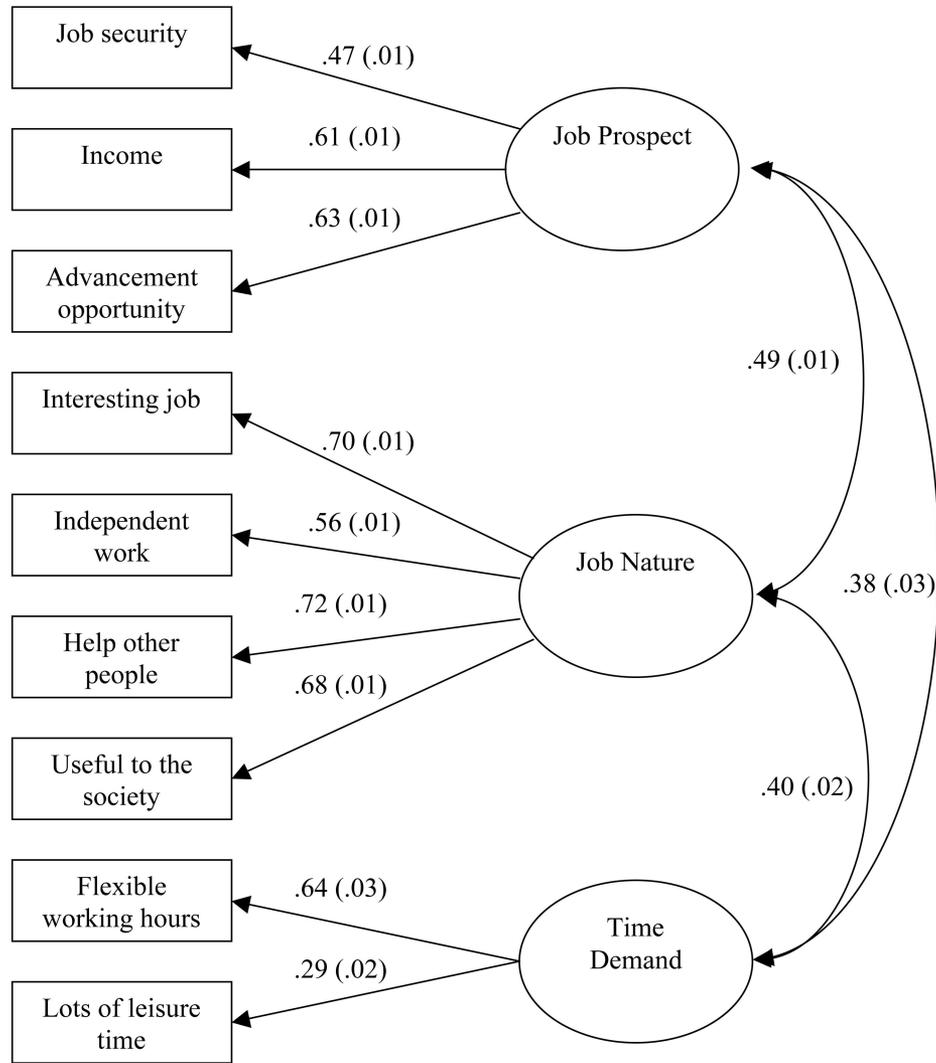


Figure 2. Confirmatory factor analytic model used in the real example with the parameter estimates with the two-stage structural equation modeling method. Measurement errors are not shown. Values in parentheses are standard errors.

Table 10
Goodness-of-Fit Indexes at Stage 2 of the Real Example

Goodness-of-fit indexes	Univariate <i>r</i>	Univariate <i>z</i>	GLS	TSSEM
$\chi^2(24, N = 7,155)$	1,607	1,622	1,282	1,277
CFI	0.83	0.83	0.88	0.87
NNFI	0.75	0.75	0.82	0.80
RMSEA	0.10	0.10	0.09	0.09
SRMR	0.06	0.06	0.07	0.07

Note. All *p* values for the χ^2 test statistics are smaller than .001. GLS = generalized least squares; TSSEM = two-stage structural equation modeling; CFI = comparative fit index; NNFI = nonnormed fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

required to verify whether well-fitted models in TSSEM or MASEM in general may or may not fit well in any or all studies.

General Discussion and Future Directions

After reviewing some current procedures (univariate *r*, univariate *z*, and GLS) and their problems for MASEM, a TSSEM approach was proposed. A multiple-group SEM approach is used to test the homogeneity of correlation matrices and to estimate the pooled correlation matrix in the first stage. The pooled correlation is used as the input, whereas its asymptotic covariance matrix is used as the weight matrix in the second stage.

Comparisons of Different Approaches

Univariate Approaches

The major advantage of the univariate approaches is their ease of application for applied researchers. Because the univariate-*r* and univariate-*z* approaches are well known, meta-analysts should have no difficulty in using them to estimate the pooled correlation matrix. Based on the current and other findings (e.g., S. F. Cheung, 2000; Hafdahl, 2001), the pooled correlation matrices in Stage 1 are generally unbiased, and the Type I error for homogeneity testing is also well-controlled for the univariate-*r* and univariate-*z* approaches, when S. F. Cheung’s BA1 approach is used.

Because of the good performance of the univariate approaches in Stage 1, it is questionable whether we can improve their poor performance in Stage 2 just by finding a better or more appropriate sample size. If this is possible, the modified univariate approaches could produce more accurate results for researchers. However, as we argue later, this may not be the best strategy. From Equation 7, it is clear that the sample size is directly involved in the calculation of all test statistics T (T_{ADF} , T_{ML} , or T_{GLS}) in SEM. Given the same data and model, the new test statistic T_{new} based on the new sample size n_{new} and the old test statistic T_{old} based on the old sample size n_{old} would be related by

$$T_{new} = \frac{n_{new}}{n_{old}} T_{old}^{.11} \quad (15)$$

It is not easy to choose the appropriate sample size for the univariate approaches. Some may argue that the chi-square test statistics (T_{ADF} , T_{ML} , or T_{GLS}) are not generally used as formal statistical tests in SEM. Different goodness-of-fit indices can be used to evaluate the model fit. However, Bollen (1990) pointed out that the sample size is also involved implicitly or explicitly in the calculation of most goodness-of-fit indices. Therefore, the sample size issue is unavoidable, even when using goodness-of-fit indices as the criterion in evaluating the model fit in SEM with the univariate approaches.

Apart from the overall model fit, the involvement of sample size in calculating the standard errors of parameter estimates complicates the issue. It can also be shown that (e.g., Bollen, 1989, Appendix 4B) when we change the sample size from n_{old} to n_{new} , the associated standard errors of the parameter estimates will also change to

$$SE_{new} = \sqrt{\frac{n_{old}}{n_{new}}} SE_{old}. \quad (16)$$

Therefore, the effects of sample size on the overall model fit and the standard errors of parameter estimates are opposite. If we want to increase the overall model fit by using a

smaller sample size, we risk losing the precision of the individual parameter estimates.

Hedges and Pigott (2001) argued that researchers should also pay more attention to the statistical power of the tests within MA. It is difficult to find an appropriate sample size that is correct for overall model fit (Type I error and power), goodness-of-fit indices, and the standard errors of parameter estimates. Based on our simulation studies, the arithmetic mean is not a good choice. The chi-square test statistics for assessing the model fit in the second stage of SEM are already inflated, and the *z* tests for the parameter estimates are also incorrect. Even though we could find a “better *N*,” the univariate approaches will still fail to take into account the variation and interdependence among the correlation coefficients and analysis of the correlation matrix as a covariance matrix.

GLS Approach

The performance of the GLS approach in estimating the pooled correlation matrix in Stage 1 depends on the statistical properties of the GLS approach and the approximation of the asymptotic covariance matrix of the correlation coefficients by Olkin and Siotani (1976). Based on previous studies (e.g., Becker & Fahrbach, 1994; S. F. Cheung, 2000; Hafdahl, 2001) and the current simulation studies, it is suggested that the finite-sample properties of this GLS approach on estimating the pooled correlation matrix are not very good.

Because the current proposed method uses the pooled correlation matrix and its estimated asymptotic covariance matrix as the input in fitting SEM in Stage 2, the poor performance of fitting SEM with GLS is expected. If the empirical performance of the GLS approach in estimating the pooled correlation and its asymptotic covariance matrix can be improved, it is possible that its performance in Stage 2 would also be improved significantly. Several suggestions have been made to improve the poor empirical performance of the GLS approach on correlation coefficients. For instance, Becker and Fahrbach (1994) suggested using the Fisher *z* scores rather than the raw correlation coefficients in the GLS approach (see also footnote 3). S. F. Cheung (2000) proposed to use some modified estimators, which are less affected by the sampling error, for the covariance matrix of the correlation coefficients. One such method that computes variances using mean correlations in place of individual study results is offered in Becker and Fahrbach (1994). It is worthwhile to pursue these directions in future MASEM research.

¹¹ More correctly, the calculation of test statistics and the standard errors are based on $n - 1$ rather than n . Because they are approximately the same for large sample sizes, n is used for simplicity.

TSSEM Approach

The TSSEM is based on the asymptotic distribution theory of the covariance matrix. Our simulation studies suggest that its finite-sample properties (for $n = 150$ to $n = 500$) are reasonably good. The ADF estimation method is used in Stage 2; therefore, a large sample size is usually required for good empirical performance (Hoogland & Boomsma, 1998; Hu, Bentler, & Kano, 1992). Several modified test statistics have been proposed to improve the finite-sample properties of the ADF method, for instance, the Satorra-Bentler scaled chi-square (T_{SB} ; Satorra & Bentler, 1988) and Yuan and Bentler's corrected chi-square (T_{YB} ; Yuan & Bentler, 1997). M. W. L. Cheung (2002) found with simulation studies that the empirical performance of T_{ADF} , T_{SB} , and T_{YB} was quite comparable in the Stage 2 of TSSEM. One speculation is that the resultant sample size in MASEM is usually large enough for the ADF method. Thus, the original T_{ADF} may be good enough for many meta-analytic applications.

Suggestions for Applied Researchers

On the basis of the results of this study, we have several suggestions for applied researchers conducting MASEM.

Stage 1: Estimating the Pooled Correlation Matrix

The results of simulation studies show that the empirical performance of the univariate- r and univariate- z and TSSEM approaches with respect to the Type I error of the homogeneity test, and their power for rejecting heterogeneous samples and parameter estimates, are very good, whereas the GLS approach is unsatisfactory in small to medium sample sizes. We suggest that the univariate approaches are usually good enough if researchers are only interested in summarizing multivariate correlation matrices.

Stage 2: Fitting SEMs

Because the goal of MASEM is not only to synthesize correlation matrices but to test proposed models using the pooled correlation matrix from Stage 1, the TSSEM approach should be used. Our simulation results reveal that TSSEM is the only method for which the test statistics for model testing are close to their theoretical values with other good statistical properties, whereas the univariate- r and univariate- z methods are too liberal in controlling the Type I error. On the other hand, the GLS approach is generally not recommended unless one has a large sample per study (e.g., $n = 1,000$).

Future Directions

Empirical Performance With Unbalanced Sample Sizes and Larger Numbers of Studies

For ease of manipulation, equal sample sizes across studies were examined in the current simulation studies. However, unbalanced sample sizes are more realistic in applied research. The relatively small number of studies ($K = 5, 10,$ and 15) compared with typical MASEM raises the issue of whether the precision (standard error) of the parameter estimates could have been impacted. The standard errors of the parameter estimates in the TSSEM approach depend on the total sample size, which is impacted by both the individual sample size per study and the number of studies. Although the number of studies in the current investigation was small, the total sample sizes were not small except in some conditions such as $K = 5$ and $n = 50$. We speculate that when the number of studies increases, the performance of TSSEM would be even better and the standard errors of the TSSEM approach would decrease. However, when the number of studies increases, it may also be more likely that the data will be drawn from heterogeneous populations. Thus, it will also be more likely to reject the homogeneity of correlation matrices. Then the usefulness of the fixed-effects models for large numbers of studies may be limited and random-effects models may be more useful.

Types of Missing Data

In MASEM correlations may be nonexistent because a particular researcher may not have measured the same variables as others did. This is not consistent with many typical mechanisms for missingness in primary studies. We suggest that these nonexistent (or nonreported) correlations can also be well defined within the framework of missing data. First, Pigott (1994) discussed the definitions of MCAR, MAR, and missing not at random (MNAR) in the context of MA. Her definitions are similar to those proposed by Little and Rubin (1987) for primary studies. Second, researchers sometimes try to maximize the quality of data with planned missing value patterns (e.g., Graham et al., 1996; Kaplan, 1995). The idea is that respondents are only required to complete some sets of items. Thus, there can be many nonexistent correlations in each subgroup of respondents because of the design. Graham et al. (1996) provided some examples of how to operationalize different types of missingness for these nonexistent variables. Indeed, research findings show that multiple-group SEM approaches are unbiased and efficient in handling such missing data if it is MCAR or MAR (Graham et al., 1996; Kaplan, 1995; B. Muthén et al., 1987; Wothke, 2000). However, it is unclear how the empirical performance of TSSEM would change when the data are MNAR.

Empirical Performance of Different Fit Indices

In the current study, the chi-square test statistics were used to evaluate model fit at Stages 1 and 2. This is appropriate because the data were generated from a multivariate normal distribution with a known structure. In real-life situations, however, it is dangerous to rely solely on the chi-square test to evaluate the model's goodness of fit because a significant chi-square may occur for a number of reasons (e.g., model misspecification, high statistical power of the test, violation of some underlying assumptions of the estimation method; Hu et al., 1992). More important, it is generally believed that models are only approximations to reality (e.g., Hu & Bentler, 1998). Simulation studies are needed to verify the empirical performance of different goodness-of-fit indices in MASEM.

Random-Effects Models for MASEM

The procedures studied here are based on the fixed-effects models in which the effect-size parameters are assumed to be fixed but have unknown constant values that can be estimated. For instance, the population model in Study 1 arose from a single population, whereas the population model in Study 2 could be considered a fixed-effects model with a binary moderator. Fixed-effects models are reasonable in MASEM when categorical moderators can be used to classify the correlation matrices into homogeneous subgroups before fitting SEM. It is possible that a model could fit well in all individual studies while the parameter estimates may actually differ across studies.¹² Using the terminology of measurement invariance, the form or configuration of the model may be the same for some individual studies, whereas the parameter estimates may be different across studies. By using the current approaches (univariate r , univariate z , GLS, and TSSEM) discussed here, such studies should be classified into different subgroups by successful identification of the moderators.

Alternatively, one may consider the random-effects models (Hedges & Vevea, 1998). Under the random-effects model, the parameters of the SEM are considered as random and varied across different studies. Hence, it would be possible to estimate the variance components of those parameters instead of testing the homogeneity of correlation matrices in Stage 1. Although the statistical development of random effects in SEM is still limited (see Rovine & Molenaar, 2000; Teachman, Duncan, Yeung, & Levy, 2001), it is our belief that as the application of MASEM becomes increasingly popular, more research studying random-effect models for MASEM can be expected.

References

- Allison, P. D. (1987). Estimation of linear models with incomplete data. *Sociological Methodology*, *17*, 71–103.
- Becker, B. J. (1992). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics*, *17*, 341–362.
- Becker, B. J. (1995). Corrections to “Using results from replicated studies to estimate linear models.” *Journal of Educational Statistics*, *20*, 100–102.
- Becker, B. J. (2000). Multivariate meta-analysis. In H. E. A. Tinsley & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 499–525). San Diego: Academic Press.
- Becker, B. J., & Fahrback, K. (1994, April). *A comparison of approaches to the synthesis of correlation matrices*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Becker, B. J., & Schram, C. M. (1994). Examining explanatory models through research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 357–381). New York: Russell Sage Foundation.
- Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Lee, S. Y. (1983). Covariance structure under polynomial constraints: Application to correlation and Alpha-type structure models. *Journal of Educational Statistics*, *8*, 207–222.
- Bentler, P. M., Lee, S. Y., & Weng, L. J. (1987). Multiple population covariance structure analysis under arbitrary distribution theory. *Communications in Statistics: Theory and Methods*, *16*, 1951–1964.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. *Psychological Bulletin*, *107*, 256–259.
- Brown, S. P., & Peterson, R. A. (1993). Antecedents and consequences of salesperson job satisfaction: Meta-analysis and assessment of causal effects. *Journal of Marketing Research*, *30*, 63–77.
- Brown, S. P., & Stayman, D. M. (1992). Antecedents and consequences of attitude toward the ad: A meta-analysis. *Journal of Consumer Research*, *19*, 34–51.
- Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*, 62–83.
- Carson, P. P., Carson, K. D., & Roe, C. W. (1993). Social power bases: A meta-analytic examination of interrelationships and outcomes. *Journal of Applied Social Psychology*, *23*, 1150–1169.
- Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2002). Improper solutions in structural equation models:

¹² We thank an anonymous reviewer for this suggestion.

- Causes, consequences, and strategies. *Sociological Methods & Research*, 29, 468–508.
- Cheung, M. W. L. (2002). *Meta-analysis for structural equation modeling: A two-stage approach*. Unpublished doctoral dissertation, Chinese University of Hong Kong.
- Cheung, M. W. L. (2003). MetaAnalysisCorr: An SAS/IML program to synthesize correlation matrices with Hunter and Schmidt, Hedges and Olkin, and generalized least squares approaches. *Applied Psychological Measurement*, 27, 234–235.
- Cheung, M. W. L., & Chan, W. (2004). Testing dependent correlation coefficients via structural equation modeling. *Organizational Research Methods*, 7, 206–223.
- Cheung, M. W. L., & Chan, W. (in press). Classifying correlation matrices into relatively homogeneous subgroups: A cluster analytic approach. *Educational and Psychological Measurement*.
- Cheung, S. F. (2000). *Examining solutions to two practical issues in meta-analysis: Dependent correlations and missing data in correlation matrices*. Unpublished doctoral dissertation, Chinese University of Hong Kong.
- Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology*, 85, 678–707.
- Conway, J. M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology*, 84, 3–13.
- Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, 105, 317–327.
- DeShon, R. P. (1998). A cautionary note on measurement error corrections in structural equation models. *Psychological Methods*, 3, 412–423.
- Duncan, S. C., & Duncan, T. E. (1994). Modeling incomplete longitudinal substance use data using latent variable growth curve methodology. *Multivariate Behavioral Research*, 29, 313–338.
- Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, 6, 161–180.
- Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling*, 7, 356–410.
- Freund, J. E. (1992). *Mathematical statistics* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Geyskens, I., Steenkamp, J. E. M., & Kumar, N. (1998). Generalizations about trust in marketing channel relationships using meta-analysis. *International Journal of Research in Marketing*, 15, 223–248.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3–8.
- Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, 31, 197–218.
- Greenwald, A. G., Pratkanis, A. R., Leippe, M. R., & Baumgardner, M. H. (1986). Under what conditions does theory obstruct research progress? *Psychological Review*, 93, 216–229.
- Hafidahl, A. R. (2001). *Multivariate meta-analysis for exploratory factor analytic research*. Unpublished doctoral dissertation, University of North Carolina at Chapel Hill.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, 6, 203–217.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486–504.
- Hershberger, S. L. (2003). The growth of structural equation modeling: 1994–2001. *Structural Equation Modeling*, 10, 35–46.
- Hom, P. W., Caranikas-Walker, F., Prussia, G. E., & Griffeth, R. W. (1992). A meta-analytical structural equations analysis of a model of employee turnover. *Journal of Applied Psychology*, 77, 890–909.
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research*, 26, 329–367.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424–453.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Hu, L., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, 112, 341–362.
- Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257–266). Hillsdale, NJ: Erlbaum.
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- Inter-University Consortium for Political and Social Research. (1989). *International Social Survey Program: Work orientation*. Ann Arbor, MI: Author.
- Jöreskog, K. G. (1972). Factor analysis by generalized least squares. *Psychometrika*, 37, 243–260.
- Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: A user's reference guide*. Chicago, IL: Scientific Software International.

- Jöreskog, K. G., Sörbom, D., Du Toit, S., & Du Toit, M. (1999). *LISREL 8: New Statistical features*. Chicago: Scientific Software International.
- Kaplan, D. (1989). A study of the sampling variability and z-values of parameter estimates from misspecified structural equation models. *Multivariate Behavioral Research*, *24*, 41–57.
- Kaplan, D. (1995). The impact of BIB spiraling-induced missing data patterns on goodness-of-fit tests in factor analysis. *Journal of Educational and Behavioral Statistics*, *20*, 69–82.
- Krane, W. R., & McDonald, R. P. (1978). Scale invariance and the factor analysis of correlation matrices. *British Journal of Mathematical and Statistical Psychology*, *31*, 218–228.
- Lee, S. Y., Poon, W. Y., & Bentler, P. M. (1995). A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, *48*, 339–358.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201–226.
- Marín-Martínez, F., & Sánchez-Meca, J. (1998). Testing for dichotomous moderators in meta-analysis. *The Journal of Experimental Education*, *67*, 69–81.
- Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling*, *5*, 22–36.
- Marsh, H. W., & Hau, K. T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 251–284). Thousand Oaks, CA: Sage.
- Miller, N., & Pollock, V. E. (1994). Meta-analytic synthesis for theory development. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 457–483). New York: Russell Sage Foundation.
- Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, *51*, 431–462.
- Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide* (3rd ed.). Los Angeles, CA: Muthén & Muthén.
- National Research Council. (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academy Press.
- Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (2003). *Mx: Statistical modeling* (6th ed.). Richmond: Virginia Commonwealth University, Department of Psychiatry.
- Olkin, I., & Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. In S. Ideka (Ed.), *Essays in probability and statistics* (pp. 235–251). Tokyo: Shinko Tsusho.
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*, 287–312.
- Pigott, T. D. (1994). Methods for handling missing data in research synthesis. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 163–175). New York: Russell Sage Foundation.
- Premack, S. L., & Hunter, J. E. (1988). Individual unionization decisions. *Psychological Bulletin*, *103*, 223–234.
- Raykov, T. (2001). Testing multivariate covariance structure and means hypotheses via structural equation modeling. *Structural Equation Modeling*, *8*, 224–256.
- Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology*, *52*, 59–82.
- Rovine, M. J., & Molenaar, P. C. M. (2000). A structural modeling approach to a multilevel random coefficients model. *Multivariate Behavioral Research*, *35*, 51–88.
- SAS Institute. (1995). *SAS/IML software: Changes and enhancements, through release 6.11*. Cary, NC: SAS Institute.
- Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *Proceedings of the American Statistical Association, Business and Economics Sections* (pp. 308–313). Alexandria, VA: American Statistical Association.
- Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist*, *47*, 1173–1181.
- Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, *71*, 432–439.
- Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods*, *1*, 47–65.
- Smith, K. W., Avis, N. E., & Assmann, S. F. (1999). Distinguishing between quality of life and health status in quality of life research: A meta-analysis. *Quality of Life Research*, *8*, 447–459.
- Teachman, J., Duncan, G. J., Yeung, W. J., & Levy, D. (2001). Covariance structure models for fixed and random effects. *Sociological Methods & Research*, *30*, 271–288.
- Tett, R. P., & Meyer, J. P. (1993). Job satisfaction, organizational commitment, turnover intention, and turnover: Path analyses based on meta-analytic findings. *Personnel Psychology*, *46*, 259–290.
- Tremblay, P. F., & Gardner, R. C. (1996). On the growth of structural equation modeling in psychological journals. *Structural Equation Modeling*, *3*, 93–104.
- Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age-cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin*, *122*, 231–249.
- Viswesvaran, C., & Ones, D. S. (1995). Theory testing: Combining

- psychometric meta-analysis and structural equations modeling. *Personnel Psychology*, 48, 865–885.
- Viswesvaran, C., & Sanchez, J. I. (1998). Moderator search in meta-analysis: A review and cautionary note on existing approaches. *Educational and Psychological Measurement*, 58, 77–87.
- Wothke, W. (1993). Nonpositive definite matrices in structural modeling. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 256–293). Newbury Park, CA: Sage.
- Wothke, W. (2000). Longitudinal and multigroup modeling with missing data. In T. D. Little, K. U. Schnable, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 219–240). Mahwah, NJ: Erlbaum.
- Yuan, K. H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvement. *Journal of American Statistical Association*, 92, 767–774.

Received October 28, 2002

Revision received September 3, 2004

Accepted October 27, 2004 ■

Low Publication Prices for APA Members and Affiliates

Keeping you up-to-date. All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the *American Psychologist* and *APA Monitor*. High School Teacher and International Affiliates receive subscriptions to the *APA Monitor*, and they may subscribe to the *American Psychologist* at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

Essential resources. APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the American Psychological Association*, and on dozens of new topical books each year.

Other benefits of membership. Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

More information. Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.