UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

THE PROBLEMS OF FIT INDICES ON REPLICATED SEM STUDIES

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

SEUNGHOO LEE
Norman, Oklahoma
2022

THE PROBLEMS OF FIT INDICES ON REPLICATED SEM STUDIES

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Jorge Mendoza, Chair
Dr. Glenn Leshner
Dr. Lori Snyder
Dr. Hairong Song
Dr. Robert Terry

Abstract

There has been a research gap in examining fit indices under the context of reproducing the result of structural equation modeling (SEM) since a replication attempt revisited not many SEM studies. Two simulation studies were conducted to examine the distribution of fit indices of SEM on replicated samples. The first simulation chose three examples from social science literature to mimic replication attempts and found that the distribution of some indices shifted away from the original value. Specifically, the fit indices that use chi-square in their formulation consistently indicated a worse fit between the model and the data in a large proportion of replication attempts. Meanwhile, relative fit indices that use log-likelihood values such as AIC and BIC were less affected by replication, showing the distribution of replicated indices centered around the value from the original sample. The chi-squared family of fit indices showed an inferior fit than the original when one tries to replicate data using the observed moment matrix, even if the model fitted well to the original data. Using a baseline model log-likelihood, a new likelihood ratio LR_0 that resists the fit-worsening effect of replications is suggested. The second simulation that varied model specification, model complexity, and sample size confirmed the finding from the first study and examined the performance of the LR_0. The new likelihood ratio was much less affected by replication than the standard likelihood ratio, but its interpretability was limited. The diminishing effect on fit indices in replicated samples implies that one should interpret them carefully.

Table of Contents

The replicability of research findings is one of the key features of science that distinguishes it from pseudo-science. However, replicability has been questioned in many areas of science, especially in Psychology, by concerned researchers in the past nearly two decades (Ioannidis, 2005; Klein et al., 2014, Open Science Collaboration, 2015). While it is still an ongoing endeavor to define replicability, two distinctive categories of replication have emerged. Direct replication emphasizes the exact do-over of research, while conceptual replication uses a different procedure from the original study and often includes an expansion of context (Koul, Becchio, & Cavallo, 2018, Zwaan, Etz, Lucas, & Donnellan, 2017). In the academic culture of novelty, replication studies have been neglected by journals. (Everett & Earp, 2015) Such trends started to change with reformation in publishing practices like result-blind peer review (Allen & Mehler, 2019) and registered reports (Simons & Holcombe, 2014).

The reformation effort is also taking place in a quantitative analytic area. While the most significant movement revolves around the usage of the *p*-value (Wasserstein & Lazar, 2016; McShane, Gal, Gelman, Robert, & Tackett, 2019), other viewpoints or resolutions, such as the role of measurement error (Loken & Gelman, 2017), the use of metascience (Schooler, 2014), and reassessing the statistical power (Anderson & Maxwell, 2017) are also being discussed. As a part of these efforts, adopting cross-validation techniques to evaluate replicability is suggested (Koul et al., 2018).

While there are apparent similarities between replicability and cross-validation, we would like to clarify their differences. They are similar in that both involve two or more samples and compare the estimates of interest. However, they differ in their purpose. To validate a set of regression weights, one might use a different sample from the same population or divide one into two parts. In either case, one sample is used to calibrate the regression equation and another to evaluate it. These regression coefficients are used to predict the criterion in another sample. The purpose of cross-validation is to examine whether the predicted values – calculated by using the coefficients from the calibration sample – correlate well to the criterion. Meanwhile, replication is a more overarching concept. Although there are different opinions about what actually counts as replicated research (Simonsohn, 2013; Verhagen & Wagenmakers, 2014), the general purpose of replication is to confirm the previous scientific findings rather than to examine the predictive effectiveness of coefficients.

Because of the conceptual differences, the interpretation of related analyses should be different. Even if the research was a direct replication attempt and the analytic strategy was identical to cross-validation, the interpretation of the result focusing on replication usually pays attention to the significance of coefficients of focal variables because that is often a criterion of whether a hypothesis is supported or not. Cross-validation studies, on the other hand, usually focus on the usefulness of regression equations, which often boils down to the evaluation of the usefulness of calibrated coefficients.

The difference between replication and cross-validation becomes more apparent when a study uses covariance structure analysis, such as confirmatory factor analysis (CFA) or structural equation modeling (SEM). Cross-validation was initially developed to examine the predictive validity of a regression equation, and the idea expanded to the covariance structure analysis (Browne, 2000). The main idea of the covariance structure analysis is to find a set of parameters $\widehat{\boldsymbol{\theta}}$ that makes the discrepancy between the model-implied covariance matrix $\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)$ and the sample covariance matrix $\boldsymbol{S}$ minimum. Cross-validation on the covariance structure analysis compares the model-implied covariance matrix estimated on a calibration sample $\boldsymbol{\Sigma}_{\mathrm{C}}\left(\widehat{\boldsymbol{\theta}}\right)$ to the covariance matrix of a validation sample $\boldsymbol{S}_{\mathrm{V}}$. (This will be discussed further in the later section.)

Ultimately, cross-validation on a covariance matrix examines the usefulness of the estimated value of the parameters. On the other hand, a replication study does not necessarily have to look into the specific value of the parameters. The research interests of a replication study can span from examining the utility of parameter estimates to investigating whether an effect(s) from the original study can be found in another context. In other words, a replication study on a moment matrix could, but does not need to, fix the parameter estimates to specific values.

Despite the recent increasing trend in replications, research that uses SEM has seen fewer attempts (Goodboy & Kline, 2017; Babin & Svensson, 2012). While there have been cautionary notes for readers and reviewers of literature using SEM that replication is strongly recommended, particularly when a model undergoes a series of modifications (Ullman & Bentler, 2009; Hermida, 2015; Babin & Svensson, 2012), such research is rarely revisited by replication attempts (Goodboy & Kline, 2017). While no extensive review sheds light on this research gap, we speculate that one of the reasons would be a scarcity of guidelines on how to replicate a finding that has been found in the SEM framework. The current study aims to provide some helpful information on SEM replication by investigating what happens when one tries to replicate an SEM study.

Although it is not a focal interest of the current research, we offer speculation on degrees of SEM replication. Assume one tries to replicate research that used SEM techniques. There are several possible scenarios one can expect. First, one can expect the configuration of the model will be replicated. In this case, all paths in the original study will be included in the replication study, but no additional constraint will be given. In other words, the measurement and structural models are the same as the original, but all parameters will be freely estimated. The researcher would conclude that the replication is successful if all the significant path coefficients in the original are also significant in the replication attempt. SEM literature on replication does not have a terminology for such an attempt, but literature on measurement invariance (MI) deals with a similar concept called configural invariance. It is similar in that it also imposes the same number of factors and patterns of loadings on different groups.

Second, one can expect to replicate the sizes of the effects. In this case, the magnitudes of effects are also of interest in addition to the configuration of the model. If the MI between the original and the replicated sample is met in the measurement part of the model, then the relationship between the latent variables can be a subject of interest. One can employ a multiple-group analysis technique to examine the equivalence of regression coefficients between variables, like the practice routinely done in measurement invariance (MI) literature.

Alternatively, one can also think of a mixture of the two cases above. An SEM study often contains more than one path between latent variables, and the MI of those variables may not always be guaranteed. In such cases, one can constrain the equivalence of estimates only where appropriate and free other estimates where coercing the equivalency is inappropriate.

The definition of successful replication depends on a goal a researcher sets according to the situation specific to the original. If one expects the pattern of significant paths to replicate but not the magnitude of the effects, then fitting only the model's skeleton and checking the hypothesized paths would determine the success of the replication attempt. On the other hand, if the importance of the study lies in the replication of effect sizes, the equivalency of the effects must be examined to confirm the successful replication.

In any case, the replication attempt is considered an independent trial of the original. A common practice of examining a model regarding its fit is required to advance further investigation on any substantial interpretation of the significances or magnitudes of effects. A model with a bad fit would be out of any further consideration. In this sense, satisfactory fit indices are vital to defining a successful replication. Before moving on to the examination of the estimation result, the fit indices are the factors to determine whether the replication attempt is successful or not. Satisfactory fit indices are not sufficient but necessary for successful replication.

Evaluations of fit indices are conventionally done by guidelines suggested by some researchers (West, Taylor, & Wu, 2012; Hu & Bentler, 1999; Kline, 2015). However, there is more than one way to evaluate fit indices when one wants to do so for the replication attempt. First, the fit indices of the replication attempt can be evaluated by comparing them to the guidelines as those of the original study had been. Second, one can compare the fit indices from the replication to the original study and look for equal or better indices. This approach seems somewhat conservative but reasonable for those looking for solid evidence of replication. However, we will argue that this is not an appropriate approach in a later section. Third, one can consider the replication attempt successful if the newly obtained fit indices fall into a range around the original fits that account for sampling error (i.e., confidence interval). This option is out of the scope of the current study for two reasons: a) not all fit indices have readily available confidence intervals, and b) building ones is not a trivial task considering that it involves not only the sampling variability but also the variability of the predicted covariance matrix. The latter will be discussed further in a later section.

The current study focuses on the model's fit indices rather than specific effects. Usually, the fit indices are a gateway to further interpretation of the SEM result. Thus, the characteristic of fit indices of replicated study is crucial in understanding and interpreting the result.

SEM and Fit Indices

The SEM, or covariance structure analysis in a more general term, is a technique that compares the observed data and the theoretical model. Usually, the theoretical model specifies the relationships between variables denoted by a set of parameters $\widehat{\boldsymbol{\theta}}$. In turn, the model is expressed as a matrix, with each element being a covariance that is a function of the parameters (the model-implied covariance matrix $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ or $\widehat{\boldsymbol{\Sigma}}$). If the parameters are unknown, they are estimated to minimize the discrepancy between $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$ and the sample covariance matrix $\boldsymbol{S}$. The most widely used discrepancy function in SEM is

$$F(\boldsymbol{S}, \widehat{\boldsymbol{\Sigma}}) = \ln|\widehat{\boldsymbol{\Sigma}}| - \ln|\boldsymbol{S}| + tr(\boldsymbol{S}\widehat{\boldsymbol{\Sigma}}^{-1}) - p, \qquad \text{(Eq 1)}$$

where p is the number of observed variables.

Chi-square (χ^2)

One way to assess the quality of the model is to examine how the model fits the data. To do so, many researchers suggested indices that measure the fit of the model and the data. Joreskog(1969) showed that f, the minimum value of Equation 1, times $(N-1)$ follows χ^2 distribution under the null that $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and the population covariance matrix $\boldsymbol{\Sigma}$ are equal with the degrees of freedom equals $p^* - q$, where p^* is the number of non-redundant elements of the covariance matrix, and q is the number of parameters in the model. (The p^* can be obtained by $p(p+1)/2$.) It is, however, not a widely used fit index by itself because the null hypothesis is likely to be rejected when N is large. Although the χ^2 test itself has limitations in empirical studies (Joreskog, 1969), it provides a basis for many fit indices.

Root Mean Square Error of Approximation (RMSEA)

The test statistic $T = (N-1)f$ follows the central χ^2 distribution under the null hypothesis, but under the alternative, it follows a non-central χ^2 distribution (Steiger & Lind, 1980), which has two parameters, the degrees of freedom k and the non-centrality parameter λ.

(As the mean of the non-central χ^2 distribution is $k + \lambda$, where the mean of the central χ^2 distribution is k.) The non-centrality parameter can be estimated by

$$\widehat{\lambda} = \frac{\left(\chi^2 - df\right)}{(N-1)},$$

where the *df* equals $p* - q$ (Steiger, 1989). This non-centrality parameter reflects how well the model fits the data. Based on this idea, the RMSEA measures the absolute badness-of-fit, decreasing with the improvement of fit. It is estimated as

$$\text{RMSEA} = \sqrt{\frac{max\left(\chi^2 - df,\, 0\right)}{df(N-1)}}\,.$$

The RMSEA is normed to be greater than or equal to 0 and is divided by the *df* to penalize overfitting.

Tucker-Lewis Index (TLI)

Another strategy to formulate the fit is to compare the hypothesized model with a baseline model, which is usually the most restrictive model. There are several ways to specify a baseline model, but the most widely used method is to make the observed variables mutually independent so that the reproduced covariances between observed variables would be zero (Widaman & Thompson, 2003). Naturally, the baseline model does not likely fit the data well, and the improvement from it is used as a measure of the fit of the hypothesized model. The TLI is one such index. It is defined as

$$\text{TLI} = \frac{\frac{\chi_0^2}{df_0} - \frac{\chi_H^2}{df_H}}{\frac{\chi_0^2}{df_0} - 1},$$

where the subscript 0 and *H* denote the baseline and the hypothesized model, respectively (Tucker & Lewis, 1973). The TLI gives the ratio of the distance of the true and the hypothesized model in terms of the ratio of χ^2 values obtained (from the central χ^2 distribution) and the *df*. The denominator is the distance between the baseline and the true model, while the numerator is the distance between the baseline and the hypothesized model. (The true model is "true" in the sense that the population would follow the model. Theoretically, it fits the data perfectly.) The index increases when the model fits the data better.

Comparative Fit Index (CFI)

Bentler (1990) suggested a fit index based on the idea of non-centrality, like the RMSEA. However, it also uses the notion of comparative fit that involves the baseline model. The comparison of non-centralities is the key feature of CFI. It is defined as

$$\mathrm{CFI} = \frac{max\left(\chi_0^2 - df_0, 0\right) - max\left(\chi_H^2 - df_H, 0\right)}{max\left(\chi_0^2 - df_0, 0\right)},$$

and increases as the model fits better to the data. The non-centrality of the true model is zero, which is hidden in the denominator. It compares two non-centrality parameters, one of the baseline and one of the hypothesized model.

Likelihood

While the likelihood is not exactly an indication of the fit between the model and the data, it is worth noting since it provides bases for a couple of indices. Maximum Likelihood (ML) estimation in a restricted parameter space finds a set of parameters $\widehat{\boldsymbol{\theta}}$ that maximize the likelihood function in Equation 1. Sometimes the statistical packages that include ML estimation give log-likelihood instead of likelihood, but they serve the same purpose in finding parameters that maximize the function. Specifically, the log-likelihood function in SEM is given

$$\mathrm{logl}(\boldsymbol{\theta} \mid data) = -\frac{N}{2} p \ln(2\pi) - \frac{N}{2} \ln|\boldsymbol{\Sigma}(\boldsymbol{\theta})| - \frac{N}{2} tr\left[\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\right], \qquad \text{(Eq 2)}$$

when the mean structure is not included in the model. Additionally, the unrestricted model assumes that the model-implied dispersion matrix equals the sample dispersion matrix. As a result, the unrestricted log-likelihood is reduced to

$$\mathrm{logl}(\boldsymbol{S}) = -\frac{N}{2}p\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{S}| - \frac{N}{2}p, \qquad \text{(Eq 3)}$$

(see Rosseel, 2021, for a detailed derivation.) The twice difference between the two log-likelihoods is equal to Equation 1 times *N*, and the minimum value of it follows χ^2 distribution. The maximized value of the likelihood function is used in several model selection indices like Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).

Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)

Based on the information theory, Akaike (1973) introduced AIC to measure the expected information loss caused by using a statistical model to represent reality in the population. The reality, or the population, is determined by an unknown underlying process. A statistical model usually tries to reveal the hidden relationships between variables. However, most of the time, the model fails to represent reality perfectly, and the loss of information occurs. The AIC estimates this information loss. It is defined as

$$AIC = -\ln\left(L_M\right) + q,$$

where L_M is the maximum value of the likelihood function. (Note that the negative log of L_M equals *f* above.) Usually, the exact amount of information loss is unknown, so the AIC can only tell the relative information loss between different models. Schwarz (1978) suggested BIC, a similar index to AIC, with an argument adopting a Bayesian perspective. The definition is

$$BIC = q\ln(N) - \ln\left(L_M\right),$$

and the only difference to the AIC is its penalty term for the model complexity.

Cross-validation and Replication

Although the cross-validation technique originated from the regression approach, it expanded to covariance structure analysis like confirmative factor analysis (CFA) or SEM (Browne, 2000). Such techniques find parameters $\widehat{\boldsymbol{\theta}}$ that minimize the discrepancy between the model-implied covariance matrix $\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)$ and the sample covariance matrix $\boldsymbol{S}$. These discrepancies are often evaluated with one or more of the fit indices. Cross-validation of the model is started by splitting the sample into two parts; the calibration and validation samples. Then the discrepancy between the model-implied covariance matrix from the calibration sample $\boldsymbol{\Sigma}_{\mathrm{C}}\left(\widehat{\boldsymbol{\theta}}\right)$ and the covariance matrix of the validation sample $\boldsymbol{S}_{\mathrm{V}}$ is examined. The subscript C and V represent that its scalar, vector, or matrix is from the calibration and validation samples. Note that a set of parameter estimates varies across samples because the calibration aims to minimize the discrepancy of the matrices. Thus, a small discrepancy between $\boldsymbol{\Sigma}_{\mathrm{C}}\left(\widehat{\boldsymbol{\theta}}\right)$ and the $\boldsymbol{S}_{\mathrm{V}}$ defines the potential usefulness of $\widehat{\boldsymbol{\theta}}$ in another sample. The cross-validation technique is used to examine the model's validity in a different context and to provide a mean of model selection alongside many fit indices (Browne, 2000).

Expected Cross-Validation Index (ECVI)

The Cross-Validation Index (CVI) uses the discrepancy function illustrated in Equation 1 as many other fit indices, but the critical difference is that it measures the discrepancy between $\boldsymbol{\Sigma}_{\mathrm{C}}\left(\widehat{\boldsymbol{\theta}}\right)$ and the $\boldsymbol{S}_{\mathrm{V}}$. However, this method split the sample into two parts, wasting the observations. To address this issue, Browne (2000) suggested the following index;

$$\mathrm{ECVI} = F\left[\boldsymbol{S}_{C}, \boldsymbol{\Sigma}_{C}\left(\widehat{\boldsymbol{\theta}}\right)\right] + N_{C}^{-1} df,$$

showed that the ECVI is the expected CVI over both calibration and validation samples when the sample sizes are equal.

Phase 1

First, we want to illustrate what a replicated study using SEM looks like. Most of the time, a replication attempt tries to answer either or both of the following questions: 1) are the hypotheses supported in the original study supported in the replication? 2) would the magnitude of effects found in the original study be replicated? In SEM studies, hypothesis testing and assessing the effect are done after fitting the model to the data. In other words, the model with bad fit indices is not a subject for further consideration. Thus, our interest lies in how the fit indices are distributed across multiple replication attempts.

Another proclaiming should be made on which type of replication we focused on in this study. As we briefly mentioned above, direct replication of a study would collect data with the same variables as the original and use the same model. Alternatively, a conceptual replication in which a researcher tries to expand the original findings to a new setting may use a different model incorporating variables that reflect the new context. While a conceptual replication might have the original model as a part of its research model, the added components will complicate the comparison of the original and the replicate. For example, a conceptual replication attempt might be interested in a boundary condition of the original phenomenon and could incorporate a moderator variable in their model. It is hard to decompose the difference in fit indices into a part due to the replication and a part due to the newly added component. For this reason, we dismiss a conceptual replication in the current study and focus on a direct replication, where the original and the replication are the same in their variables and the model specification.

Method

We chose three published studies using SEM to test their hypotheses and randomly generated data that resembles the original. They were selected to illustrate situations where the fit indices' levels differ from exceptional to acceptable to mediocre. Although the selection process was not exhaustive, no other factors were considered except that their method was SEM and their level of fit indices. The first example is Guido, Marcati, & Peluso's (2011) research on the marketing conception of entrepreneurs using Azjen's (1991) Theory of Planned Behavior. Their model includes four latent variables (Attitude, Subjective Norm, Perceived Behavioral Control, and Intention to perform the behavior) indicated by twelve observed variables. The model is specified to predict the Intention by three other latent variables with covariances between the predictors. A detailed model specification can be found in Figure A1 in Appendix. Their original data consists of 188 observations. Because the data was unavailable to the public, we used their reported means, standard deviations, and correlation matrix to conduct further simulations. The model shows excellent fit (CFI = 0.996, RMSEA = 0.023) and serves as an example of exceptional fit.

The second example was taken from Huth-Bocks, Levendosky, Bogat, and von Eye's (2004) infant-mother attachment research. Their final model contains 21 observed variables as the indicators of six latent variables (Maternal Attachment Experience, Prenatal Social Support, Postnatal Social Support, Prenatal Representations of Caregiving, Prenatal Risk Factors, and Infant-Mother Attachment). In their model, the Experience indirectly predicts Attachment via two paths: a) Prenatal Social Support, which leads to Postnatal Social Support, and b) Caregiving. In addition, the Risk Factors affect Caregiving. Figure A2 in Appendix shows a measurement and the structural model specification. Their model fitted the data with the size of 204 and showed acceptable fit indices values; RMSEA = .04 and CFI = .97.

We took research about mathematics learning by Passolunghi, Vercelloni, and Schadee (2007) as our third example. The research aimed to identify the precursors of mathematics learning and went through extensive model modifications and comparisons, cutting down more than half of the variables they collected in their final model. We used the full model (rather than the final model) as our example. This decision provided an illustrative example where the fit indices are insufficient. Despite the trend that research with such a bad fit would rarely get published, their full model is theoretically defensible while showing an inadequate fit. Therefore, we took the full model (Model 5, Passolunghi et al., 2007). It has 21 observed variables, including Vocabulary and Block Design, and seven latent variables, which are Span Forward, Phonology, Number, Comparison, Working Memory, Counting, and Mathematics, fitted to a sample data size of 170. (See Passolunghi et al., 2007 for detailed descriptions.) Their model specifies that Vocabulary predicts Span Forward, Phonology, Number, Comparison, Working Memory, and Counting, while Block Design only predicts Comparison, Working Memory, and Counting. Furthermore, Working Memory and Counting jointly predict Mathematics, creating indirect paths from the exogenous variables to the final outcome, Mathematics. The more detailed model specification can be found in Figure A3 in Appendix. It is worth noting that we failed to reproduce the same result as theirs in terms of the fit indices. We found worse fit index values than they did (CFI = .85, TLI = .81, RMSEA = .07). Again, we proceed with this model as it serves our purpose of illustrating the situation where the fit indices are mediocre.

Simulation

All simulations have two population moment matrix conditions across six different sample sizes (N = 100, 200, 400, 1,000, 2,000, and the original sample size). The first condition uses the observed covariance matrix to generate simulated data (and will be referred to as “observed” condition hereafter), while the second condition uses the predicted covariance matrix (also referred to as model-implied covariance matrix) for the same purpose (and will be called as "perfect-fit" condition afterward). The model will have a perfect fit to the predicted covariance matrix. As a result, the second condition illustrates "the upper bound" when one tries to replicate research.

The first ("observed") condition of simulations using each of the three examples above was done by following these steps: 1) reconstruct the covariance matrix using the standard deviations and the correlation matrix, 2) generate random numbers that follow a multivariate normal distribution with the means and the reconstructed covariance matrix, with the size according to the sample size conditions, 3) fit the original model to the replicated data and record the fit indices, 4) repeat the step two and three 1,000 times.

The main differences between the second ("perfect-fit") condition to the "observed" condition is in the first few steps described above. In the "perfect-fit" condition, we 1) reconstructed the moment matrix from the reported means, standard deviations, and correlation matrix, 2) fitted the original model to it to generate the predicted covariance matrix, 3) generated random numbers that follow a multivariate normal distribution, using the observed means and the predicted covariance matrix, with the size corresponding to the sample size conditions, 4) fitted the replicated data generated in step 3 and record the fit indices, and 5) repeated the step 3 and 4 1,000 times. Note that none of the models include a mean structure, so the means of the observed variable are not estimated. Therefore, we used the observed mean vector for data generation.

Each fitting to the replicated sample used ML estimation provided by the 'lavaan' package (Rosseel, 2012) in R (R Core Team, 2022). The replication attempts that failed to reach convergence are excluded from the results. We examined the influence of sample size on the distribution of the fit indices by setting it at the sample size of the original study, 100, 200, 400, 1,000, and 2,000, intended to cover the range of sample sizes that can often be found in the SEM literature. Note that since some fit indices are sensitive to sample size, we specified a sample size corresponding to each condition when we fit the model to the original covariance matrix for comparable results. For example, for the N = 100 condition, we used the original covariance matrix for fitting the model, but we specified the number of sample observations as 100 instead of the original sample size. The only condition identical in every aspect of the original is the original sample size condition.

Research Outcome

To examine how replication attempts affect the fit indices, we propose two types of satisfactory fit: a) a better index value than the conventional guideline of evaluating fit, and b) a better index value than the original. We use the recommendation suggested by West, Taylor, and Wu (2012). Note that not all fit indices have fixed recommended value; for example, less AIC indicates a better fit, but there is no specific value for AIC that tell readers a good fit. The threshold of acceptable values for CFI is .95 or greater. It is also .95 for TLI. The RMSEA threshold is .06 or smaller. Then, we calculated the percentage of each type of satisfactory fit for each fit index.

Results

The result showed an interesting phenomenon in the distribution of fit indices. Generally, the fit indices of simulated replications were worse than the original in the first two examples when the replicated samples were generated using the observed moment matrix of the original study. Notably, any chi-square-based fit indices, such as CFI, TLI, and RMSEA, showed worse performance on the replication. On the other hand, the fit indices based on the log-likelihood value (AIC and BIC) are less impacted by replication.

The "Observed" Condition

Because of the difference in the bases of data generation, the patterns in both conditions are quite different. We focus on the result of the "observed" condition first. In Example 1, χ^2, CFI, TLI, RMSEA, and ECVI of the original study were 54.21, 0.99, 0.99, 0.03, and 0.61, respectively, while the means of 1,000 replication from the observed covariance matrix and the original sample size (N = 188) were 104.47, 0.95, 0.94, 0.08 and 0.87, respectively, showing that the fit indices of replicated samples are worse than the original. A complete simulation result of the "observed" condition in Example 1 can be found in Table 1. Similar patterns can be found in Example 2. As the simulation result of the "observed" condition on Example 2 in Table 2 shows, the mean fit indices of the replicated sample (M_{χ^2} = 553.93, M_{CFI} = 0.86, M_{TLI} = 0.83, M_{RMSEA} = 0.10, and M_{ECVI} = 3.15) are worse than the fit indices of the original study (χ^2 = 363.58, CFI = 0.92, TLI = 0.91, RMSEA = 0.07, and ECVI = 2.23) when the sample size was equal to the original (N = 206).

The estimation details of Example 3 are not identical to Example 1 and Example 2, where the non-convergence replication and the inadmissible solutions (e.g., negative variance estimates) are dropped. Example 3 was selected to illustrate when the model fit is below the acceptable range. In the original study (Passolunghi et al., 2007), the model we selected was not their final model and had suboptimal fit indices. In consequence, we had trouble fitting the model to their original data. Specifically, the converged solution contained a negative variance on one of the latent variables. Typically, such a solution is not admissible, and the model would likely undergo some modifications. However, considering the current study's exploratory nature in a phenomenon that has not been reported before, we decided to report the fit indices of the converged solution and proceed with subsequent replication attempts. Like the original data producing an impermissible solution, most replicated samples generated unacceptable solutions. We have not included such replications in the results of Examples 1 and 2, but for illustration purposes, we decided to include the non-admissible results in Example 3. In Example 3, the fit indices (M_{χ^2} = 571.30, M_{CFI} = 0.64, M_{TLI} = 0.55, M_{RMSEA} = 0.12, and M_{ECVI} = 4.05) are worse than the original (χ^2 = 317.40, CFI = 0.83, TLI = 0.79, RMSEA = 0.07, and ECVI = 2.56) in the "observed" condition with the sample size of the original (N = 170). A detailed simulation result on Example 3 can be found in Table 3.

[INSERT TABLE 1 HERE.]

[INSERT TABLE 2 HERE.]

[INSERT TABLE 3 HERE.]

In contrast to the χ^2 family fit indices, the fit indices calculated from the maximized log-likelihood value do not show a diminishing pattern by replication attempt. As shown in Table 1, in the "observed" condition in Example 1, where the sample size is equal to the original (N = 188), the original value of the log-likelihood is -7719.76 while the mean log-likelihood of replicated samples is -7705.59 with a standard deviation (SD) of 33.87. The original value fell into an interval surrounded by less than a half SD from the replicated means. The AIC and BIC, consequently, showed similar patterns (the original AIC is 15499.52 while the mean AIC replicated is 15471.17 with an SD of 67.74, and the original BIC is 15596.61, while the mean BIC replicated is 15568.26 with an SD of 67.74). Similarly, in the original sample size condition of Example 2 in Table 2, the value of original log-likelihood, AIC, and BIC is -7868.80, 15833.59, and 15993.33, respectively, while the mean and the SD of those from replicated samples is M_{logL} = -7843.04, SD_{logL} = 48.75, M_{AIC} = 15782.07, SD_{AIC} = 97.50, M_{BIC} = 15941.81, and SD_{BIC} = 97.50.

Example 3 showed a vastly different pattern in the log-likelihood, AIC, and BIC. As shown in Table 3, when the sample size was equal to the original (N = 170), the values of the original are logL = -7483.77, AIC = 15085.54, and BIC = 15270.55, while the replicated distributions have M_{logL} = -6768.07, SD_{logL} = 47.20, M_{AIC} = 13654.14, SD_{AIC} = 94.41, M_{BIC} = 13839.15, and SD_{BIC} = 94.41. Notice that the original fit values were much worse than the replicated fits. We believe that this is not a subject of serious interpretation. Instead, we consider this as a quirk that happened to be in the original data. This matter will be addressed in the next section.

We calculated the percentage of replication attempts with a better fit than the fit from the original covariance matrix. We observed two common trends in this satisfactory fit percentage when we varied the sample sizes within the "observed" condition across all examples. First, none to a tiny percentage of replication attempts have better chi-square and chi-square-based fits (CFI, TLI, RMSEA, and ECVI) than the original when the sample size is equal to or under 400, as shown in Table 1 and Table 2. The percentage increased when the sample sizes were 1,000 and 2,000, but not by much. The largest percentage was 18.04%, observed in CFI and TLI in Example 1 when the sample size was 2,000. Second, the percentages of a better fit in the log-likelihood and the likelihood-based fits (AIC and BIC) tended to decrease as the sample size increased. For example, the percentage of replication with better AIC in the "observed" condition in Example 1 was 66.32% when the sample size was 188, and it decreased by 55.04% when the sample size was 2,000.

We compared the replicated fit indices to specific values considered "cut-off" criteria. West et al. (2012) suggested such criteria on many fit indices, including CFI, TLI, and RMSEA, to be 0.95, 0.95, and 0.06, respectively, to be considered a good fit. We defined another type of satisfactory fit according to this guideline in the previous section. Accordingly, we calculated the percentages of replications that had fit values exceeding these criteria. Generally, the magnitudes of this rate depended on the fit of the original study and the type of fit index. In Example 1, which was selected as a study with excellent fit, the rate of satisfactory fit ranged from 54.84% to 94.15% on CFI as the sample size increased, as shown in Table 1. On the other hand, the rate decreased from 32.97% to 2.62% on TLI with the increasing sample size. The satisfactory fit rate for RMSEA also decreased from 16.94% to 0.00%.

In Table 2 that contains the results of Example 2, the satisfactory fit rate using the threshold showed 0.00% in all χ^2-based fit indices in all "observed" conditions. That is, no replication attempts we tried reached a satisfactory level commonly used in SEM literature. We emphasize the importance of this result. In a small sample size condition (N = 100), the CFI, TLI, and RMSEA indicated excellent fit (1.00, 1.01, and 0.00, respectively) of the model to the data. Yet, when one tried to replicate the sample, the fit of the same model went down to an unacceptable level (M_{CFI} = 0.84, M_{TLI} = 0.82, and M_{RMSEA} = 0.10). Although they differ in magnitude, the deteriorating fit happened in all sample size conditions, leading to the result that no condition achieved satisfactory χ^2-based fits.

The pattern described above can be better understood with Figure 1, which visualizes the empirical cumulative distribution of fit indices for the replication samples. The x-axis represents each fit indices' value, while the y-axis represents the percentile of the value. Also, the solid curve lines represent the cumulative distribution of the fit index in the "observed" condition, while the dots on the lines signify the fit index value from the original covariance matrix. As shown in Figure 1, the chi-square values for the replicated samples are almost always distributed on the right-hand side of the original, showing that few replicated samples have better chi-square than the original. This pattern of worsened fit also can be found in the distributions of CFI, TLI, and RMSEA, compared to the dots in Figure 1.

On the other hand, the distribution of AIC and BIC in Figure 1 centered around the original values, showing clear contrasts to the shifted distribution of the chi-square-based fit indices away from the original values. Note that the seemingly vertical lines in the graphs for AIC and BIC is due to vast differences in the range of each distribution and the overall range of the scale. The patterns identical to Example 1 were found in Example 2. In Example 2, the chi-square and its descendants (CFI, TLI, and RMSEA) showed worse fits than the original, while the distribution of AIC and BIC are centered around in Figure 2. In Example 3, where the model had an inadmissible solution on the original data, the chi-square and the fit indices stemming from it showed shifted distributions away from the original values in Figure 3. Interestingly, the AIC and BIC distribution are much better than the original values. This will be discussed in the later section.

[INSERT FIGURE 1 HERE.]

[INSERT FIGURE 2 HERE.]

[INSERT FIGURE 3 HERE.]

A final note about the visualizations of the "observed" condition is on the pattern with varying sample sizes that were described above. As the sample size increased, the shift of the chi-square distribution away from its original value became less salient. (see Figure 1, Figure 2, and Figure 3.) In other words, the distribution of the chi-square moved toward its original value as the sample size increased. A similar pattern was found in the likelihood fit indices (AIC and BIC). Specifically, the centers of distribution of AIC and BIC are located slightly left of the original values when the sample size was small, but they moved toward the original values as the sample size increased.

The "perfect-fit" Condition

The purpose of running a simulation on the "perfect-fit" condition was to illustrate what one can expect to see under the "ideal" situation when one tries to replicate an SEM. In the "perfect-fit" condition, we used a model-implied (predicted) covariance matrix as our bases for generating replication samples. Considering that the function in Equation 1 measures the discrepancy between two covariance matrices, fitting the model to its own predicted matrix would show a perfect fit. Consequently, the discrepancy observed in fitting the replication data generated from the predicted matrix would be solely due to the sampling error.

As expected, the distributions of replication fit from the "perfect-fit" condition are generally better than the "observed" condition distribution. In the "perfect-fit" condition of Example 1, as shown in Table 1, the mean of χ^2 replicated was 49.71 with the SD of 10.15 when N = 188, but in the "observed" condition the mean of χ^2 was 104.47 with the SD of 16.84. In all other sample size conditions, the mean χ^2 were smaller in the "perfect-fit" condition than those of the "observed" condition. The same pattern was found in Example 2. The mean of χ^2 replicated was 192.19 in the "perfect-fit" condition, while it was 553.93 in the "observed" condition when N = 206.

An interesting trend is that the mean of χ^2 and its descendants (CFI, TLI, and RMSEA) did not change its magnitude as the sample size changed. Table 1 shows that the mean χ^2 of replicated ranged from 47.61 to 50.66, the mean of CFI from 0.99 to 1.00, the mean of TLI from 0.99 to 1.00, and the mean of RMSEA from 0.02 to 0.00. This consistency is especially unexpected for χ^2 as it is known to be sensitive to *N*. The result of Example 2 in Table 2 demonstrated a similar pattern. The χ^2 means ranged from 184.44 to 201.71 across varying sample sizes.

Figure 1 contained empirical cumulative distribution functions of two conditions superimposed on one another. The χ^2 distribution of the “perfect-fit” condition (the dotted curve) did not change as the sample size varied, while one of the “observed” condition (the solid curve) moved away from the “perfect-fit” distribution. The CFI, TLI, and RMSEA do not share this pattern. The χ^2 distributions in Example 2 were depicted in Figure 2, showing the same pattern as Example 1: the distribution of the “perfect-fit” did not move, but the distribution of the “observed” increased in value. The χ^2 family fit indices (CFI, TLI, and RMSEA) showing similar patterns in Example 2 to those in Example 1 were represented in Figure 2.

Another noteworthy characteristic of AIC and BIC is revealed in Figure 1. The distributions of AIC and BIC from both conditions largely overlapped. For example, the means and SDs of the “observed” condition and the “perfect-fit” condition in Example 1 in Table 1 are M_{AIC} (SD_{AIC}) = 15471.17 (67.74) and M_{AIC} (SD_{AIC}) = 15454.46 (68.58), respectively, when N = 188. The same pattern was observed in Example 2. In Figure 2, the distributions of AIC and BIC from the “observed” condition and the “perfect-fit” condition overlapped each other (e.g., M_{BIC} (SD_{BIC}) = 15941.81 (97.50) and M_{BIC} (SD_{BIC}) = 15923.32 (90.00), respectively, when N = 206, as shown in Table 2). This result implies that the log-likelihood-based fit indices like AIC and BIC are not only insensitive to the replication, but also insensitive to the bases of data generation; the observed or the predicted covariance matrix.

While the result of Example 3 is based on an impermissible solution, it provided useful insights. First, the worsening pattern of the χ^2 family fit indices in the “observed” condition and the insensitivity of χ^2 to the *N* in the “perfect-fit” condition (the χ^2 means ranged from 180.64 to

216.17, while N varied from 100 to 2,000, as shown in Table 3) were observed in here as well. Second, unlike the patterns observed in Example 1 and Example 2, the distributions of AIC and BIC from both conditions were not overlapped but were placed far apart. Moreover, the distribution of the "perfect-fit" condition showed worse fits (larger values). This seemingly inconsistent result can be understood considering the problem we encountered in fitting the original data. The predicted covariance matrix, the result of the original model fitting, was not a viable solution for the original data. Hence, the replication data generated from the problematic predicted covariance matrix showed a much worse fit than the replication from the observed covariance matrix.

Phase 2

The purpose of the second part of the current study is to examine the phenomenon observed in Phase 1 in controlled settings. The examples in Phase 1 were from empirical research. Although the parameters that might impact on the outcome of the study (such as the number of observed variables, the number of parameters to be estimated, the degrees of freedom, or the sample size) are known for the examples, using the real-world samples inhibits the flexibility of manipulating the conditions. In Phase 2, we built a simulation with varying conditions. For simplicity, we employed an intercorrelated factor model commonly used in confirmatory factor analysis (CFA). It can be considered a simplified SEM model where the measurement models are specified, but the structural model is just correlated factors. Figure 4 depicts a general model we used in Phase 2. A detailed model specification will be introduced in the simulation section.

[INSERT FIGURE 4 HERE.]

Specifically, we manipulated the number of observed variables *p*, the number of latent factors *f*, and the sample size *N*. The number of parameters and the degree of freedom is functions of *p* and *f* when the models are inter-correlated factor models, which we used. These factors are commonly used in the literature to examine the impact of model size on various SEM outcomes, such as fit indices and empirical rejection rates of the chi-square (Moshagen, 2012; Shi, Lee, & Terry, 2018). Additionally, we added a misspecified model condition to examine the fit indices when the model fit is not optimal.

Another purpose of Phase 2 simulation is to suggest a new index useful in replication attempts. The fit indices such as CFI, TLI, and RMSEA provide easy-to-use metrics and guidelines to evaluate a model according to its values. However, as observed in Phase 1, conventional fit indices based on χ^2 such as CFI, TLI, ECVI, and RMSEA deteriorate when one tries to replicate a study. Meanwhile, the fit indices based on the likelihood, such as AIC and BIC, are much less affected by replication, meaning that their expected values of a replicated sample are close to those of the original. Yet, their values are not easy to interpret when they are used alone. In fact, they estimate the relative amount of information lost by using a model. Their intended usage is to compare the model's AIC or BIC to each other and select the best model.

A constant pattern observed in Phase 1 is that the fit indices based on χ^2 worsen on replication attempts while the fit indices based on likelihood do not. Recall that χ^2 is the likelihood ratio of the saturated (unrestricted) and the hypothesized model. From this, we speculate that the likelihood of a saturated model affects the worsening effects of the χ^2-based fit indices. To avoid such an effect, we suggest an alternative likelihood ratio. Instead of the likelihood ratio of the saturated and the hypothesized model, the likelihood ratio of the baseline model and the hypothesized model is suggested.

A theorem offered by Wilks (Wilks, 1938) shows that twice the likelihood ratio will asymptotically follow χ^2 distribution with degrees of freedom *df* equal to the difference of *df*s of two likelihoods as the sample size *N* approaches infinity. In theory, we can formulate the likelihood ratio of the baseline and the hypothesis model and expect the ratio to be distributed as χ^2 distribution. The new ratio does not involve the unrestricted likelihood so that it will be less impacted by replication attempts. Therefore, the baseline likelihood ratio is defined as

$$LR_0 = -2\left[\ln L\left(\boldsymbol{\theta}_0\right) - \ln L\left(\widehat{\boldsymbol{\theta}}\right)\right] = 2\left[\ln L\left(\widehat{\boldsymbol{\theta}}\right) - \ln L\left(\boldsymbol{\theta}_0\right)\right]$$

where $\ln L(\boldsymbol{\theta}_0)$ is the log-likelihood of the baseline model, and $\ln L(\widehat{\boldsymbol{\theta}})$ is the log-likelihood of the hypothesized model.

Using a baseline model to formulate a fit index is not unusual. The CFI and TLI, introduced in the previous section, utilize the idea of the baseline model. To recap, the TLI compares the baseline chi-square χ_0^2 and its *df* ratio to the hypothesized model chi-square χ^2 and its *df*. The CFI compares the non-centrality of the baseline and the hypothesized model. As another example that is more directly related to *LR*0, Bentler & Bonett (1980) suggested the normed fit index (NFI) that is defined as

$$NFI = \frac{\chi_0^2 - \chi^2}{\chi_0^2}, \qquad \text{(Eq 4)}$$

where χ_0^2 is the likelihood ratio of the baseline and the saturated model,

$$\chi_0^2 = -2\left[\ln L\left(\boldsymbol{\theta}_0\right) - \ln L(\boldsymbol{S})\right] = 2\left[\ln L(\boldsymbol{S}) - \ln L\left(\boldsymbol{\theta}_0\right)\right],$$

and χ^2 is the likelihood ratio of the hypothesized and the saturated model,

$$\chi^2 = -2\left[\ln L\left(\widehat{\boldsymbol{\theta}}\right) - \ln L(\boldsymbol{S})\right] = 2\left[\ln L(\boldsymbol{S}) - \ln L\left(\widehat{\boldsymbol{\theta}}\right)\right].$$

Expanding Equation 4 in terms of log-likelihoods, we get

$$
\begin{aligned}
\frac{\chi_0^2-\chi^2}{\chi_0^2} &= \frac{2\left[\ln L(\boldsymbol{S})-\ln L\left(\boldsymbol{\theta}_0\right)\right]-2\left[\ln L(\boldsymbol{S})-\ln L\left(\widehat{\boldsymbol{\theta}}\right)\right]}{2\left[\ln L(\boldsymbol{S})-\ln L\left(\boldsymbol{\theta}_0\right)\right]} \\
&= \frac{2\ln L(\boldsymbol{S})-2\ln L\left(\boldsymbol{\theta}_0\right)-2\ln L(\boldsymbol{S})+2\ln L\left(\widehat{\boldsymbol{\theta}}\right)}{2\left[\ln L(\boldsymbol{S})-\ln L\left(\boldsymbol{\theta}_0\right)\right]} \\
&= \frac{-2\ln L\left(\boldsymbol{\theta}_0\right)+2\ln L\left(\widehat{\boldsymbol{\theta}}\right)}{2\left[\ln L(\boldsymbol{S})-\ln L\left(\boldsymbol{\theta}_0\right)\right]} \\
&= \frac{2\left[\ln L\left(\widehat{\boldsymbol{\theta}}\right)-\ln L\left(\boldsymbol{\theta}_0\right)\right]}{2\left[\ln L(\boldsymbol{S})-\ln L\left(\boldsymbol{\theta}_0\right)\right]} \\
&= \frac{LR_0}{\chi_0^2}
\end{aligned}
$$

Per our speculation that the saturated model's log-likelihood is the reason for the deteriorating fit, the NFI would also show the worsened fits on the replicated samples since it has the log-likelihood of the saturated model as its components. It is also worth noting that the log-likelihood of the saturated model cancels out on the numerator of NFI, which is the difference between the baseline chi-square and the hypothesized model chi-square. We speculate that NFI would suffer from the replication while LR_0 would resist the replication effect on the chi-square-based fits.

To summarize, the goal of Phase 2 is to investigate the generalizability of the diminishing phenomenon of fit indices observed in Phase 1 in terms of several factors, such as the number of observed variables and the model complexity, and to formulate a new index that can be used in replication attempts.

Simulation

The simulation strategy was identical to Phase 1 in the broad sense that it also mimicked a replication attempt. We generated many simulated replication samples from the observed covariance matrix to examine the distribution of fit indices. The main difference in Phase 2 is that we also generated the (original) observed covariance matrix to gain more control of the model specifications. Specifically, we imposed a population model to generate data that serves as the original. Then, we used the original to generate many datasets as the replication samples. We justify the reasons for adopting the two-step data generation scheme as followings: 1) it granted the control of the population, and 2) it imitated a replication attempt in a typical scenario that a researcher does not have information about the population and relies on the observed data.

Four levels of the number of observed variables (p = 15, 30, 45, and 60), two levels of the number of latent factors (f = 3 and p/3), four levels of the sample size (N = 200, 400, 1000, and 2000), and two conditions of the model specification had crossed each other to build 64 different models. A commonly used confirmatory factor analysis (CFA) with a congeneric measurement model where each observed variable is loaded to a single corresponding latent factor was used to generate data for each condition. A generalized depiction of the model specification is represented in Figure 4. For example, in a condition where p = 30 and f = p/3, each of the ten factors is indicated by three observed variables while those factors are correlated with each other.

Data Generation

The first step of the two-step procedure was generating original data. This step was done by imposing population relationships onto a generated dataset. The intercorrelated factor model is specified as follows:

$$\boldsymbol{X} = \boldsymbol{\Lambda\eta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{X}$ is a vector of observed variables, $\boldsymbol{\Lambda}$ is a factor loading matrix, $\boldsymbol{\eta}$ are the latent variables, and $\boldsymbol{\varepsilon}$ is a residual vector. In this notation, two random components cause the value of X; the value of the latent variable and the unexplained residual. Note that only one latent factor is loaded to an observed variable. We assumed both are normally distributed, independent of each other. Thus, we generated a sample size of *N* that contains f normally distributed factor scores and *p* normally distributed residuals. For simplicity, we set all factor loadings to 0.7 and the residual variances to 0.51. Additionally, we set all latent factor variances to 1, and all factor correlations are set to 0.3. We obtained a dataset of observed variables by applying the relationship to the generated factor scores and residuals.

Although we took the strategy that generates unobserved variables and imposes a population relationship on them, note that this is equivalent to generating observed random variables that follow multivariate normal distribution directly from a population moment matrix corresponding to a model specification. It is because a sum of normally distributed random variables is also normally distributed. In our configuration of the factor loadings set to 0.7 and the factor correlations of 0.3, a correlation between the observed variables that are loaded on the same factor is the factor loading λ times the variance of latent variable times λ again, 0.7 × 1 × 0.7 = 0.49. A correlation between observed variables loaded on different factors is λ times the correlation between those factors times λ, 0.7 × 0.3 × 0.7 = 0.147. Since we simplified the factor loadings and the factor correlations, the value of each non-diagonal element of the population correlation matrix is either 0.49 when the indicators are on the same factor or 0.147 when the indicators are on different factors. Again, our strategy of generating observed variables from latent variables is equivalent to a strategy that constructs the population moment matrix first and generates random variables directly from it.

The first step was to obtain the “original” data to be replicated. The second step was to generate many “replicated” samples from the original data. Similar to what has been done in Phase 1, we calculated the observed covariance matrix of the original and generated 1,000 sets of multivariate normal random variables per condition that follows it.

Model Fitting

Aside from the model complexity conditions (e.g., *p* and *f*), we also created two conditions for how a model is specified. In a well-specified condition, we used the same model that generates the dataset, except the model parameters (e.g., factor loadings, residual variances, and factor correlations) are freely estimated. In contrast, we also created a misspecified condition where the fitted model is not identical to the population model. In the misspecified condition, a single path to an observed variable is moved from one latent variable to another so that that observed variable is loaded to a different factor than what it supposed to be loaded. For example, in the well-specified condition where $p = 15$ and $f = 3$, the first five observed variables (X_{11} to X_{15}) are loaded to the first factor (F_1), the next five (X_{21} to X_{25}) are loaded to the second factor (F_2) and so on, where Xij represents the *i*th indicator of *j*th latent factor. In the misspecified condition, the X_{12} to X_{15} are loaded to F_1 while X_{11}, X_{21}, X_{22}, X_{23}, X_{24}, and X_{25} are loaded to F_2. Figure 5 contains a generalized diagram of the model that is fitted in the misspecified condition. Only one observed variable was moved to the other factor to create a misalignment in the data regardless of other conditions. That is, no matter how complex the model is, only one indicator-factor misalignment occurred.

[INSERT FIGURE 5 HERE.]

Since we generated all models in the same manner (a congeneric correlated factor model), the number of parameters estimated *q* and the degree of freedom *df* are functions of *p* and *f*. The number of unique information in the dispersion matrix is $p(p+1)/2 = 120$ when $p = 15$. When $f = 3$, *q* is equal to 33 and consists of 15 factor loadings, 15 residual variances, and three correlations between factors, whereas when $f = 5$, *q* is equal to 40, consists of the same numbers of factor loadings and the residual variances but now ten correlations between factors. The *df*s are equal to 87 and 80, respectively.

Once the model is specified according to the condition, it is fitted to the replicated sample. The remaining steps for the simulation are similar to the simulation in Phase 1. After the model fitting is completed and converged, the fit indices are recorded, and the process of generating the replication sample and the model fitting is repeated. Another main difference between Phase 1 and Phase 2 is that Phase 2 does not have the "perfect-fit" condition. The scheme for data generation is already known in Phase 2, and the purpose of the "perfect-fit" condition in Phase 1, which was to give illustrations of the "upper-bound" of replication results, can be achieved in the well-specified condition in Phase 2. In other words, the well-specified condition fits a "true" model to data, and theoretically, the imperfection of fit is only due to the sampling error.

The simulations in Phase 1 and Phase 2 have another meaningful difference. In Phase 1, we used examples from the literature. Therefore, we do not know what the true model that generates data in the population is. On the contrary, we know the true model that fits perfectly to the population in Phase 2. This difference makes the well-specified condition in Phase 2 unique compared to any conditions in the Phase 1 simulation. The difference between the well-specified condition and the "perfect-fit" condition is that the well-specified condition uses the two-step data generation so that it does not sample from the population but from the observed covariance matrix, like researchers who do not know the population would do.

Additionally, the well-specified condition differs from the "observed" condition in Phase 1 in that we know the former fits the true model to the data, while we do not know the true model in the latter. It is highly likely that the model we fitted in the "observed" condition in Phase 1 is not the true model. Thus, the “observed” condition in Phase 1 is comparable to the misspecified condition in Phase 2 because they both fit a misspecified model to the data.

Results

The baseline likelihood ratio LR_0 and the standard likelihood ratio LR were collected across the simulated originals and replications. All fit indices included in Phase 1 can be derived from either the log-likelihood value of the hypothesized model or the χ^2, which is equal to LR. Therefore, we included only these two as our results. Note that we will use χ^2 as the notation of the standard likelihood ratio instead of LR to maintain consistency with the notation we introduced and used in Phase 1. Also, note that LR_0 is an index of goodness-of-fit. A small value indicates that the discrepancy between the baseline and hypothesized models is small. This is different from the interpretation of χ^2. The χ^2 is an index of badness-of-fit, which means the larger value indicates the hypothesized model is far from the “true” model. In turn, we consider the replication attempt with a larger value of LR_0 than the original has a satisfactory fit, while the replication with a smaller value of χ^2 than the original has a satisfactory fit.

As expected, LR_0 is not heavily affected by replication, while χ^2 has a 0% of satisfactory fit in most conditions, with a few exceptions in the misspecified condition. In the well-specified condition, none of the model complexity parameters (p and f, and consequently q and df) impacted the percentage of satisfactory χ^2; in all conditions, the rate was 0.00%. In other words, the χ^2 always showed a worse fit than the original when the model was well-specified. Figure 6 contains empirical cumulative distribution functions of χ^2 in the well-specified condition. The curves representing the replicated distribution are placed on the right side of the dots, representing the value from the original covariance matrix. On the other hand, the LR_0 exhibited a different property in replicated samples. More than half of the replication attempts in the well-specified condition had better LR_0 across all other conditions, ranging from 50.90% to 92.59%, as shown in Table 4 and Figure 7. Considering the sampling error, the ideal would be a satisfactory rate of 50%. While the LR_0 was consistently overestimating the difference between the baseline and the hypothesized log-likelihood, it has a much better property than the χ^2 in terms of diminishing the fit in replication. Another pattern that can be found in Figure 7, consistent with the observation made in Phase 1, was that the overestimation tended to decrease with increasing sample size. For instance, the proportions of satisfactory LR_0 gradually decreased (57.70%, 52.90%, 51.90%, and 50.90%) as we increased the sample size (N = 200, 400, 1,000, and 2,000) when p = 60 and f = 3, as shown in Table 4. Lastly, the overestimation of LR_0 was larger when the number of parameters estimated q was greater. We indirectly manipulated q by imposing different numbers of latent factors f. For example, q = 63 when p = 30 and f = 3, while q = 105 when p = 30 and f = 10. For these conditions, the satisfactory LR_0 rates were 51.20% and 58.40%, respectively, when N = 2,000. Parallel patterns can be observed across other conditions. A complete simulation result for the well-specified condition can be found in Table 4.

[INSERT TABLE 4 HERE.]

[INSERT FIGURE 6 HERE.]

[INSERT FIGURE 7 HERE.]

The pattern of non-replication of χ^2 repeated in the misspecified condition, with a few exceptions. As we can see in Table 5, only in the conditions where p = 15 the satisfactory χ^2 rate showed a slightly increasing pattern with increasing sample sizes (the percentages of χ^2 smaller than the original were 0.10%, 0.00%, 2.10%, and 5.80% when f = 3 and N = 200, 400, 1,000, and 2,000, respectively, while the percentages were 0.00%, 0.90%, 3.20%, and 6.60%, respectively, when f = 5.) The pattern was not observed in any other conditions on p, as shown in Figure 8. On the contrary, the *LR_0* showed a similar pattern observed in the well-specified condition. It tended to overestimate the *LR_0* observed in the original sample, but the overestimation decreased with increasing sample size, as shown in Figure 9. For example, the satisfactory *LR_0* rate decreased (58.10%, 56.60%, 55.00%, and 51.90%) when the sample size increased (N = 200, 400, 1,000, and 2,000, respectively) when p = 15 and f = 3. The effect of the number of parameters on the satisfactory *LR_0* rate was also observed in the misspecified condition. For instance, the satisfactory *LR_0* rates were 56.70%, 54.50%, 56.10%, and 52.00% when p = 45, f = 3, q = 93, and N increased from 200 to 2,000, respectively, while the rates were 84.40%, 77.80%, 67.20%, and 61.30%, respectively, when p = 45, f = 15, and q = 195. A detailed result of the misspecified conditions is contained in Table 5.

[INSERT TABLE 5 HERE.]

[INSERT FIGURE 8 HERE.]

[INSERT FIGURE 9 HERE.]

One of the main findings in Phase 2 was a confirmation of what had been observed in Phase 1. We have observed that the fit indices with χ^2 in their formulation deteriorated when one tries to replicate an SEM from the observed moment matrix, while the fit indices stemmed from the log-likelihood value did not. Based on this observation, we also have speculated that the likelihood of the saturated model was causing the deterioration since the χ^2 is the ratio of two likelihood values, one for the hypothesized model and the other for the saturated model. Another main finding of Phase 2 was that the likelihood ratio that involves the unrestricted likelihood showed a weaker fit in the context of replication of a study, while the *LR_0*, which is the ratio of the likelihood of the baseline model and the likelihood of the hypothesized model and does not involve the likelihood of the saturated model, was less affected by replication.

It is worth noting that replicated LR_0 overestimates its original value, especially when p and q are large and N is small. While the LR_0 is newly introduced and its distributional properties have not been scrutinized, some helpful research findings are available to explain the such deviation. In their investigation of the model size effect of SEM, Shi et al. (2018) found that the standard likelihood ratio test statistic does not follow the asymptotic chi-square distribution well when the number of observed variables is large ($p \geq 60$). Moreover, Jackson (2003) also found an association between the number of parameters estimated and the poor approximation of the likelihood ratio to the chi-square distribution. From these findings, we can speculate that LR_0 also suffered from the small sample size and the large model complexity.

An important advantage of LR_0 over AIC and BIC is that it performed better than the likelihood-based fit indices in distinguishing a better model-data fit from a worse one. A comparison of Figure 10 and Figure 11 revealed that the distributions of BIC are identical whether the model is well-specified or misspecified. In other words, BIC showed limited usefulness in distinguishing a better fit from a worse one. On the contrary, a comparison of Figure 7 and Figure 9 showed a general tendency of larger LR_0 on the well-specified condition than the misspecified condition, with a few exceptions. Considering that the Phase 2 simulation results depended on a single random draw from the population that serves as the original dataset, this potential advantage over BIC is worth investigating in future research.

[INSERT FIGURE 10 HERE.]

[INSERT FIGURE 11 HERE.]

Finally, we speculated that Bentler & Bonett's (1980) NFI would show the worsening effect caused by replication attempts despite its apparent similarity with LR_0. This speculation has been supported by the simulation result done in Phase 2. In short, the worsening pattern is parallel to the pattern of χ^2 shown in Table 4 and Table 5. Almost no conditions have shown a better NFI than the original in the well-specified condition, while small percentages of replications with better NFI were observed in the misspecified condition. A complete result regarding the NFI is contained in Table B1 in Appendix B.

Discussion

The pattern we observed in both phases is unexpected because there should be no difference in the fit indices of the original and the replicates on average. Fit indices are intended to indicate how well (or poorly) the model fits the data. In the current context, the model is fixed, while the data is stochastic, following a population distribution. All samples should be exchangeable, including the original. Considering the sampling error, some of the replicated data should show a better fit than the original while others worse. Note that we are not constraining the estimates to the original; all parameters are freely estimated, corresponding to the data in each replication. It also means that the discrepancy between $\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)$ and $\boldsymbol{S}$ is minimized in each replication and is reflected in the fit indices.

The difference between the theoretical χ^2 distribution and the simulated χ^2 distribution is worth noting. Joreskog (1969) indicated that *N* times the minimized discrepancy follows the χ^2 distribution. In other words, the theoretical χ^2 distribution can be obtained once the parameter estimates fix the minimum value of the discrepancy function. The sampling variability, then, makes the theoretical χ^2 distribution. On the other hand, the simulated distribution in the current study differs in that the value of the discrepancy function and the sample both vary. Although the model specification is identical throughout the simulation, each replication has its own set of parameter estimates and, therefore, $\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)$ differs from others. This critical difference makes the theoretical χ^2 distribution unusable to make any inference about the current result. The simulated distribution does not follow the χ^2 distribution, and it is difficult to determine what kind of distribution it follows.

The result of the simulation in Phase 2 showed that the standard likelihood ratio that compares the unrestricted and the restricted (hypothesized) log-likelihood does not replicate. Especially the replication of the unrestricted log-likelihood was problematic. Recall that Equation 2 is reduced to Equation 3 when the model-implied moment matrix equals the sample moment matrix. In turn, the unrestricted log-likelihood in Equation 3 is independent of parameter estimates. The source of variability in Equation 3 is the determinant of the sample moment matrix, given that the dimension and the sample size are fixed. We can conclude that the determinant of the sample moment matrix diminishes when we generate multivariate normal data from other sources. One could speculate that generating random numbers based on simplified assumptions (such as assuming the sample follows the multivariate normal distribution) may lose some information in the original sample or even the population. In fact, the differential entropy (which is a measure of information) for a multivariate normal distribution is defined as

$$H(\boldsymbol{\Sigma}) = \frac{p}{2} + \frac{1}{2}p\ln(2\pi) + \frac{1}{2}\ln|\boldsymbol{\Sigma}|,$$

which can be obtained by dividing Equation 3 by negative of the sample size. We dismiss this speculation for the results in Phase 2. Unlike the simulation in Phase 1, where we do not know how the observed variables were distributed, the simulation in Phase 2 generated the original data that follows a multivariate normal distribution. No unknown causes or relationships affected the generation of observed variables. However, the diminishing effect of χ^2 was still observable in the Phase 2 simulation. Thus, we rule out the violation of the multivariate normal assumption as a possible reason for the fit-worsening effect.

Despite of unclarified reason for the phenomena, the results showed a clear pattern of worsening in specific fit indices. The chi-square and its descendants (CFI, TLI, and RMSEA) on the replicated sample showed worse fits than the original most of the time, particularly when the sample size was not large. This finding invites caution in interpreting the result of any replication attempts using these SEM indices. The current observation warns that if a study tries to replicate previous findings on SEM, the fit indices such as CFI, TLI, RMSEA, and ECVI would likely fail to replicate the original fit index. Thus, it requires further investigation before dismissing the attempt due to its (possibly) inadequate fit indices.

Moreover, the risk of rejecting the replication is amplified with a conjunction of a common SEM practice that puts clear "cut-off" points for the indices (see West, Taylor, & Wu, 2012). In some cases, a researcher may face a situation where the original research was acceptable in terms of fit, but the replication was not.

Another suggestion the current finding could make is to use the chi-square family indices in conjunction with the likelihood family (AIC and BIC). They are less influenced by replication, so they can provide additional information to the widely used chi-square family. For example, if the CFI of replication is worse than the original, but the AIC is close to the original, it may indicate a successful replication that the CFI failed to capture.

Nevertheless, a cautionary note should be made about using relative fits such as AIC and BIC. As stated above, the relative fits compare model-data fit among different models to select a model that loses a minimum amount of information. However, the performance of AIC and BIC in selecting a model has been questioned in recent research (Sen & Bradshaw, 2017). Specifically, the accuracy of the relative fits in selecting the true model is concerningly low, especially when the measurement properties of indicators are not optimal. The context of usage of AIC and BIC, however, is different in the current study. Relative fits are conventionally used to select a model among many, but here it is used to compare the same model with a different dataset.

We also formulated an index to determine whether the replicated sample fits the model. Although the newly suggested baseline likelihood ratio LR_0 showed the property of resistance to the fit-worsening effect of replication attempt, it has a clear limitation to be used in assessing the fit between the model and the data. Like the standard likelihood ratio test, it follows χ^2 distribution when the sample size approaches infinity. It is not very useful by itself as the null that the baseline and the hypothesized model are the same would almost always be rejected. Moreover, it is difficult to determine the distribution the test statistic follows under the alternative. The attempts to formulate fit indices that resemble the formula of CFI, TLI, and RMSEA were unsuccessful in discerning well- and misspecified models from each other. This failure is partially due to the lack of criteria that indicate model-data fit in an absolute sense (i.e., a CFI value over .95 indicates a good fit) that is not influenced by the replication. In such circumstances where one should doubt the credibility of most indices, any evaluation of newly suggested goodness- or badness-of-fit would be a self-reference. We suggest a further examination of the LR_0 in terms of the characteristics of its distribution by empirical methods as well as mathematical derivations for future research.

On a final note, the current study sheds light on a gap in the research area of replication of SEM. The finding that many fit indices show worse fit in replicated samples offers helpful information to whoever tries to reproduce a result from SEM. Furthermore, researchers who use covariance structure analysis with generating multivariate normal random variables would be directly affected by the current finding. In such a case, one should be aware that the observed χ^2

value may not represent the value that could be observed in the result from the original moment matrix.

Reference

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike* (pp. 199-213). Springer, New York, NY.

Allen, C., & Mehler, D. M. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS biology*, *17*(5), e3000246.

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*(3), 305-324.

Babin, B. J., & Svensson, G. (2012). Structural equation modeling in social science research: Issues of validity and reliability in the research process. *European Business Review*, *24*(4), 320-330

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin*, *88*(3), 588.

Browne, M. W. (2000). Cross-validation methods. *Journal of mathematical psychology*, *44*(1), 108-132.

Everett, J. A., & Earp, B. D. (2015). A tragedy of the (academic) commons: Interpreting the replication crisis in psychology as a social dilemma for early-career researchers. *Frontiers in psychology*, *6*, 1152.

Goodboy, A. K., & Kline, R. B. (2017). Statistical and practical concerns with published communication research featuring structural equation modeling. *Communication Research Reports*, *34*(1), 68-77.

Guido, G., Marcati, A., & Peluso, A. M. (2011). Nature and antecedents of a marketing approach according to Italian SME entrepreneurs: A structural equation modeling approach. *International Journal of Entrepreneurial Behavior & Research*.

Hermida, R. (2015). The problem of allowing correlated errors in structural equation modeling: concerns and considerations. *Computational Methods in Social Sciences*, *3*(1), 5.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural equation modeling: a multidisciplinary journal*, *6*(1), 1-55.

Huth-Bocks, A. C., Levendosky, A. A., Bogat, G. A., & Von Eye, A. (2004). The impact of maternal characteristics and contextual variables on infant–mother attachment. *Child development*, *75*(2), 480-496.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124.

Jackson, D. L. (2003). Revisiting Sample Size and Number of Parameter Estimates: Some Support for the N:q Hypothesis. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(1), 128–141.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183-202.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr, R. B., Bahník, Š., Bernstein, M. J., ... & Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social psychology*, *45*(3), 142.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. *Frontiers in Psychology, 9*, 1117.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584-585.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235-245.

Moshagen, M. (2012). The model size effect in SEM: Inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(1), 86-98.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Passolunghi, M. C., Vercelloni, B., & Schadee, H. (2007). The precursors of mathematics learning: Working memory, phonological ability and numerical competence. *Cognitive development*, *22*(2), 165-184.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of statistical software*, *48*, 1-36.

Rosseel, Y. (2021). Evaluating the observed log-likelihood function in two-level structural equation modeling with missing data: From formulas to R code. *Psych*, *3*(2), 197-232.

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature, 515*(7525), 9-9.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461-464.

Sen, S., & Bradshaw, L. (2017). Comparison of relative fit indices for diagnostic model selection. *Applied psychological measurement*, *41*(6), 422-438.

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 21-40.

Simons, D. J., & Holcombe, A. O. (2014). Registered replication reports. *APS Observer*, *27*.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological science*, *24*(10), 1875-1888.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate behavioral research*, *25*(2), 173-180.

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

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*(1), 1-10.

Ullman, J. B., & Bentler, P. M. (2012). Structural equation modeling. *Handbook of Psychology, Second Edition*, *2*.

Verhagen, J., & Wagenmakers, E. J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*(4), 1457.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, *70*(2), 129-133.

West, S. G., Taylor, A. B., & Wu, W. (2012). Model fit and model selection in structural equation modeling. *Handbook of structural equation modeling*, *1*, 209-231.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological methods*, *8*(1), 16.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, *9*(1), 60-62.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*.

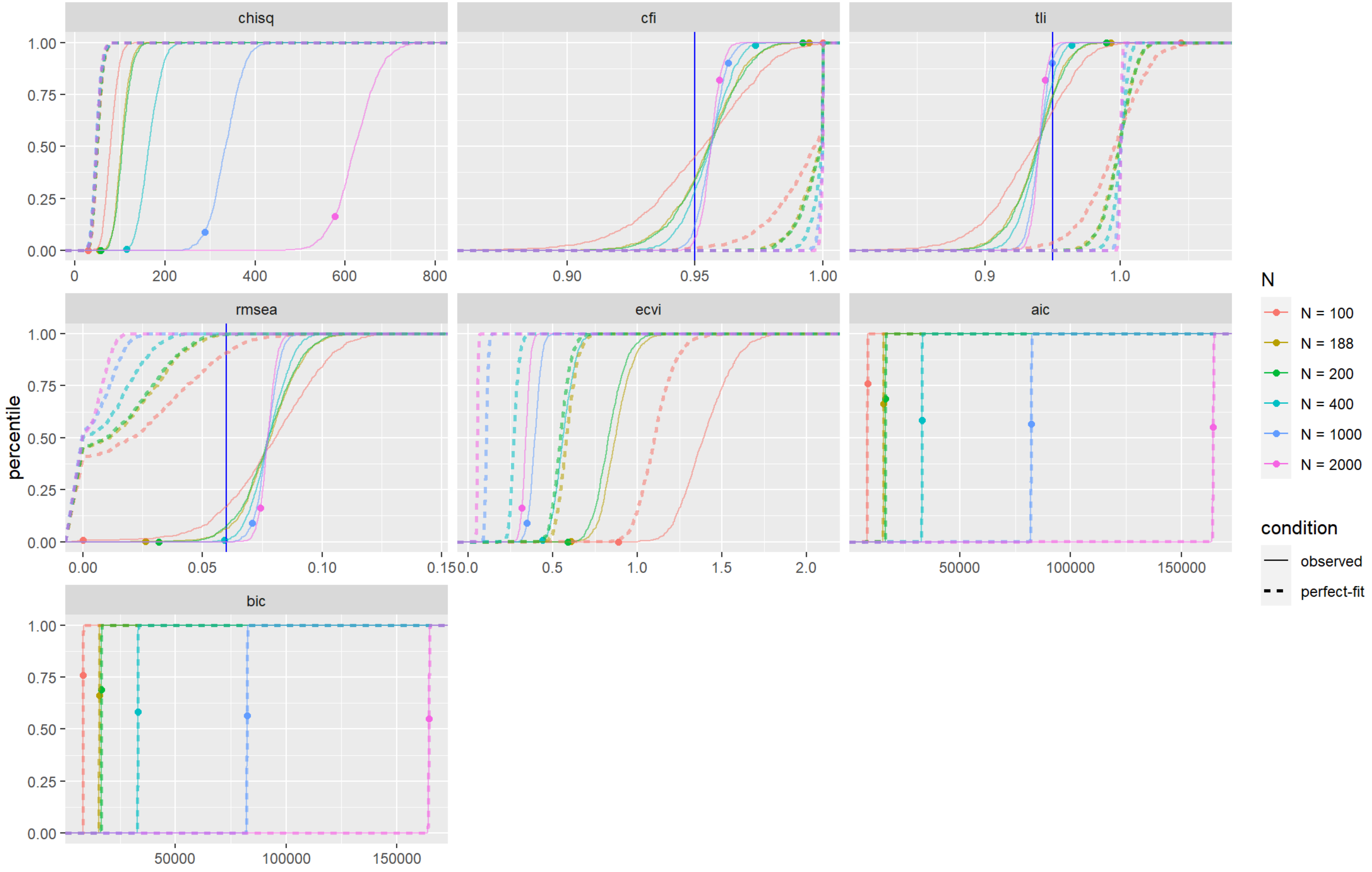

Figure 1. Empirical cumulative distribution functions of fit indices on Example 1. The dots represent the value from the original moment matrix. The blue vertical lines on CFI, TLI, and RMSEA mark the recommendation values by West, Taylor, and Wu (2012).

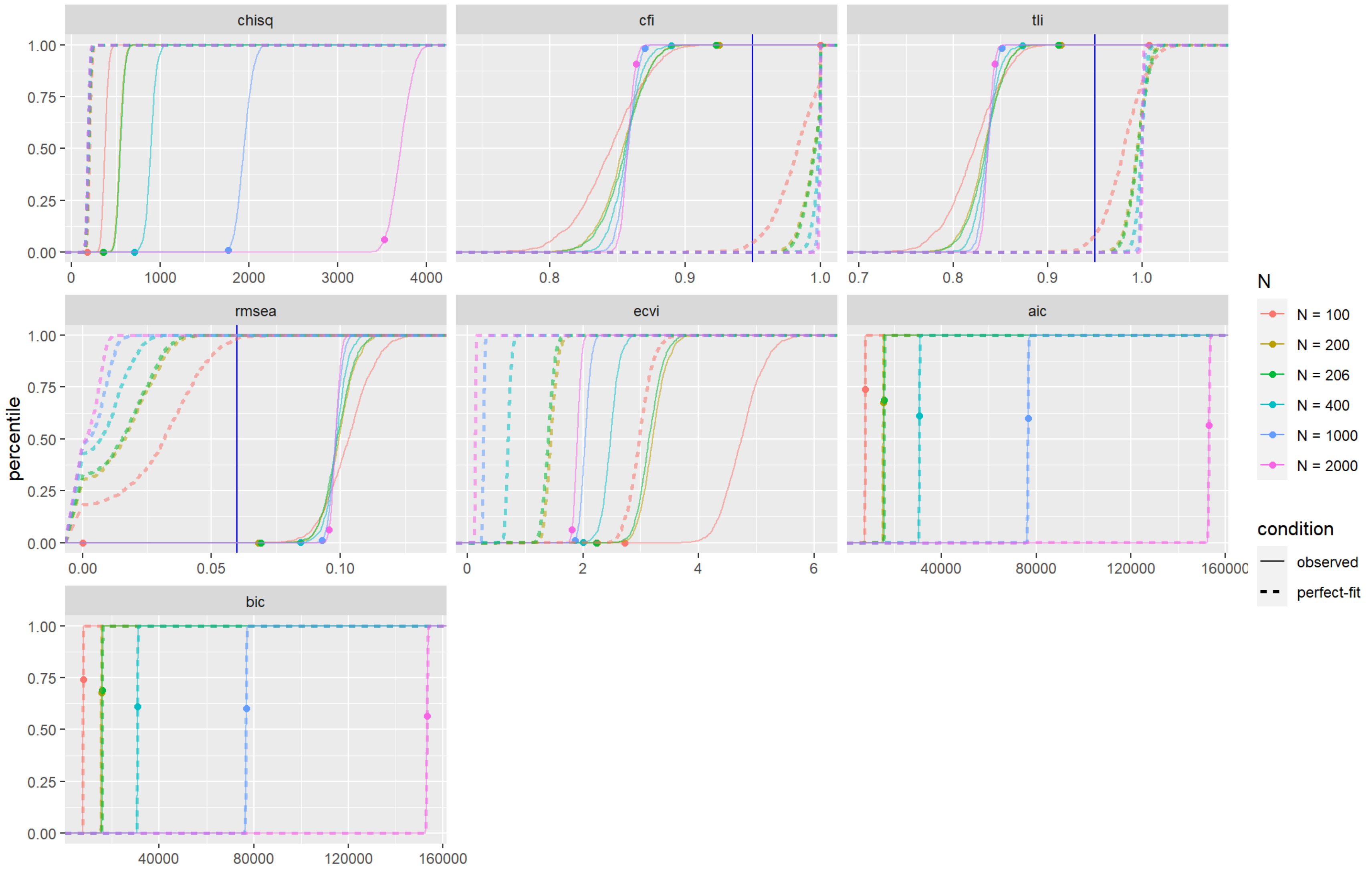

Figure 2. Empirical cumulative distribution functions of fit indices on Example 2. The dots represent the value from the original moment matrix. The blue vertical lines on CFI, TLI, and RMSEA mark the recommendation values by West, Taylor, and Wu (2012).

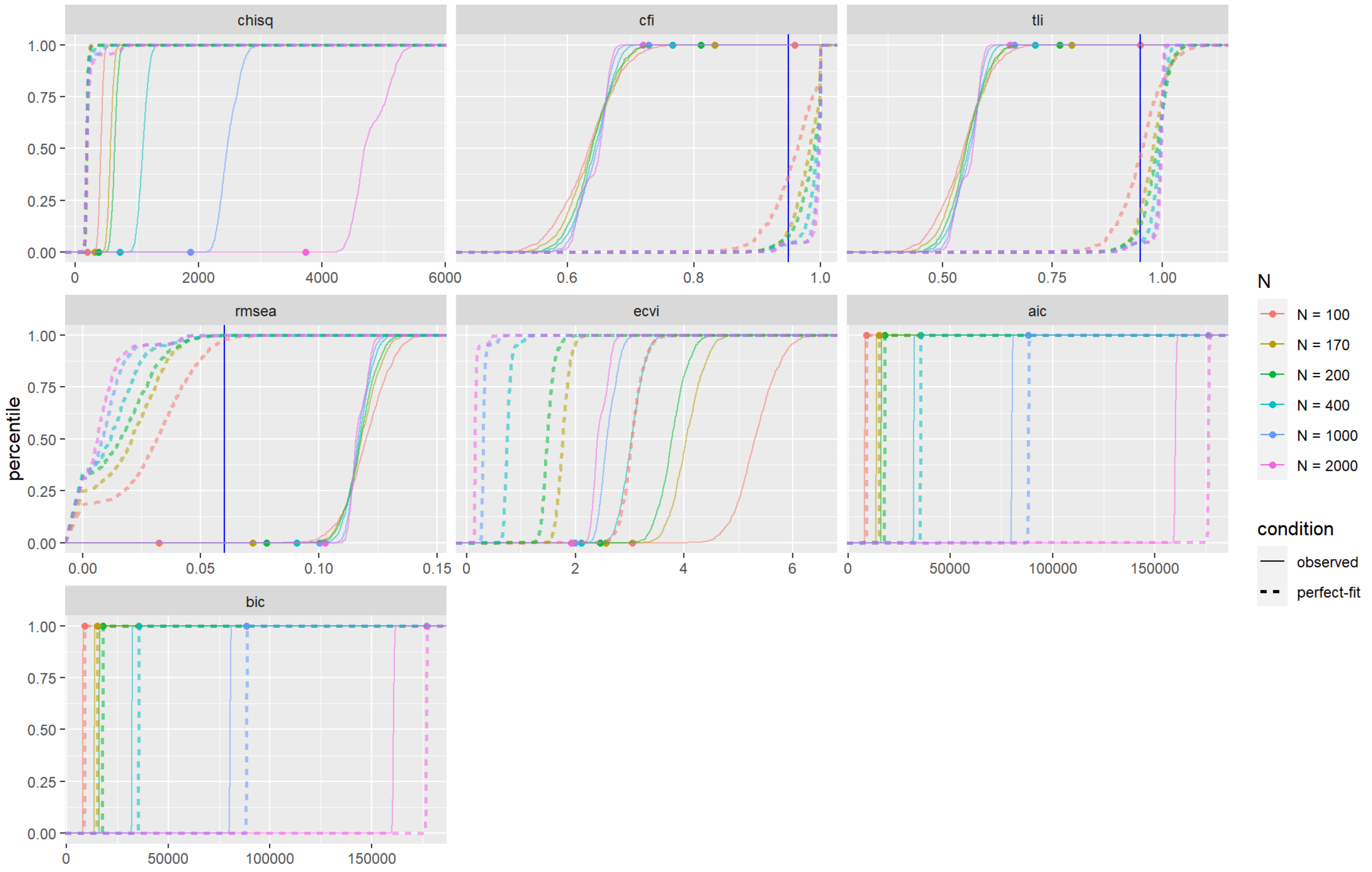

Figure 3. Empirical cumulative distribution functions of fit indices on Example 3. The dots represent the value from the original moment matrix. The blue vertical lines on CFI, TLI, and RMSEA mark the recommendation values by West, Taylor, and Wu (2012).

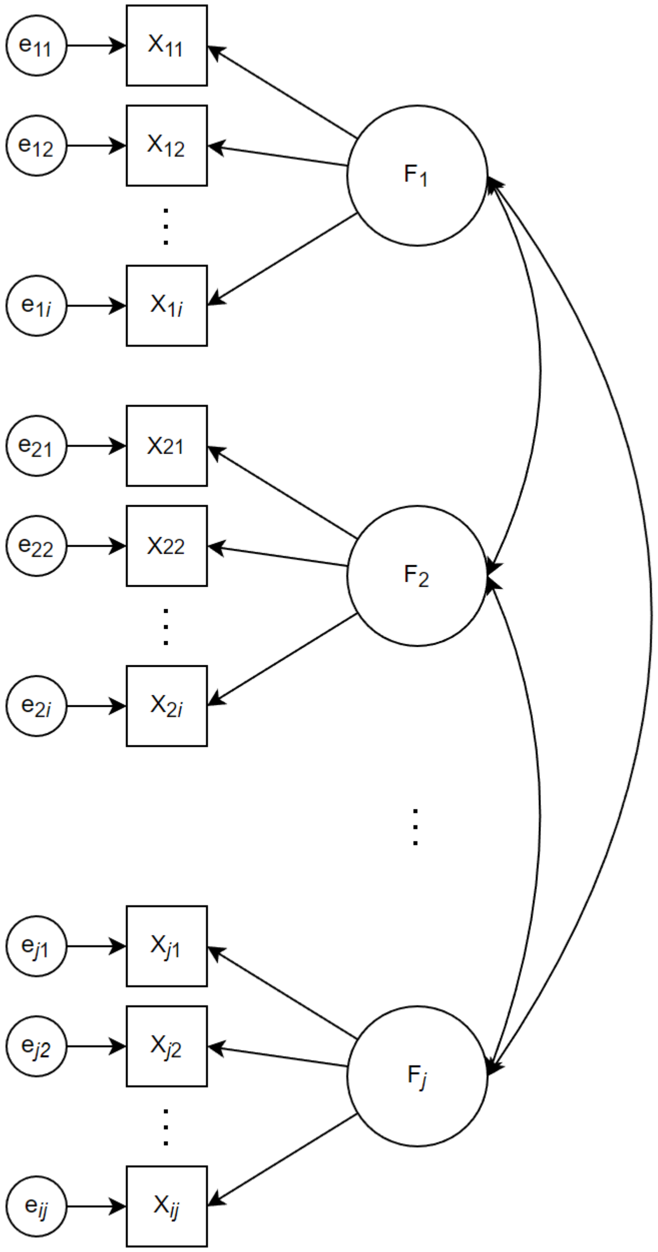

Figure 4. An intercorrelated factor model with *i* times *j* observed variables and *j* latent factors.

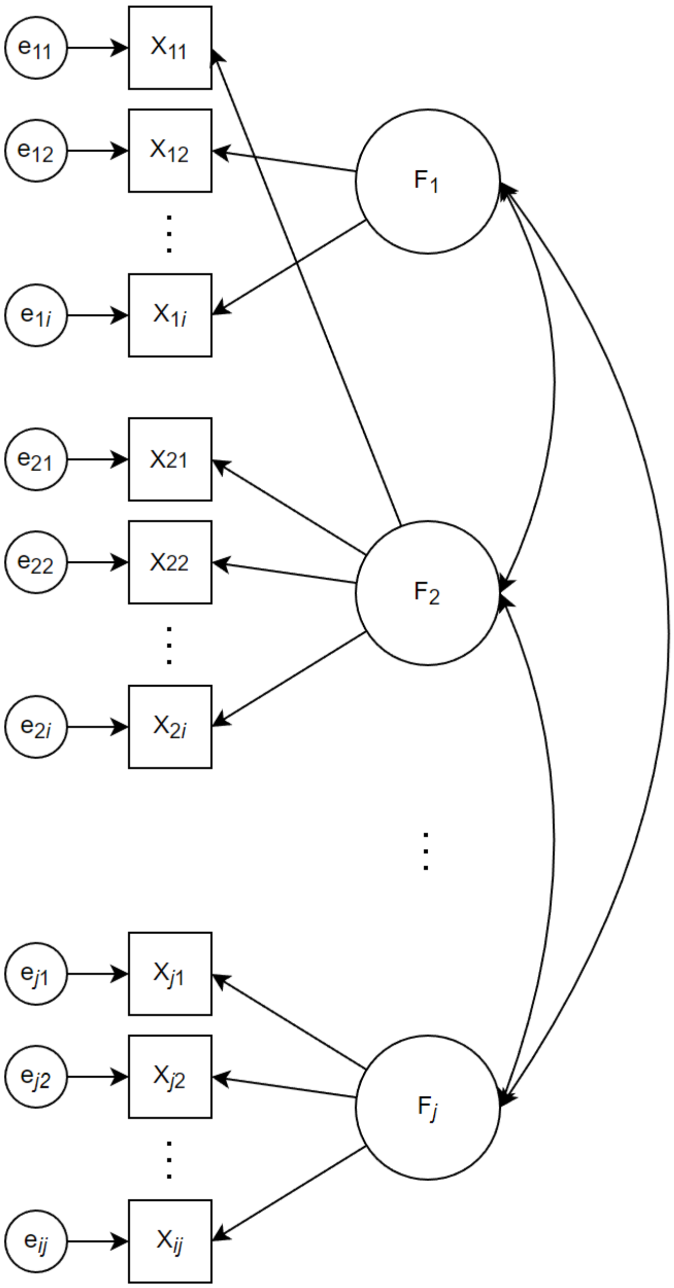

Figure 5. A diagram for the misspecified condition in Phase 2 simulation.

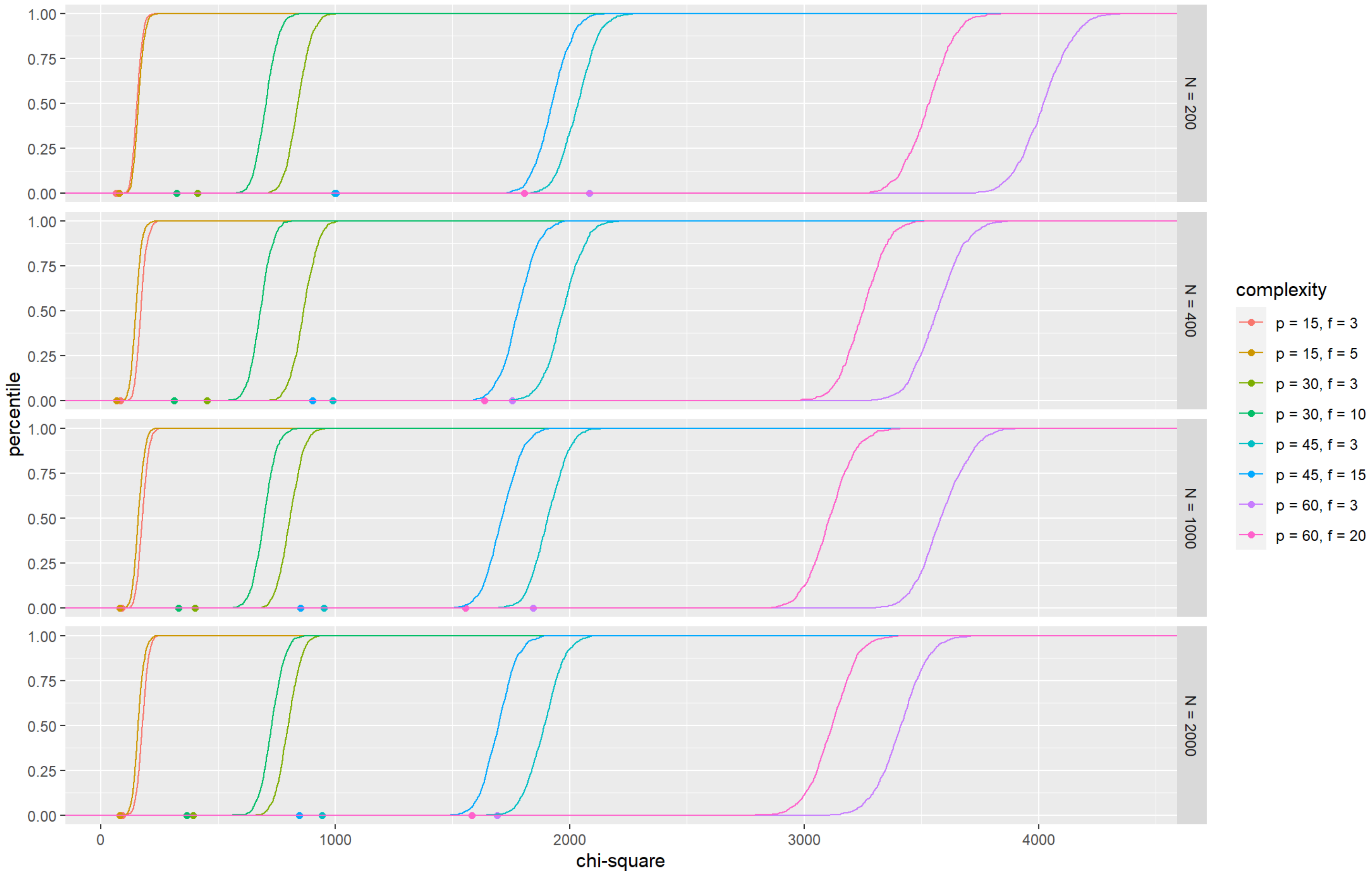

Figure 6. Empirical cumulative distribution functions of chi-square in the well-specified condition in Phase 2 simulation. The dots represent the value from the original covariance matrix.

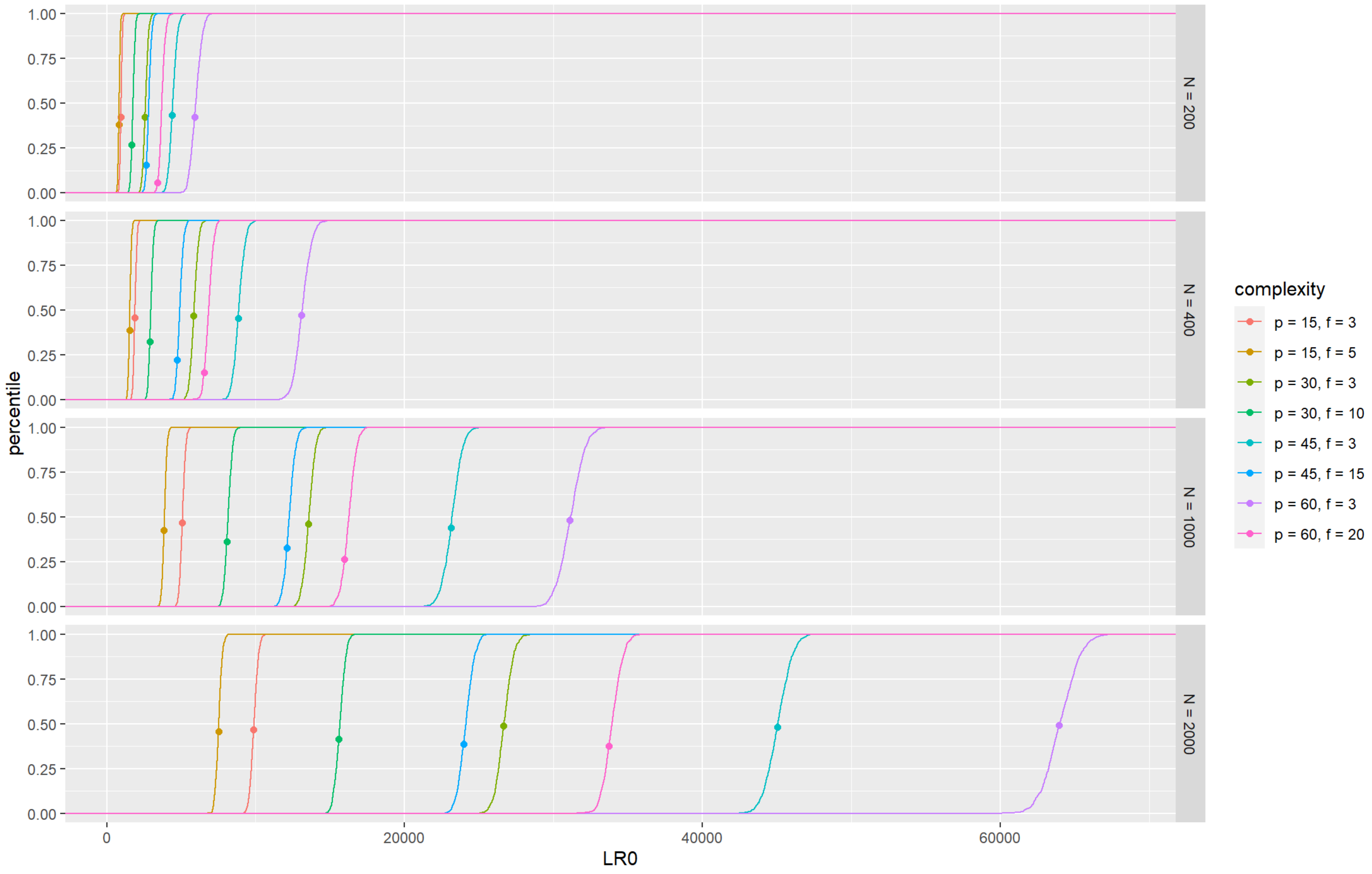

Figure 7. Empirical cumulative distribution functions of LR_0 in the well-specified condition in Phase 2 simulation. The dots represent the value from the original covariance matrix.

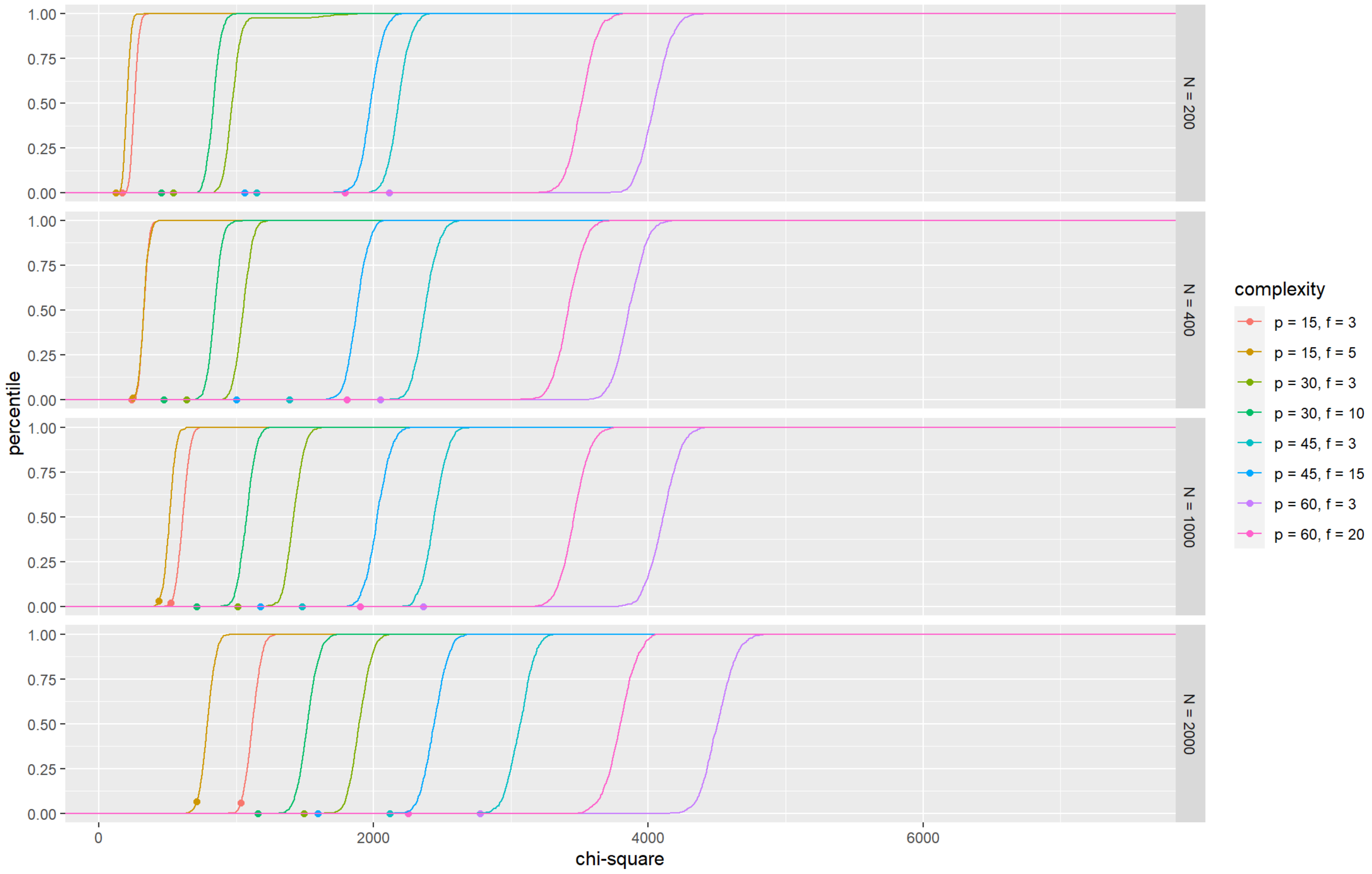

Figure 8. Empirical cumulative distribution functions of chi-square in the misspecified condition in Phase 2 simulation. The dots represent the value from the original covariance matrix.

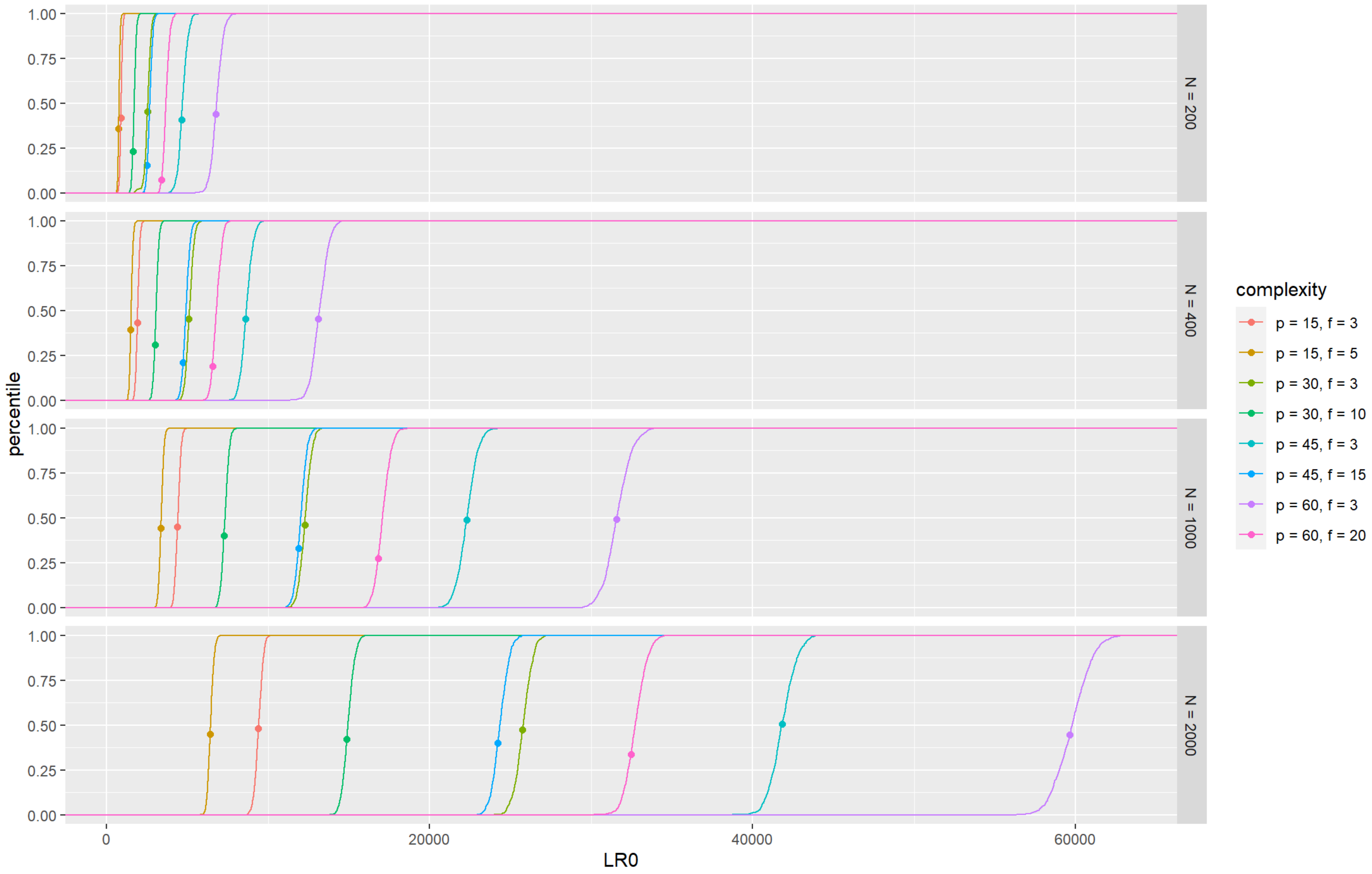

Figure 9. Empirical cumulative distribution functions of LR_0 in the misspecified condition in Phase 2 simulation. The dots represent the value from the original covariance matrix.

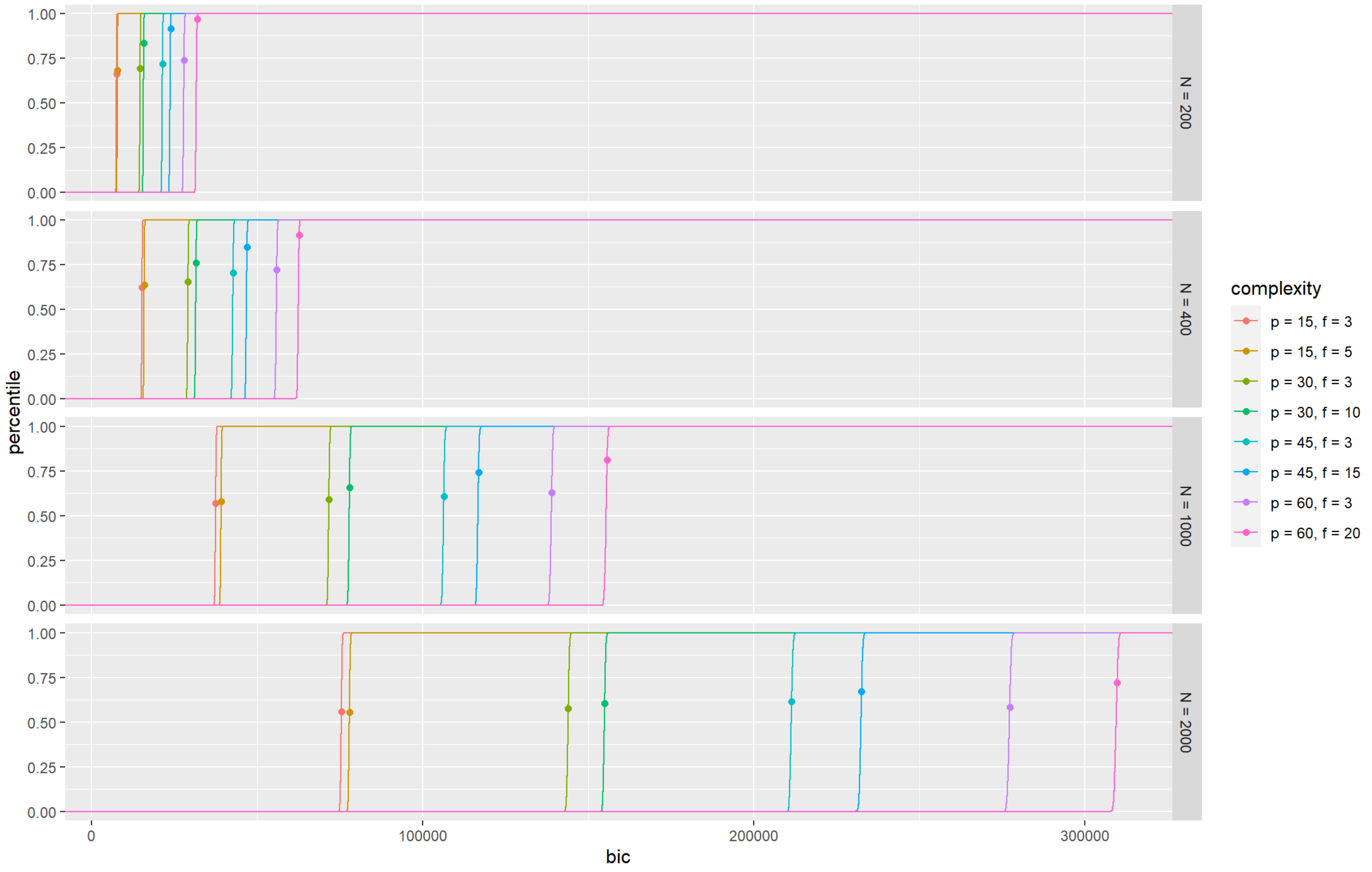

Figure 10. Empirical cumulative distribution functions of BIC in the well-specified condition in Phase 2 simulation. The dots represent the value from the original covariance matrix.

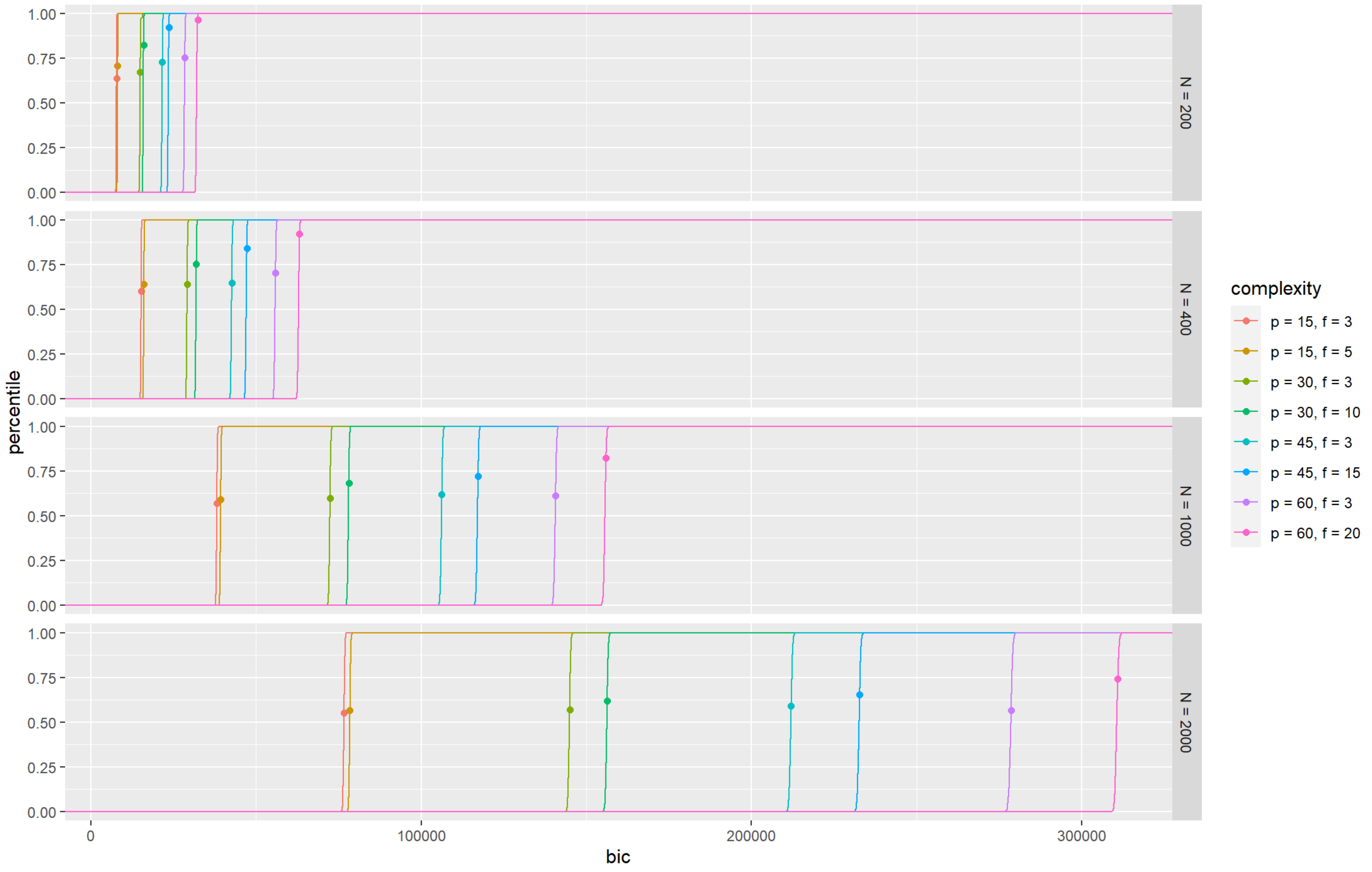

Figure 11. Empirical cumulative distribution functions of BIC in the misspecified condition in Phase 2 simulation. The dots represent the value from the original covariance matrix.

Table 1. *Result of simulated replication attempts on Guido et al. (2011)*

Condition	*N*		χ^2	CFI	TLI	RMSEA	ECVI	LogLik.	AIC	BIC
Observed	100	ORIG	28.83	1.00	1.04	0.00	0.89	-4103.42	8266.85	8345.00
		M	80.09	0.95	0.93	0.08	1.40	-4087.06	8234.12	8312.28
		SD	15.21	0.02	0.03	0.02	0.15	24.99	49.98	49.98
		%Vs.THR	-	54.84%	32.97%	16.97%	-	-	-	-
		%Vs.ORIG	0.00%	0.65%	0.00%	0.65%	0.00%	75.95%	75.95%	75.95%
	188[a]	ORIG	54.21	0.99	0.99	0.03	0.61	-7719.76	15499.52	15596.61
		M	104.47	0.95	0.94	0.08	0.87	-7705.59	15471.17	15568.26
		SD	16.84	0.01	0.02	0.01	0.09	33.87	67.74	67.74
		%Vs.THR	-	65.60%	25.52%	6.20%	-	-	-	-
		%Vs.ORIG	0.10%	0.10%	0.10%	0.10%	0.10%	66.32%	66.32%	66.32%
	200	ORIG	57.67	0.99	0.99	0.03	0.59	-8212.90	16485.79	16584.74
		M	107.16	0.95	0.94	0.08	0.84	-8197.45	16454.91	16553.86
		SD	17.60	0.01	0.02	0.01	0.09	35.93	71.87	71.87
		%Vs.THR	-	66.19%	26.58%	7.54%	-	-	-	-
		%Vs.ORIG	0.00%	0.00%	0.00%	0.00%	0.00%	68.84%	68.84%	68.84%
	400	ORIG	115.33	0.97	0.96	0.06	0.44	-16431.81	32923.63	33043.37
		M	165.97	0.96	0.94	0.08	0.56	-16418.52	32897.04	33016.78
		SD	23.80	0.01	0.01	0.01	0.06	52.07	104.15	104.15
		%Vs.THR	-	71.40%	19.34%	1.01%	-	-	-	-
		%Vs.ORIG	0.70%	1.21%	1.21%	0.70%	0.70%	58.41%	58.41%	58.41%
	1000	ORIG	288.33	0.96	0.95	0.07	0.35	-41088.55	82237.09	82384.33
		M	334.91	0.96	0.94	0.08	0.39	-41076.98	82213.96	82361.19
		SD	34.66	0.01	0.01	0.00	0.03	81.31	162.62	162.62
		%Vs.THR	-	87.93%	7.95%	0.00%	-	-	-	-
		%Vs.ORIG	8.85%	9.76%	9.76%	8.85%	8.85%	56.64%	56.64%	56.64%
	2000	ORIG	576.65	0.96	0.94	0.07	0.32	-82183.10	164426.20	164594.23
		M	623.12	0.96	0.94	0.08	0.34	-82167.63	164395.26	164563.28
		SD	49.42	0.00	0.01	0.00	0.02	111.73	223.47	223.47
		%Vs.THR	-	94.15	2.62	0.00	-	-	-	-
		%Vs.ORIG	16.33	18.04	18.04	16.33	16.33	55.04	55.04	55.04
Perfect-fit	100	ORIG	28.83	1.00	1.04	0.00	0.89	-4103.42	8266.85	8345.00
		M	50.66	0.99	0.99	0.02	1.11	-4080.93	8221.85	8300.01
		SD	10.71	0.01	0.02	0.02	0.11	25.40	50.81	50.81
		%Vs.THR	-	98.92%	96.31%	91.00%	-	-	-	-
		%Vs.ORIG	1.19%	41.00%	1.41%	41.00%	1.19%	81.02%	81.02%	81.02%
	188[a]	ORIG	54.21	0.99	0.99	0.03	0.61	-7719.76	15499.52	15596.61
		M	49.71	1.00	1.00	0.02	0.58	-7697.23	15454.46	15551.55
		SD	10.15	0.01	0.01	0.02	0.05	34.29	68.58	68.58
		%Vs.THR	-	100.00%	100.00%	99.90%	-	-	-	-
		%Vs.ORIG	69.05%	69.05%	69.05%	69.05%	69.05%	74.22%	74.22%	74.22%
	200	ORIG	57.67	0.99	0.99	0.03	0.59	-8212.90	16485.79	16584.74
		M	49.56	1.00	1.00	0.02	0.55	-8192.39	16444.79	16543.74
		SD	10.33	0.01	0.01	0.02	0.05	35.85	71.69	71.69
		%Vs.THR	-	100.00%	100.00%	99.69%	-	-	-	-
		%Vs.ORIG	79.11%	79.11%	79.11%	79.11%	79.11%	71.87%	71.87%	71.87%

Table 1. *Continued*

Condition	*N*		χ^2	CFI	TLI	RMSEA	ECVI	LogLik.	AIC	BIC
	400	ORIG	115.33	0.97	0.96	0.06	0.44	-16431.81	32923.63	33043.37
		M	48.68	1.00	1.00	0.01	0.27	-16412.48	32884.97	33004.71
		SD	10.30	0.00	0.01	0.01	0.03	48.98	97.95	97.95
		%Vs.THR	-	100.00%	100.00%	100.00%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	64.97%	64.97%	64.97%
	1000	ORIG	288.33	0.96	0.95	0.07	0.35	-41088.55	82237.09	82384.33
		M	47.61	1.00	1.00	0.01	0.11	-41066.01	82192.03	82339.26
		SD	10.01	0.00	0.00	0.01	0.01	79.29	158.59	158.59
		%Vs.THR	-	100.00%	100.00%	100.00%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	59.50%	59.50%	59.50%
	2000	ORIG	576.65	0.96	0.94	0.07	0.32	-82183.10	164426.20	164594.23
		M	48.23	1.00	1.00	0.00	0.05	-82168.19	164396.38	164564.41
		SD	9.62	0.00	0.00	0.00	0.00	109.12	218.23	218.23
		%Vs.THR	-	100.00%	100.00%	100.00%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	54.77%	54.77%	54.77%

a. Sample size of the original study is 188.

Note. ORIG = statistics of the original study; M and SD are of the replication attempts; Vs.THR is the percentage of replications that showed a better fit than thresholds suggested in West et al., (2012); Vs.ORIG is the percentage of replications that showed a better fit than the ones from the original dispersion matrix.

Table 2. *Result of simulated replication attempts on Huth-Bocks et al. (2004)*

Condition	*N*		χ^2	CFI	TLI	RMSEA	ECVI	LogLik.	AIC	BIC
Observed	100	ORIG	176.49	1.00	1.01	0.00	2.72	-3814.36	7724.72	7849.77
		M	378.86	0.84	0.82	0.10	4.75	-3791.46	7678.92	7803.96
		SD	35.13	0.03	0.03	0.01	0.35	35.30	70.59	70.59
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	0.00%	0.00%	0.00%	0.00%	0.00%	74.01%	74.01%	74.01%
	200	ORIG	352.99	0.93	0.91	0.07	2.24	-7639.30	15374.60	15532.92
		M	546.70	0.85	0.83	0.10	3.21	-7616.03	15328.07	15486.39
		SD	43.34	0.02	0.02	0.01	0.22	50.33	100.66	100.66
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	0.00%	0.00%	0.00%	0.00%	0.00%	67.60%	67.60%	67.60%
	206[a]	ORIG	363.58	0.92	0.91	0.07	2.23	-7868.80	15833.59	15993.33
		M	553.93	0.86	0.83	0.10	3.15	-7843.04	15782.07	15941.81
		SD	42.86	0.02	0.02	0.01	0.21	48.75	97.50	97.50
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	0.00%	0.00%	0.00%	0.00%	0.00%	68.88%	68.88%	68.88%
	400	ORIG	705.98	0.89	0.87	0.08	2.00	-15289.14	30674.28	30865.87
		M	894.87	0.86	0.84	0.10	2.48	-15269.67	30635.34	30826.93
		SD	58.18	0.01	0.01	0.00	0.15	69.25	138.49	138.49
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	0.10%	0.40%	0.40%	0.10%	0.10%	61.10%	61.10%	61.10%
	1000	ORIG	1764.94	0.87	0.85	0.09	1.86	-38238.63	76573.26	76808.83
		M	1947.83	0.86	0.84	0.10	2.04	-38214.22	76524.44	76760.01
		SD	86.70	0.01	0.01	0.00	0.09	110.51	221.02	221.02
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	1.00%	1.60%	1.60%	1.00%	1.00%	60.10%	60.10%	60.10%
	2000	ORIG	3529.89	0.86	0.84	0.10	1.81	-76487.77	153071.54	153340.38
		M	3714.00	0.86	0.84	0.10	1.91	-76465.26	153026.51	153295.36
		SD	121.30	0.00	0.01	0.00	0.06	154.87	309.74	309.74
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	6.10%	9.30%	9.30%	6.10%	6.10%	56.60%	56.60%	56.60%
Perfect-fit	100	ORIG	176.49	1.00	1.01	0.00	2.72	-3814.36	7724.72	7849.77
		M	201.71	0.98	0.98	0.03	2.98	-3778.62	7653.24	7778.29
		SD	21.05	0.02	0.02	0.02	0.21	33.51	67.01	67.01
		%Vs.THR	-	95.41%	91.85%	98.57%	-	-	-	-
		%Vs.ORIG	11.72%	18.25%	11.82%	18.25%	11.72%	85.73%	85.73%	85.73%
	200	ORIG	352.99	0.93	0.91	0.07	2.24	-7639.30	15374.60	15532.92
		M	193.61	0.99	0.99	0.02	1.45	-7604.71	15305.42	15463.74
		SD	20.24	0.01	0.01	0.01	0.10	47.30	94.61	94.61
		%Vs.THR	-	100.00%	100.00%	100.00%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	76.40%	76.40%	76.40%
	206[a]	ORIG	363.58	0.92	0.91	0.07	2.23	-7868.80	15833.59	15993.33
		M	192.19	0.99	1.00	0.01	1.40	-7833.79	15763.58	15923.32
		SD	20.13	0.01	0.01	0.01	0.10	45.00	90.00	90.00
		%Vs.THR	-	100.00%	100.00%	100.00%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	78.68%	78.68%	78.68%

Table 2. *Continued*

Condition	*N*		χ^2	CFI	TLI	RMSEA	ECVI	LogLik.	AIC	BIC
	400	ORIG	705.98	0.89	0.87	0.08	2.00	-15289.14	30674.28	30865.87
		M	187.52	1.00	1.00	0.01	0.71	-15256.32	30608.64	30800.23
		SD	19.77	0.00	0.01	0.01	0.05	66.77	133.54	133.54
		%Vs.THR	-	100.00%	100.00%	100.00%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	67.00%	67.00%	67.00%
	1000	ORIG	1764.94	0.87	0.85	0.09	1.86	-38238.63	76573.26	76808.83
		M	185.22	1.00	1.00	0.00	0.28	-38200.64	76497.29	76732.86
		SD	18.77	0.00	0.00	0.01	0.02	105.61	211.23	211.23
		%Vs.THR	-	100.00%	100.00%	100.00%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	63.70%	63.70%	63.70%
	2000	ORIG	3529.89	0.86	0.84	0.10	1.81	-76487.77	153071.54	153340.38
		M	184.44	1.00	1.00	0.00	0.14	-76454.19	153004.39	153273.23
		SD	18.88	0.00	0.00	0.00	0.01	141.32	282.65	282.65
		%Vs.THR	-	100.00%	100.00%	100.00%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	57.80%	57.80%	57.80%

a. Sample size of the original study is 206.

Note. ORIG = statistics of the original study; M and SD are of the replication attempts; Vs.THR is the percentage of replications that showed a better fit than thresholds suggested in West et al., (2012); Vs.ORIG is the percentage of replications that showed a better fit than the ones from the original dispersion matrix.

Table 3. *Result of simulated replication attempts on Passolunghi et al. (2007)*

Condition	*N*		χ^2	CFI	TLI	RMSEA	ECVI	LogLik.	AIC	BIC
Observed	100	ORIG	186.71	0.96	0.95	0.03	3.05	-4398.27	8914.55	9068.25
		M	413.88	0.63	0.55	0.12	5.32	-3963.91	8045.82	8199.53
		SD	38.60	0.05	0.06	0.01	0.39	34.03	68.06	68.06
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
	170[a]	ORIG	317.40	0.83	0.79	0.07	2.56	-7483.77	15085.54	15270.55
		M	571.30	0.64	0.55	0.12	4.05	-6768.07	13654.14	13839.15
		SD	49.27	0.04	0.05	0.01	0.29	47.20	94.41	94.41
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
	200	ORIG	373.43	0.81	0.77	0.08	2.46	-8806.13	17730.25	17924.85
		M	638.36	0.64	0.55	0.12	3.78	-7968.14	16054.28	16248.88
		SD	52.19	0.03	0.04	0.01	0.26	50.08	100.15	100.15
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
	400	ORIG	724.11	0.77	0.71	0.09	2.11	-17610.41	35338.82	35574.32
		M	1096.58	0.64	0.56	0.12	3.04	-15986.12	32090.25	32325.74
		SD	83.93	0.03	0.04	0.01	0.21	73.75	147.51	147.51
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
	1000	ORIG	1867.16	0.73	0.66	0.10	1.99	-44068.75	88255.49	88545.05
		M	2486.03	0.64	0.56	0.12	2.60	-40024.32	80166.63	80456.19
		SD	166.18	0.02	0.03	0.00	0.17	128.72	257.44	257.44
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
	2000	ORIG	3734.21	0.72	0.65	0.10	1.93	-88146.95	176411.90	176742.35
		M	4786.94	0.64	0.56	0.12	2.45	-80085.02	160288.04	160618.49
		SD	293.34	0.02	0.03	0.00	0.15	206.17	412.34	412.34
		%Vs.THR	-	0.00%	0.00%	0.00%	-	-	-	-
		%Vs.ORIG	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
Perfect-fit	100	ORIG	186.71	0.96	0.95	0.03	3.05	-4398.27	8914.55	9068.25
		M	186.87	0.96	0.95	0.03	3.05	-4358.24	8834.48	8988.18
		SD	20.01	0.04	0.05	0.02	0.20	32.62	65.24	65.24
		%Vs.THR	-	62.75%	53.91%	97.54%	-	-	-	-
		%Vs.ORIG	52.35%	54.14%	54.14%	52.35%	52.35%	88.48%	88.48%	88.48%
	170[a]	ORIG	317.40	0.83	0.79	0.07	2.56	-7483.77	15085.54	15270.55
		M	182.72	0.98	0.98	0.02	1.77	-7446.01	15010.02	15195.03
		SD	20.49	0.02	0.03	0.01	0.12	39.61	79.23	79.23
		%Vs.THR	-	89.04%	82.01%	99.88%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	83.97%	83.97%	83.97%
	200	ORIG	373.43	0.81	0.77	0.08	2.46	-8806.13	17730.25	17924.85
		M	180.64	0.98	0.98	0.02	1.49	-8766.78	17651.56	17846.16
		SD	23.16	0.02	0.03	0.01	0.12	45.98	91.97	91.97
		%Vs.THR	-	92.28%	87.10%	99.88%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	80.18%	80.18%	80.18%

Table 3. *Continued*

Condition	*N*		χ^2	CFI	TLI	RMSEA	ECVI	LogLik.	AIC	BIC
	400	ORIG	724.11	0.77	0.71	0.09	2.11	-17610.41	35338.82	35574.32
		M	183.99	0.99	0.99	0.01	0.75	-17577.53	35273.05	35508.55
		SD	34.65	0.02	0.02	0.01	0.09	63.26	126.53	126.53
		%Vs.THR	-	95.83%	94.93%	100.00%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	68.21%	68.21%	68.21%
	1000	ORIG	1867.16	0.73	0.66	0.10	1.99	-44068.75	88255.49	88545.05
		M	195.20	0.99	0.99	0.01	0.31	-44040.72	88199.45	88489.01
		SD	64.51	0.01	0.02	0.01	0.06	103.77	207.54	207.54
		%Vs.THR	-	95.70%	95.57%	100.00%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	61.21%	61.21%	61.21%
	2000	ORIG	3734.21	0.72	0.65	0.10	1.93	-88146.95	176411.90	176742.35
		M	216.17	0.99	0.99	0.01	0.17	-88132.52	176383.04	176713.49
		SD	133.98	0.02	0.02	0.01	0.07	148.02	296.05	296.05
		%Vs.THR	-	95.73%	95.73%	99.85%	-	-	-	-
		%Vs.ORIG	100.00%	100.00%	100.00%	100.00%	100.00%	55.82%	55.82%	55.82%

a. Sample size of the original study is 170.

Note. ORIG = statistics of the original study; M and SD are of the replication attempts; Vs.THR is the percentage of replications that showed a better fit than thresholds suggested in West et al., (2012); Vs.ORIG is the percentage of replications that showed a better fit than the ones from the original dispersion matrix.

Table 4. Simulation results of well-specified models.

p	f	q	df	df_0	N	df_{LR0}	*Orig. LR_0*	M_{LR0}	SD_{LR0}	LR_0 > orig.	*Orig. χ^2*	$M_{\chi 2}$	$SD_{\chi 2}$	χ^2 < orig.
15	3	33	87	105	200	18	950.22	966.07	81.08	57.70%	64.09	154.65	20.81	0.00%
15	3	33	87	105	400	18	1882.40	1901.81	114.60	54.40%	82.79	171.58	22.33	0.00%
15	3	33	87	105	1000	18	5076.10	5092.03	184.18	53.30%	89.60	177.94	22.93	0.00%
15	3	33	87	105	2000	18	9870.07	9891.97	260.19	53.30%	90.43	177.38	22.88	0.00%
15	5	40	80	105	200	25	810.55	835.01	72.76	62.10%	78.97	161.56	21.50	0.00%
15	5	40	80	105	400	25	1526.54	1556.22	100.21	61.40%	68.20	150.10	21.74	0.00%
15	5	40	80	105	1000	25	3846.52	3879.43	151.86	57.60%	80.30	160.88	22.52	0.00%
15	5	40	80	105	2000	25	7503.75	7524.30	213.67	54.30%	80.17	160.47	22.87	0.00%
30	3	63	402	435	200	33	2573.29	2609.46	175.64	57.60%	413.37	842.22	49.74	0.00%
30	3	63	402	435	400	33	5842.40	5866.75	259.89	53.10%	451.31	866.04	51.94	0.00%
30	3	63	402	435	1000	33	13537.09	13578.78	385.26	54.00%	402.08	807.97	49.23	0.00%
30	3	63	402	435	2000	33	26658.88	26684.78	553.02	51.20%	392.32	799.38	47.85	0.00%
30	10	105	360	435	200	75	1685.70	1767.18	130.12	73.30%	322.61	704.89	46.05	0.00%
30	10	105	360	435	400	75	2892.77	2967.14	151.75	67.40%	310.94	680.78	45.10	0.00%
30	10	105	360	435	1000	75	8066.13	8155.19	251.82	63.50%	331.76	696.41	46.24	0.00%
30	10	105	360	435	2000	75	15583.71	15659.03	348.99	58.40%	366.41	729.78	46.77	0.00%
45	3	93	942	990	200	48	4390.68	4440.83	268.65	56.70%	996.27	2034.90	76.72	0.00%
45	3	93	942	990	400	48	8823.67	8872.84	372.85	54.50%	990.12	1973.48	78.73	0.00%
45	3	93	942	990	1000	48	23119.85	23191.22	617.06	56.10%	950.99	1909.05	73.72	0.00%
45	3	93	942	990	2000	48	45020.93	45061.98	833.92	52.00%	943.67	1891.28	76.14	0.00%
45	15	195	840	990	200	150	2663.32	2828.51	167.70	84.40%	1003.88	1926.37	73.70	0.00%
45	15	195	840	990	400	150	4744.68	4911.41	214.04	77.80%	903.39	1782.10	73.83	0.00%
45	15	195	840	990	1000	150	12101.70	12263.76	363.27	67.20%	852.91	1713.03	73.19	0.00%
45	15	195	840	990	2000	150	23957.63	24115.69	507.87	61.30%	846.91	1696.55	70.06	0.00%
60	3	123	1707	1770	200	63	5910.93	5987.55	371.08	57.70%	2084.73	4022.48	110.04	0.00%
60	3	123	1707	1770	400	63	13091.80	13135.06	542.32	52.90%	1754.77	3568.70	103.06	0.00%
60	3	123	1707	1770	1000	63	31101.95	31152.11	827.55	51.90%	1842.89	3590.12	107.40	0.00%
60	3	123	1707	1770	2000	63	63950.35	64016.64	1151.44	50.90%	1691.22	3412.26	102.50	0.00%
60	20	310	1520	1770	200	250	3425.09	3698.32	196.18	92.59%	1806.96	3532.62	98.72	0.00%
60	20	310	1520	1770	400	250	6548.69	6811.93	290.71	82.00%	1636.73	3247.14	94.08	0.00%
60	20	310	1520	1770	1000	250	15973.99	16256.06	449.68	73.50%	1555.06	3109.80	95.46	0.00%
60	20	310	1520	1770	2000	250	33728.41	33934.77	634.05	61.70%	1582.25	3117.57	95.95	0.00%

Note. p = no. of observed variables; f = no. of latent factors; q = no. of parameters estimated; df_0 = degrees of freedom of baseline model; N = sample size; LR_0 and subscript LR0 = statistics regarding baseline likelihood ratio; χ^2 and subscript χ^2 = statistics regarding standard likelihood ratio. M and SD are of replication attempts.

Table 5. Simulation results of misspecified models.

p	*f*	*q*	*df*	*df*$_0$	*N*	*df*$_{\mathrm{LR0}}$	*Orig. LR*$_0$	*M*$_{\mathrm{LR0}}$	*SD*$_{\mathrm{LR0}}$	*LR*$_0$ > orig.	*Orig.* χ^2	*M* $_{\chi 2}$	*SD* $_{\chi 2}$	χ^2 < orig.
15	3	33	87	105	200	18	884.18	904.07	83.39	58.10%	168.99	260.12	28.65	0.10%
15	3	33	87	105	400	18	1910.89	1932.40	121.46	56.60%	238.72	328.74	31.23	0.00%
15	3	33	87	105	1000	18	4396.98	4417.38	175.74	55.00%	524.83	612.00	44.27	2.10%
15	3	33	87	105	2000	18	9405.63	9420.50	256.86	51.90%	1033.78	1120.17	56.58	5.80%
15	5	40	80	105	200	25	747.83	775.16	66.31	63.65%	123.44	205.42	25.47	0.00%
15	5	40	80	105	400	25	1502.10	1531.27	103.30	60.60%	249.69	328.55	33.95	0.90%
15	5	40	80	105	1000	25	3363.37	3387.93	146.35	55.80%	435.51	514.87	40.85	3.20%
15	5	40	80	105	2000	25	6418.09	6446.85	188.40	54.90%	710.99	790.93	52.42	6.60%
30	3	63	402	435	200	33	2530.41	2543.26	206.27	54.50%	541.43	987.73	124.82	0.00%
30	3	63	402	435	400	33	5097.86	5140.79	242.68	54.50%	636.40	1051.04	58.32	0.00%
30	3	63	402	435	1000	33	12268.45	12303.00	370.94	53.80%	1012.65	1420.01	65.20	0.00%
30	3	63	402	435	2000	33	25762.45	25795.47	531.92	52.70%	1493.13	1898.82	74.52	0.00%
30	10	105	360	435	200	75	1626.16	1712.11	117.98	76.23%	452.41	835.55	49.42	0.00%
30	10	105	360	435	400	75	2992.64	3068.53	156.66	69.20%	471.05	842.74	50.97	0.00%
30	10	105	360	435	1000	75	7267.15	7334.32	231.86	59.90%	709.85	1075.23	59.62	0.00%
30	10	105	360	435	2000	75	14880.16	14960.11	362.84	57.90%	1159.99	1520.66	72.20	0.00%
45	3	93	942	990	200	48	4642.07	4711.78	300.98	59.18%	1147.94	2184.03	86.75	0.00%
45	3	93	942	990	400	48	8618.30	8663.12	362.15	54.50%	1386.15	2376.97	84.26	0.00%
45	3	93	942	990	1000	48	22286.97	22314.06	608.76	51.00%	1481.32	2443.44	83.34	0.00%
45	3	93	942	990	2000	48	41832.42	41846.59	797.55	49.30%	2119.90	3065.31	93.38	0.00%
45	15	195	840	990	200	150	2513.28	2680.75	160.66	84.31%	1063.41	1981.58	77.83	0.00%
45	15	195	840	990	400	150	4743.11	4921.76	226.01	78.80%	1001.85	1879.02	73.94	0.00%
45	15	195	840	990	1000	150	11880.79	12028.47	340.07	67.00%	1174.97	2030.04	83.27	0.00%
45	15	195	840	990	2000	150	24228.40	24369.54	499.16	59.80%	1596.53	2442.22	86.57	0.00%
60	3	123	1707	1770	200	63	6757.76	6818.17	376.64	55.75%	2112.18	4048.26	124.00	0.00%
60	3	123	1707	1770	400	63	13113.47	13189.15	526.26	54.67%	2050.35	3866.88	155.83	0.00%
60	3	123	1707	1770	1000	63	31566.85	31618.94	813.09	50.80%	2362.52	4108.38	111.25	0.00%
60	3	123	1707	1770	2000	63	59643.02	59780.73	1104.56	55.40%	2773.87	4507.71	114.33	0.00%
60	20	310	1520	1770	200	250	3402.38	3688.05	204.06	92.59%	1794.09	3514.61	99.86	0.00%
60	20	310	1520	1770	400	250	6552.68	6812.01	292.15	81.00%	1804.50	3419.24	101.65	0.00%
60	20	310	1520	1770	1000	250	16823.49	17101.74	459.96	72.50%	1902.60	3460.84	101.18	0.00%
60	20	310	1520	1770	2000	250	32480.35	32751.85	657.87	66.30%	2250.86	3795.68	108.90	0.00%

Note. *p* = no. of observed variables; *f* = no. of latent factors; *q* = no. of parameters estimated; *df*$_0$ = degrees of freedom of baseline model; *N* = sample size; *LR*$_0$ and subscript LR0 = statistics regarding baseline likelihood ratio; χ^2 and subscript χ^2 = statistics regarding standard likelihood ratio. M and SD are of replication attempts.

Appendix A

Diagrams For the Models Used in Examples

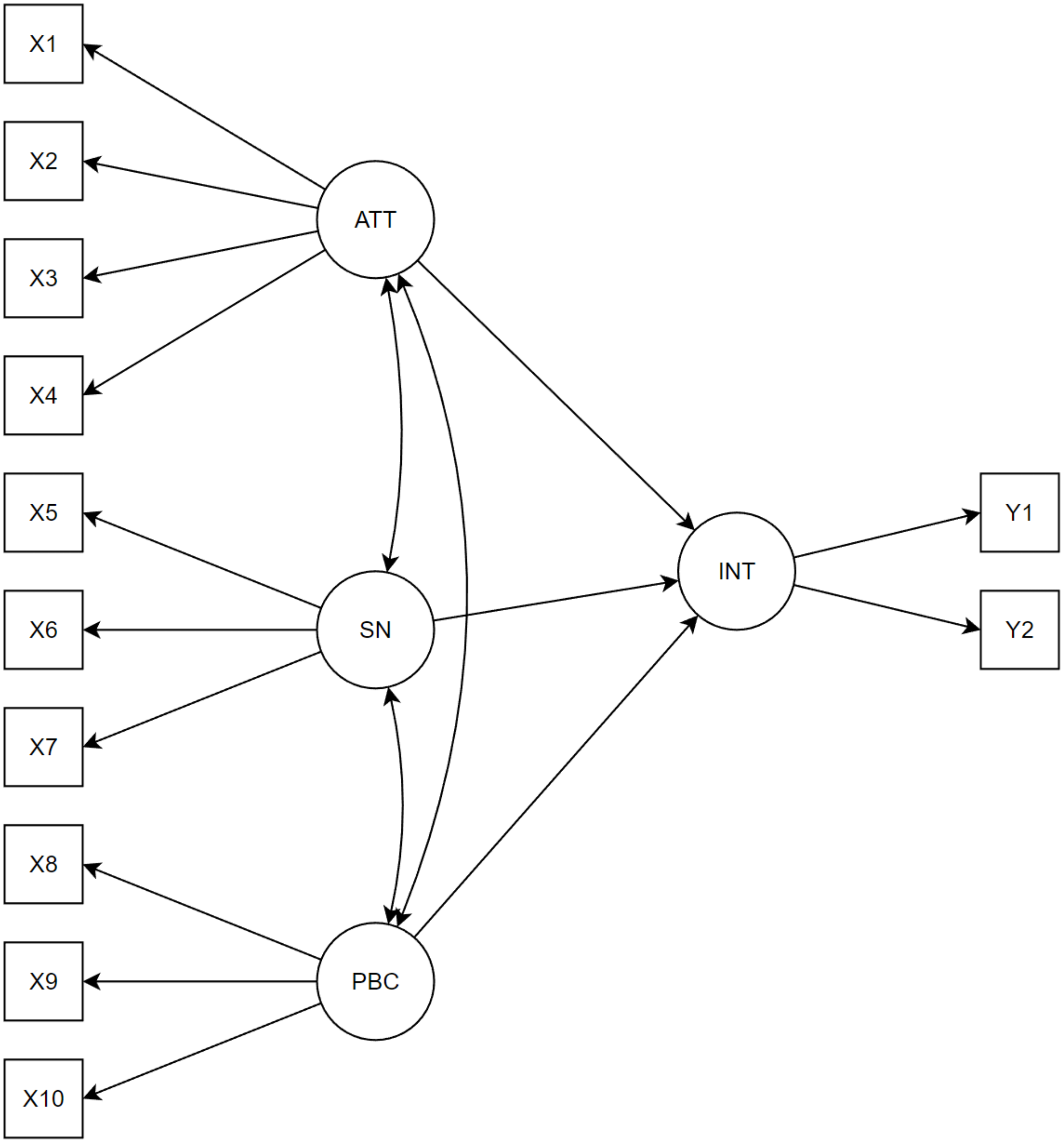

Figure A1. The model specification of Guido et al., (2017). The number of parameters estimated was 30, and the degrees of freedom was 48.

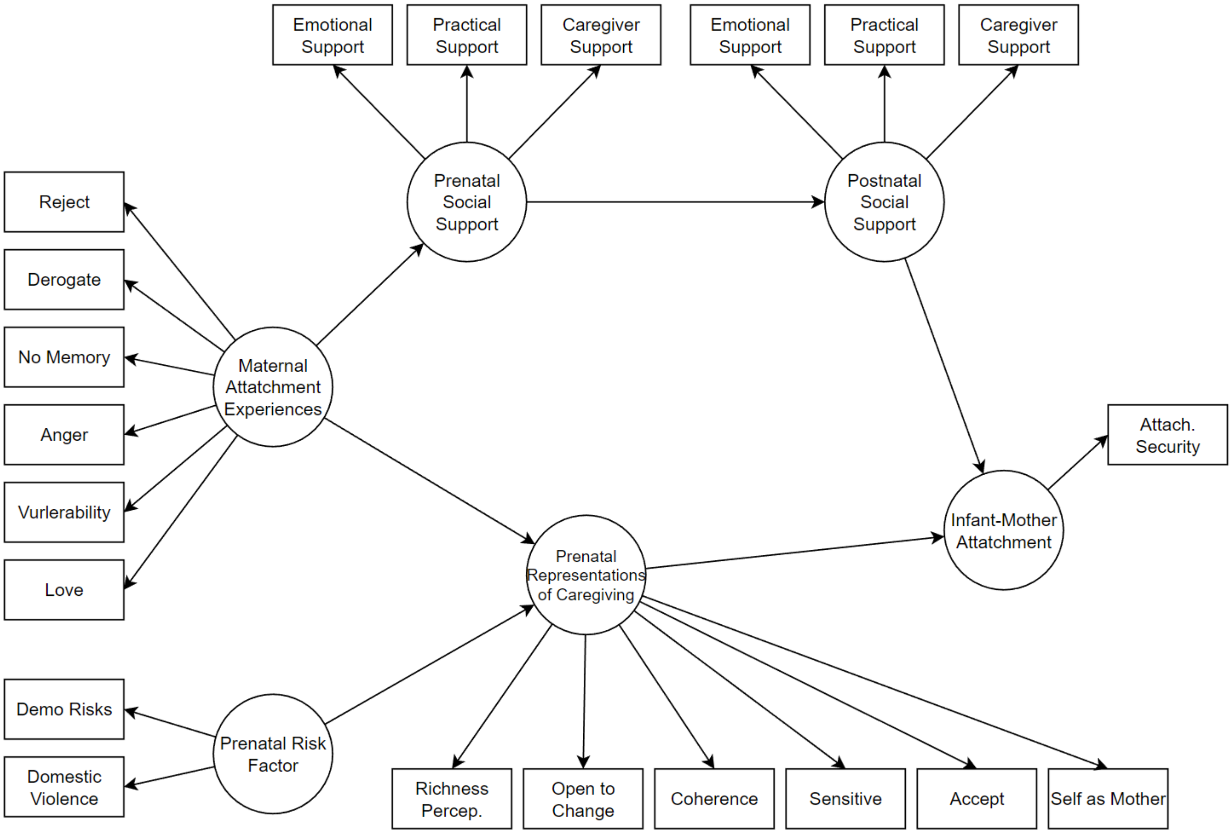

Figure A2. The model specification on Huth-Bocks et al., (2004). The number of parameters estimated was 48, and the degrees of freedom was 183.

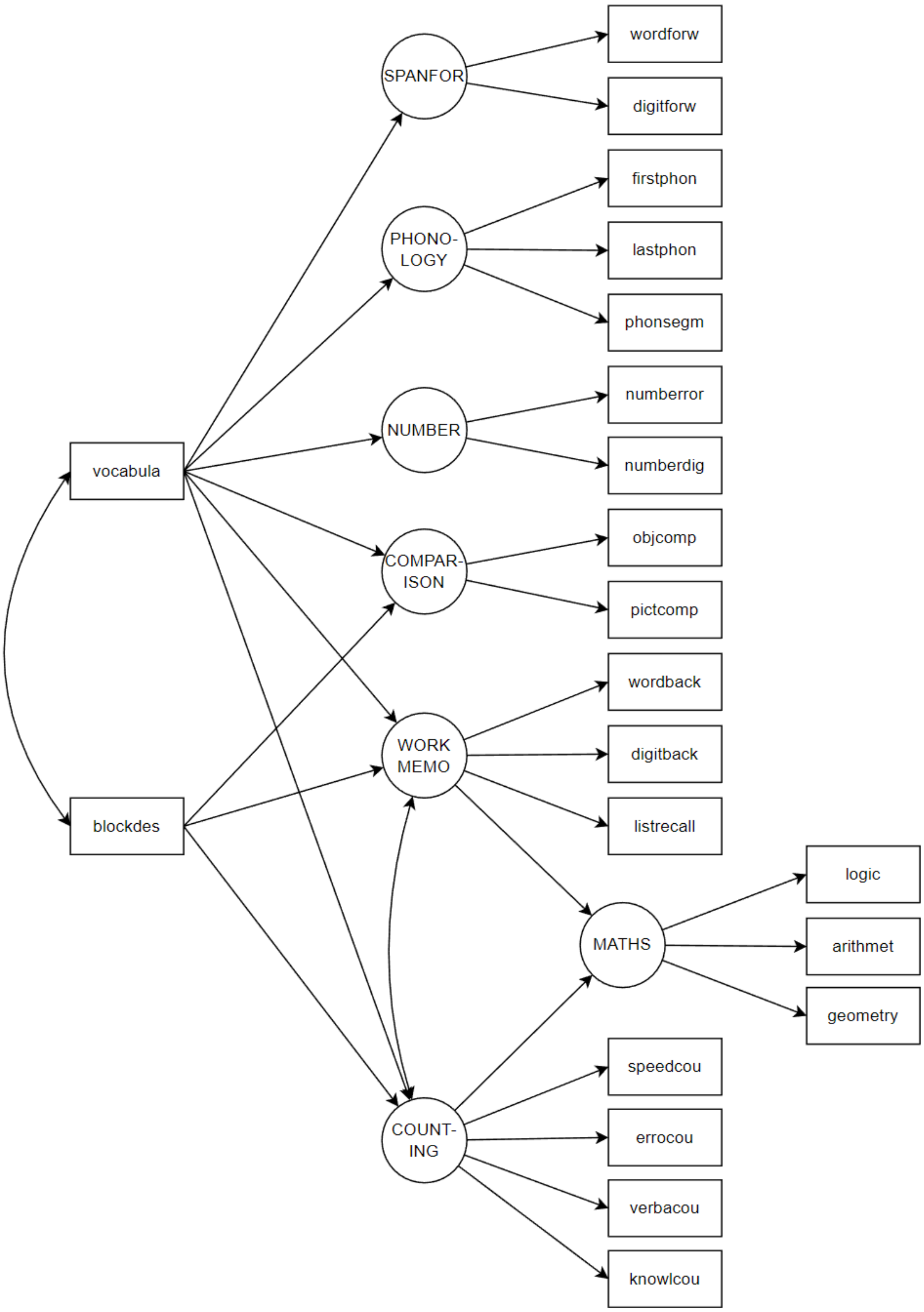

Figure A3. The model specification of Model 5 on Passolunghi et al., (2007). The number of parameters estimated was 59, and the degrees of freedom was 169.

Appendix B

Supplement Information on Simulation in Phase 2

Table B1. The simulation result for NFI in Phase 2.

					Well-specified Condition					Misspecified Condition				
p	*f*	*q*	*df*	*N*	Orig. NFI	M_{NFI}	SD_{NFI}	% v.Og.	% v.THR.	Orig. NFI	M_{NFI}	SD_{NFI}	% v.Og.	% v.THR.
15	3	33	87	200	0.937	0.862	0.020	0.00%	0.00%	0.840	0.776	0.025	0.30%	0.00%
15	3	33	87	400	0.958	0.917	0.011	0.00%	0.00%	0.889	0.854	0.013	0.20%	0.00%
15	3	33	87	1000	0.983	0.966	0.004	0.00%	100.00%	0.893	0.878	0.009	3.30%	0.00%
15	3	33	87	2000	0.991	0.982	0.002	0.00%	100.00%	0.901	0.894	0.005	7.70%	0.00%
15	5	40	80	200	0.911	0.837	0.022	0.00%	0.00%	0.858	0.790	0.025	0.00%	0.00%
15	5	40	80	400	0.957	0.912	0.013	0.00%	0.00%	0.857	0.823	0.017	1.50%	0.00%
15	5	40	80	1000	0.980	0.960	0.006	0.10%	95.30%	0.885	0.868	0.010	4.80%	0.00%
15	5	40	80	2000	0.989	0.979	0.003	0.00%	100.00%	0.900	0.891	0.007	8.60%	0.00%
30	3	63	402	200	0.862	0.756	0.017	0.00%	0.00%	0.824	0.720	0.036	0.00%	0.00%
30	3	63	402	400	0.928	0.871	0.009	0.00%	0.00%	0.889	0.830	0.010	0.00%	0.00%
30	3	63	402	1000	0.971	0.944	0.004	0.00%	3.60%	0.924	0.896	0.005	0.00%	0.00%
30	3	63	402	2000	0.985	0.971	0.002	0.00%	100.00%	0.945	0.931	0.003	0.00%	0.00%
30	10	105	360	200	0.839	0.714	0.020	0.00%	0.00%	0.782	0.672	0.021	0.00%	0.00%
30	10	105	360	400	0.903	0.813	0.013	0.00%	0.00%	0.864	0.784	0.013	0.00%	0.00%
30	10	105	360	1000	0.960	0.921	0.005	0.00%	0.00%	0.911	0.872	0.007	0.00%	0.00%
30	10	105	360	2000	0.977	0.955	0.003	0.00%	97.00%	0.928	0.908	0.004	0.00%	0.00%
45	3	93	942	200	0.815	0.685	0.016	0.00%	0.00%	0.802	0.683	0.016	0.00%	0.00%
45	3	93	942	400	0.899	0.818	0.009	0.00%	0.00%	0.861	0.785	0.009	0.00%	0.00%
45	3	93	942	1000	0.960	0.924	0.003	0.00%	0.00%	0.938	0.901	0.004	0.00%	0.00%
45	3	93	942	2000	0.979	0.960	0.002	0.00%	100.00%	0.952	0.932	0.002	0.00%	0.00%
45	15	195	840	200	0.726	0.595	0.017	0.00%	0.00%	0.703	0.575	0.018	0.00%	0.00%
45	15	195	840	400	0.840	0.734	0.012	0.00%	0.00%	0.826	0.723	0.012	0.00%	0.00%
45	15	195	840	1000	0.934	0.877	0.006	0.00%	0.00%	0.910	0.856	0.006	0.00%	0.00%
45	15	195	840	2000	0.966	0.934	0.003	0.00%	0.00%	0.938	0.909	0.003	0.00%	0.00%
60	3	123	1707	200	0.739	0.598	0.016	0.00%	0.00%	0.762	0.627	0.015	0.00%	0.00%
60	3	123	1707	400	0.882	0.786	0.008	0.00%	0.00%	0.865	0.773	0.011	0.00%	0.00%

					Well-specified Condition					Misspecified Condition				
p	*f*	*q*	*df*	*N*	Orig. NFI	M_{NFI}	SD_{NFI}	% v.Og.	% v.THR.	Orig. NFI	M_{NFI}	SD_{NFI}	% v.Og.	% v.THR.
60	3	123	1707	1000	0.944	0.897	0.004	0.00%	0.00%	0.930	0.885	0.004	0.00%	0.00%
60	3	123	1707	2000	0.974	0.949	0.002	0.00%	34.00%	0.956	0.930	0.002	0.00%	0.00%
60	20	310	1520	200	0.655	0.511	0.015	0.00%	0.00%	0.655	0.512	0.015	0.00%	0.00%
60	20	310	1520	400	0.800	0.677	0.011	0.00%	0.00%	0.784	0.666	0.012	0.00%	0.00%
60	20	310	1520	1000	0.911	0.839	0.006	0.00%	0.00%	0.898	0.832	0.006	0.00%	0.00%
60	20	310	1520	2000	0.955	0.916	0.003	0.00%	0.00%	0.935	0.896	0.003	0.00%	0.00%

Note. Orig. NFI: The “original” value of NFI. The original sample was generated under population, while the replication samples are generated using the original covariance matrix.

MNFI: The mean of NFI observed in replicated samples.

SDNFI: The standard deviation of NFI in replicated samples.

% v.Og.: The percentage of replication attempts that have a better fit than the original.

% v.THR.: The percentage of replication attempts that have a better fit than the original.