Structural Equation Modeling With Many Variables: A Systematic Review of Issues and Developments

Lifang Deng[1]*, Miao Yang[2] and Katerina M. Marcoulides[3]

[1] Department of Psychology, Beihang University, Beijing, China, [2] Department of Psychology, University of Notre Dame, Notre Dame, IN, United States, [3] School of Human Development and Organizational Studies in Education, University of Florida, Gainesville, FL, United States

Survey data in social, behavioral, and health sciences often contain many variables (p). Structural equation modeling (SEM) is commonly used to analyze such data. With a sufficient number of participants (N), SEM enables researchers to easily set up and reliably test hypothetical relationships among theoretical constructs as well as those between the constructs and their observed indicators. However, SEM analyses with small N or large p have been shown to be problematic. This article reviews issues and solutions for SEM with small N, especially when p is large. The topics addressed include methods for parameter estimation, test statistics for overall model evaluation, and reliable standard errors for evaluating the significance of parameter estimates. Previous recommendations on required sample size N are also examined together with more recent developments. In particular, the requirement for N with conventional methods can be a lot more than expected, whereas new advances and developments can reduce the requirement for N substantially. The issues and developments for SEM with many variables described in this article not only let applied researchers be aware of the cutting edge methodology for SEM with big data as characterized by a large p but also highlight the challenges that methodologists need to face in further investigation.

Keywords: structural equation modeling, small sample size, parameter estimates, test statistics, stand errors

1. INTRODUCTION

Many important attributes in the social, behavioral, and health sciences cannot be observed directly. Examples of such attributes include happiness, depression, anxiety, cognitive and social competence, etc. They are typically measured by multiple indicators that are often subject to measurement errors. Structural equation modeling (SEM) has become a major tool for examining and understanding relationships among latent attributes. Existing SEM methods are developed using asymptotics by assuming a large number of observations (N) and a small number of variables (p). However, with survey data or data collected using questionnaires, p can be rather large while N may be limited due to the high costs associated with obtaining a sufficient number of participants in the data collection. In such instances, blindly applying SEM methods developed using asymptotics can easily either result in misleading results or in unattainable parameter estimates due to non-convergences in computation. This article reviews the development of methods that aim to address small sample issues in SEM particularly when large numbers of variables are involved.

Note that a small or large sample size in SEM and other statistical modeling is relative to the number of variables, a method that works well with a small sample size is also expected to work well with a large number of variables. While our focus is mainly on methodology development, we will also discuss various recommendations on required sample size N in the SEM literature. These recommendations are typically based on limited simulation results using the method of normal-distribution-based maximum likelihood (NML). In particular, the requirement for N with the conventional NML method can be a lot more than expected to obtain reliable results, whereas new advances and developments can reduce the requirement for N substantially. We hope that our discussion of the various pros and cons of different methods will make researchers be aware of the potential problems and issues that might arise when using SEM with many variables, and the information offered would allow them to identify a method that is most suitable for their research when facing a data set with either a small N or a large p.

Let $\mathbf{S}$ be a sample covariance matrix based on a sample of size N with p variables. Currently, the most widely used method for SEM analysis is to fit $\mathbf{S}$ by a structural model using NML. Overall model evaluation is conducted by comparing the likelihood ratio statistic T_{ml} against the nominal chi-square distribution or by some fit indices such as root mean square error of approximation (RMSEA, Steiger and Lind, 1980), and/or the comparative fit index (CFI, Bentler, 1990). Standard errors of parameter estimates are obtained by inverting the normal-distribution-based information matrix. When data are normally distributed and the sample size is sufficiently large, the NML procedure is expected to perform well. In practice, data tend to be non-normally distributed (Micceri, 1989). Under such circumstance one may choose the Satorra-Bentler rescaled and adjusted statistics for overall model evaluation, and the sandwich-type covariance matrix for computing standard errors of parameter estimates (Satorra and Bentler, 1994). However, these procedures are not reliable when the number of variables p is relatively large, since their validity is justified by asymptotics. In particular, when p is large, the conventional SEM methods as implemented in most available software may fail to generate a set of parameter estimates due to the non-convergence that might occur during computation. Additionally, the likelihood ratio statistic may reject the correct model 100% of the time even when data are normally distributed.

The issue of small N with a relatively large p has been discussed by many authors in the extant literature (e.g., Barrett and Kline, 1981; Bentler and Chou, 1987; Jackson, 2001; de Winter et al., 2009; Xing and Yuan, 2017), in some cases aiming to reduce the requirement for N or to give a rule of thumb on the required sample size for properly conducting SEM. We try to provide an overview of up-to-date developments in methodology by addressing small sample issues in SEM. They include procedure for addressing the problems of near singular covariance matrix due to small N, obtaining more efficient parameter estimates with non-normally distributed data, improving the performance of test statistics, and procedures for more accurate standard errors (SEs). In the rest of the article, we

first discuss various recommendations related to required sample sizes, including those proposed for exploratory factor analysis. Then we turn our attention to parameter estimation, where we discuss recent developments related to the use of ridge methods as well as penalized likelihoods. Overall model evaluation is discussed next, whereupon we review *ad-hoc* corrections as well as principled corrections to the likelihood ratio statistic and to the Satorra and Bentler's rescaled statistic. Standard errors are discussed subsequently, and includes a discussion of some advances that have been recently made. A real data example is presented next, contrasting conventional methods with recently developed new methods. Finally, we present some general recommendations, highlight some limitations, and conclude with emphasizing important remaining challenging issues in need of future attention.

2. *AD-HOC* RECOMMENDATIONS CONCERNING SAMPLE SIZE

Many scholars have studied sample size issues in SEM and factor analysis. Earlier research noted that reasonable results could be obtained in SEM analyses when N is <200 (Gerbing and Anderson, 1985), or at least above 100 (Boomsma, 1985). While these recommendations were supported by Monte Carlo results, the number of variables in these studies was rather small. Bentler and Chou (1987) subsequently noted that sample size N should instead be considered relative to the number of parameters q, and the ratio of $N:q$ can be as low as 5:1 for normally distributed data, and 10:1 for arbitrary distributions. However, recent studies have suggested that, as p increase, we in fact need an N that is much large than 200 in order for T_{ml} to perform as expected. In particular, it was found that at the nominal level of 5%, T_{ml} rejected a correct model 100% when $N = 200$ and $p = 90$ (Moshagen, 2012); and rejected correct models around 85% when $N = 1,000$ and $p = 120$ (Shi et al., 2018). In addition, the studies by Jackson (2001), Moshagen and Shi et al. indicated that the behavior of T_{ml} is little affected by the number of parameters.

The issue of not having sufficiently large sample sizes has also been a major concern in exploratory factor analysis (EFA) since most well-known psychological scales contain many items, Within the context of EFA, the recommendations are quite varied, including for example for N to be above 100 and even up to above 1000 (Guilford, 1954; Kline, 1979; Gorsuch, 1983; Comrey and Lee, 1992), In terms of the ratio N/p, it has been suggested as needing to be anywhere from above 1.2 to up to 10 (Everitt, 1975; Cattell, 1978; Barrett and Kline, 1981; Gorsuch, 1983; Arrindell and van der Ende, 1985). MacCallum et al. (1999), however, argued that the necessary N is in fact dependent on other conditions in addition to N/p, including communality and the number of indicators per factor. These results were also confirmed by Preacher and MacCallum (2002) and Mundfrom et al. (2005). Interestingly, de Winter et al. (2009) noted that N can be much smaller than p if both the size of communality and the number of indicators per factor are high. However, the discussion and recommendations concerning the required sample size needed in EFA are mostly for the purpose of factor

recovery[1] instead of overall model evaluation or inference of parameter estimates.

While factor recovery in EFA might seem conceptually different from the issue of statistical inference in SEM or CFA, and the main focus in this article is model estimation and evaluation, our discussion and review will inevitably provide insight on inference issues in EFA as well. In particular, because factor recovery is closely related to standard errors of factor loading estimates, conditions or methods corresponding to more efficient parameter estimates will yield better factor recovery in EFA. For example, Yuan et al. (2010) have shown that standard errors (SEs) of factor loading estimates in CFA increase with the size of error variances, and decreases with the number of indicators per factor and the size of factor loadings other than that corresponding to the SE itself. We would expect these results also hold for standard errors of factor loading estimates in EFA. However, smaller values of SEs in EFA and more accurate estimates of SEs are two different concepts. Also, the conditions on p and m (the number of factors) that yield smaller SEs for parameter estimates are different from the conditions that lead to more reliable test statistics for overall model evaluation. For example, Shi et al. (2018) found that the distribution of T_{ml} is little affected by the number of indicators per factor (p/m), and the performance of T_{ml} becomes worse when p increases while holding N constant.

In summary, recommendations on required sample sizes in the literature of SEM and EFA are all simply *ad-hoc* conjectures. Although they might be based on small simulation studies with varied N but limited values of p, they are not justified statistically, nor generalizable to conditions with large p, as also noted by Yang et al. (2018).

3. PARAMETER ESTIMATION

A SEM model can be formulated according to theory or information gained at the exploratory stage with EFA. Once a model is formulated, we need to obtain parameter estimates before model evaluation. We discuss several major methods in this section, including normal distribution based maximum likelihood (NML), generalized least squares (GLS), the Bayesian approach, penalized likelihood, and some related methods to yield more stable/efficient parameter estimates.

3.1. NML and Ridge ML

The most widely used method for parameter estimation in SEM is NML, which is equivalent to minimizing the discrepancy function

$$F_{ml}(\boldsymbol{\theta}) = \text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})) - \log|\mathbf{S}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})| - p \qquad (1)$$

---

[1]Factor recovery is commonly measured by a so-called congruence coefficient (https://en.wikipedia.org/wiki/Congruence_coefficient), which is a formula parallel to the product-moment correlation between the population factor loadings and their sample counterparts without centralizing at the mean. Preacher and MacCallum (2002) used root mean squared differences between the population correlations and model implied correlations for measuring factor recovery.

for estimating $\boldsymbol{\theta}$, where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the structural model. Unlike in linear regression where there is an analytical formula for estimating the regression coefficients, we have to use an iterative procedure to minimize $F_{ml}(\boldsymbol{\theta})$, and the Fisher-scoring algorithm (see Lee and Jennrich, 1979) is typically used for such a purpose. Small N or large p can cause various problems when minimizing $F_{ml}(\boldsymbol{\theta})$, including near singular covariance matrices due to not having enough distinct observations and slower convergence due to large sampling errors. When a sample does not contain a sufficient number of distinct cases, the sample covariance matrix $\mathbf{S}$ is near singular (not full rank). Then, the iteration process for computing the NML estimates can be very unstable, and it may take literally hundreds of iterations to reach a convergence (Yuan and Bentler, 2017). When $\mathbf{S}$ is literally singular, equation (1) is not defined, and other methods for parameter estimation will likely break down as well. Even when the sample size is quite large, $\mathbf{S}$ can be near singular in real data analysis due to multi-collinearity (Wothke, 1993), especially when p is large. A great deal of research has been directed to address the problems encountered with near singular covariance matrices, and to increase the chance and speed of convergence in computation at the same time.

When $\mathbf{S}$ is singular, the program LISREL (Jöreskog and Sörbom, 1993) provides an option of ridge SEM by replacing the $\mathbf{S}$ in the NML discrepancy function in Equation (1), with $\mathbf{S} + k\text{diag}(s_{11}, \ldots, s_{pp})$, where $k > 0$ and s_{jj} is the sample variance of the jth variable. However, statistical properties of the resulting ridge parameter estimates and test statistics have not been obtained analytically or asymptotically. Through empirical studies, McQuitty (1997, p. 251) concluded that "there appears to be ample evidence that structural equation models should not be estimated with LISREL's ridge option unless the estimation of unstandardized factor loadings is the only goal." A ridge technique that is different from the one implemented in LISREL was developed by Yuan and Chan (2008). They proposed to replace $\mathbf{S}$ in equation (1) by $\mathbf{S}_a = \mathbf{S} + a\mathbf{I}$ and recommended choosing $a = p/N$. Let the resulting discrepancy function be denoted by $F_{mla}(\boldsymbol{\theta})$, and call the procedure of minimizing $F_{mla}(\boldsymbol{\theta})$ ridge ML. Yuan and Chan (2008) showed that both the speed of convergence and convergence rate of the Fisher-scoring algorithm in ridge ML are much higher than in ML. They also showed that, with smaller Ns, ridge ML yields consistent and more efficient parameter estimates $\hat{\boldsymbol{\theta}}_a$ than ML even when data are normally distributed. They further proposed a rescaled statistic and sandwich-type SEs for model and parameter evaluation. However, with this approach, the inferences are still based on asymptotics. Furthermore, with typically non-normally distributed data (Micceri, 1989), minimizing $F_{mla}(\boldsymbol{\theta})$ may not generate most efficient parameter estimates of $\boldsymbol{\theta}$, because the function does not account for possible skewness and kurtosis in the sample.

In item factor analysis with ordinal data, the polychoric correlation matrix $\mathbf{R}$ tends to be nonpositive definite because different elements are computed using different pairs of variables. One method of item factor analysis is to treat $\mathbf{R}$ as a covariance matrix in the NML based method, and constrain all the diagonal

elements of the structural model implied covariance matrix Σ at 1.0. This method has been recommended by Lei (2009) for SEM analysis when the sample size N is not large enough. But it fails when **R** is not positive definite. Yuan et al. (2011) generalized the ridge ML method in Yuan and Chan (2008) for continuous data to ordinal data by fitting a structural model to $\mathbf{R}_a = \mathbf{R} + a\mathbf{I}$ via minimizing NML discrepancy function F_{mla}. Empirical results indicated that ridge ML for ordinal data yields much more efficient parameter estimates than ML for ordinal data when modeling polychoric correlations. While Yuan et al. (2017) did not compare ridge ML directly with the widely used methods of least squares (LS) and diagonally weighted least squares (DWLS), we would expect ridge ML to yield more efficient parameter estimates than LS and DWLS, especially when N is small and p is relatively large. However, like for NML with continuous data, the sampling property of **R** is not accounted for in the stage of parameter estimation with the ridge ML method.

3.2. GLS and Ridge GLS

Arruda and Bentler (2017) studied the normal-distributed-based generalized-least-squares (GLS) method, and proposed to replace the sample covariance matrix in the GLS weight matrix by a regularized covariance matrix. While their development is to improve the conditions of the weight matrix, they mostly focused on the performance of GLS-type test statistics for overall model evaluation with normally distributed data. We would expect that this regularized GLS method would yield more efficient parameter estimates than NML or GLS with large p. Further study is needed.

A well-known development that considers the sampling variability of the sample covariance matrix **S** is the generalized-least-squares (GLS) method developed by Browne (1984), which is also called the asymptotically distribution free (ADF) method. Let **s** be the vector of non-duplicated elements in **S**, $\sigma(\theta)$ be the structured counterpart of **s**, and $\hat{\Gamma}$ be a consistent estimator of the covariance matrix of $N^{1/2}\mathbf{s}$. The GLS/ADF discrepancy function is given by

$$F_{gls}(\theta) = (\mathbf{s} - \sigma(\theta))'\mathbf{W}(\mathbf{s} - \sigma(\theta)), \qquad (2)$$

where $\mathbf{W} = \hat{\Gamma}^{-1}$, with $\hat{\Gamma}$ being the sample 4th-order moment matrix (Mooijaart and Bentler, 2011). The GLS/ADF parameter estimates are obtained by minimizing the function $F_{gls}(\theta)$ in equation (2). While the GLS/ADF method enjoys the property of yielding asymptotically most efficient estimator among all methods of modeling **S**, its performance is rather poor unless the sample size N is rather large and p is relatively small. Further studies show that this is because $\hat{\Gamma}$ is very unstable, especially when $\hat{\Gamma}$ is near singular with a large p and a relatively small N (Yuan and Bentler, 1998; Huang and Bentler, 2015).

Note that the function $F_{gls}(\theta)$ in equation (2) becomes the LS discrepancy function when **W** is replaced by the identity matrix **I**. In contrast to the instability of $\hat{\Gamma}$, the weight matrix **I** in the LS method is most stable because it does not depend on data. However, there is no mechanism in the LS method to account for the variances of the elements in **S**, and consequently the LS estimator does not possess the desired asymptotic properties

of the GLS/ADF estimator. Considering the pros and cons of LS and ADF method, Yuan and Chan (2016) proposed a ridge GLS method in which the weight matrix **W** in Equation (2) is replaced by $\mathbf{W}_a = [a\hat{\Gamma} + (1-a)\mathbf{I}]^{-1}$, where a is a scalar that can be tuned according to certain conditions. Clearly, GLS/ADF corresponds to $a = 1$ while LS corresponds to $a = 0$. We can choose an a in between to yield the most efficient parameter estimates. Results in Yuan and Chan (2016) indicate that, with typically non-normally distributed data in practice, ridge GLS yields uniformly more efficient parameter estimates than LS, GLS/ADF, and NML. While ridge GLS enjoys various advantages over other well-known methods, it involves the determination of the tuning parameter a. Currently, there does not exist an effective method for choosing a to yield most efficient parameter estimates. Also, as we will discuss in the next section, the test statistics following ridge methods also need to be calibrated, especially with large p and small N.

3.3. The Bayesian Method

In addition to ML and GLS, another major method for parameter estimation is the Bayesian approach. In particular, for some nonlinear models (e.g., models involving interactions among latent variables), the likelihood function of the model parameters might be hard to specify or becomes too complicated to work with. It is relatively easy to specify a conditional distribution of the parameters via data augmentation (Gelman et al., 2014). Thus, one might be able to obtain results close to those by ML or GLS using the Bayesian approach that is facilitated by Gibbs sampling or Markov chain Monte Carlo (MCMC). Another advantage of the Bayesian approach is that one can include prior information by properly specifying prior distributions for the model parameters. However, properly specifying prior distributions needs skills, especially when the prior information does not come in the form of an inverted Wishart distribution or inverted gamma distribution that are needed in most developments of Bayesian methodology (Scheines et al., 1999; Lee, 2007), and inaccurate specification of the prior distributions can result in biased estimates (Baldwin and Fellingham, 2013; Depaoli, 2014; McNeish D., 2016). While the covariance matrix $\mathbf{S}_a = \mathbf{S} + a\mathbf{I}$ in ridge ML can also be regarded as a Bayesian estimate by specifying a prior distribution for the saturated covariance matrix, the effect of a is removed from the estimates of error variances (Yuan and Chan, 2008). This is why ridge ML yields more accurate parameter estimates than NML even when data are normally distributed.

Unlike the methods of ML and GLS that are justified by asymptotics, the modern Bayesian approach to estimation and inference is based on sampling from the posterior distribution. Thus, the Bayesian method has been suggested to deal with issues for for small N. Indeed, the MCMC method has been shown to outperform ML and GLS in small sample contexts (Lee and Song, 2004; Zhang et al., 2007; Moore et al., 2015; van de Schoot et al., 2015; McNeish D. M., 2016). However, informative priors are used in these studies. In particular, when N is small, the priors are expected to dominate the results. Actually, one may get satisfactory results even when $N = 0$ if accurate priors are specified. Thus, to a certain degree, the small

sample advantage of a Bayesian method is subjective. With flat or Jeffreys noninformative priors, Bayesian methods are expected to yield equivalent results to ML. However, empirical results do not endorse the Bayesian methods with small N (Baldwin and Fellingham, 2013; McNeish D. M., 2016).

In conclusion, when prior information is indeed available and one can include it in the current study via an accurate specification of priors and the distribution of current data given the parameter, Bayesian method is preferred and will be able to successfully address the problem for SEM with small N and/or large p. One needs to be cautions when either of the specifications is not proper. In particular, Bayesian methods "will not address the small sample issue and that ML with small sample alterations typically produce estimates with quality that can equal or often surpass MCMC methods that do not carefully consider the prior distributions" (McNeish D. M., 2016, p. 753).

3.4. Penalized Likelihood

The issue of large p and small N also poses challenges to other conventional statistical methods. In the literature of regression analysis, lasso methodology has been shown to be a viable method with big data or when p is rather large but N is not sufficiently large (Tibshirani, 1996). As a generalization of ridge regression, the idea of lasso is to squeeze parameters with small values so that they are equivalent to being removed from the model. When there are too many predictors to consider and when their relevance is unclear, lasso regression provides a viable tool for conducting regression analysis and variable selection at the same time. Lasso methodology has also been generalized to SEM via penalized likelihood (Jacobucci et al., 2016; Huang et al., 2017). However, although there can be many items with survey data, SEM is generally conceptualized as a confirmatory methodology. In particular, it is expected that both the measurement and the structural parts of an SEM model be established in priori, and while the model parameters are freely estimated they are not free to be removed, simply because the resulting model might correspond to a completely different theoretical hypothesis. Also, the size of a parameter in SEM is scale dependent. The estimates with smallest values might be statistically most significant.

Parallel to factor rotation, lasso methodology might be a more useful technique for exploratory factor analysis when applied to standardized variables, because the scales of the measured variables become irrelevant. While the corresponding parameter estimates with standardized variables are more comparable, it does not imply that their standard errors also are comparable or become irrelevant (Cudeck, 1989). One still needs to be cautious when using the lasso methodology for big data in SEM or factor analysis, especially when items that are theoretically important for measuring an underlying construct might have smaller loadings.

In summary, various methods have been proposed in SEM to yield more accurate/efficient parameter estimates. Currently, the ridge ML and ridge GLS appear to be the most promising methods. In particular, when combining ridge ML with robust transformation (Yuan et al., 2000), ridge ML may in fact yield estimates that are close to full information maximum likelihood estimates. Further study in this direction is clearly needed.

4. TEST STATISTICS

In addition to making parameter estimation difficult, a large p also causes problems to the overall model evaluation, which is considered by many researchers a key aspect of SEM (e.g., Marcoulides and Yuan, 2017). Model evaluation has also gained more extensive studies than other aspects of SEM, simply because any elaboration on parameter estimates or causal relationship among the variables is conditional upon determination of an adequate model. Because there are many extensive developments in this direction that have appeared in the literature, here we only discuss the pros and cons of methods connected to the issue of the effect of small N and/or large p. Most of these studies on advancing model inference in SEM within this context follow two directions. One is to account for non-normally distributed data; and the other is to account for small sample sizes or large number of variables. Of course, because statistics that account for non-normally distributed data also face the challenge of large number of variables, we inevitably also focus our review on their behaviors with small N and/or large p.

4.1. Correction to T_{ml} Under the Normality Assumption

The most widely used test statistic in SEM is $T_{ml} = (N - 1)F_{ml}$, mostly because it is the default statistic in available software, not because normally distributed data are common or T_{ml} provides more reliable model evaluation. Under the normality assumption and a correct model, T_{ml} approaches the nominal chi-square distribution χ^2_{df} as the sample size N increases while p is fixed. However, this result does not tell us how large N needs to be at a given value of p for T_{ml} to approximately follow χ^2_{df}. As we noted earlier, the statistic T_{ml} can reject the correct model 100% at the nominal level of 5% even when data are normally distributed. While there are a lot of efforts to improve the performance of T_{ml} by many authors, most are ad hoc corrections rather than principled ones. Consequently, the behavior of the corrected statistics can vary as conditions change.

It is a general and well-known phenomenon that the likelihood ratio statistic tends to reject the correct model more often than expected when N is not sufficiently large, not just in SEM. As a consequence, statisticians have developed a systematic approach for correcting the likelihood ratio statistic, and it is called the Bartlett correction (Bartlett, 1937, 1954; Box, 1949; Lawley, 1956). Wakaki et al. (1990) obtained the Bartlett correction to T_{ml} for a class of covariance structural models with normally distributed data. However, the corrected statistic is rather complicated even for a relatively simple model, and it is impractical to implement the Bartlett correction on T_{ml} for general SEM models. In the context of exploratory factor analysis (EFA), Bartlett (1951, see also Lawley and Maxwell, 1971, p. 36) proposed a simplified formula to correct the likelihood ratio statistic, which is to replace $(N - 1)$ in T_{ml} with $N_b = N - (2p + 11)/6 - 2m/3$, where m is the number of factors. Because the

corrected statistic, $T_{mlb} = N_b F_{ml}$, is easy to implement, Nevitt and Hancock (2004, see also Fouladi, 2000) proposed to apply the simple correction to confirmatory factor analysis (CFA) and SEM, where m is still the number of latent factors. However, studies by Nevitt and Hancock (2004) and Herzog et al. (2007) indicate that type I errors with $T_{mlb} \sim \chi^2_{df}$ in SEM tend to be much lower than the nominal level.

Considering that the number of free parameters in SEM is much smaller than that in EFA when m is large, Yuan (2005) proposed to replace $(N-1)$ in the definition of T_{ml} with $N_y = N - (2p + 13)/6 - m/3$. However, this proposal is only a heuristic rather than one that is statistically justified. A more complicated correction was originally offered by Swain (1975), who proposed to replace $(N-1)$ in T_{ml} by

$$N_s = N - 1 - [p(2p^2 + 3p - 1) - h_q(2h_q^2 + 3h_q - 1)]/(12df),$$

where $h_q = [(1 + 8q)^{1/2} - 1]/2$ and q is the number of free parameters in the structural model. Studies by Fouladi (2000), Herzog et al. (2007) and Herzog and Boomsma (2009) indicate that the performance of test from best to worst are $T_{mls} = N_s F_{ml}$, $T_{mly} = N_y F_{ml}$, and T_{mlb}. Although the performance of T_{mls} is potentially promising, the correction is not statistically justified.

Parallel to the Bartlett correction, Yuan et al. (2015) developed a procedure that involved an empirical correction. In particular, they proposed to estimate the coefficient $\boldsymbol{\beta}$ in

$$T_{mle} = (N - \mathbf{c}'\boldsymbol{\beta})F_{ml}$$

by matching the empirical mean of T_{mle} with the nominal degrees of freedom or the mean of χ^2_{df}, where $\mathbf{c}$ is a vector whose elements are different combinations of p, q, and m. Using Monte Carlo results across 342 conditions of N, p, q, and m, they estimated the vector $\boldsymbol{\beta}$ by maximum likelihood. One of the statistics they recommended is $T_{mle} = N_e F_{ml}$, where $N_e = N - (2.381 + 0.361p + 0.003q)$. Yuan et al. (2015) noted that T_{mle} can be properly used when $N > \max(50, 2p)$. Recently, Shi et al. (2018) conducted a rather comprehensive simulation study and showed that T_{mle} performed better than T_{ml}, T_{mlb}, and T_{mls}. However, type I error rates of T_{mle} can still be inflated when p is extremely large (e.g., $p = 90$), even when $N = 200$. They noted that for normally distributed data N needs to be $>4p$ in order for T_{mle} to properly control type I errors if p is over 100.

4.2. Corrections to Test Statistics That Account for Non-normality

In addition to correcting T_{ml} for normally distributed data with small N and/or large p, various developments on test statistics with non-normally distributed data were also made, including modifying the statistic T_{ml} following NML as well as working with GLS, ridge GLS or a robust estimation method. We discuss next their properties with small N as well as those of their modified versions.

The most widely used statistics that account for non-normality are the rescaled statistic T_{rml} and the adjusted statistic T_{aml}, developed by Satorra and Bentler (1994). The statistic T_{rml}

has the property that its mean asymptotically equals that of the nominal chi-square distribution χ^2_{df}, and the statistic T_{aml} has the property that both its mean and variance asymptotically equal those of the approximating chi-square distribution $\chi^2_{df_a}$. Note that the value of df_a (the degrees of freedom) for the reference distribution of T_{aml} is determined by both the model and the underlying population distribution, and needs to be estimated in practice. Among the two statistics, T_{rml} is more widely used and is called a robust chi-square statistic by some authors (Bentler and Yuan, 1999). Although the exact distribution of neither T_{rml} nor T_{aml} is known even asymptotically, Monte Carlo results in Fouladi (2000) at $p = 6$ and 12, and in Hu et al. (1992) at $p = 15$ indicate that they perform reasonably well for medium to large N. However, there exists evidence that T_{rml} and T_{aml} do not work well with small N (Bentler and Yuan, 1999; Nevitt and Hancock, 2004). In particular, when p is relatively large, results in Yuan et al. (2017) suggested that T_{rml} can reject the correct model from 0 to 100% while the nominal rate is 5%.

Earlier studies indicated that T_{rml} tend to over-reject the correct model when N is not sufficiently large (e.g., Hu et al., 1992; Bentler and Yuan, 1999). Many *ad-hoc* corrections have been proposed to correct such behavior, including $T^{(b)}_{rml}$ obtained by replacing the $(N-1)$ in the formulation of T_{rml} with N_b, which is the formula proposed by Bartlett (1951) for correcting the behavior of T_{ml} in EFA. Aiming to improve the behavior of T_{rml} in over-rejecting corrected models at smaller N, Jiang and Yuan (2017) proposed four statistics to further modify the behavior of T_{rml}. However, these statistics may reject the correct model 0 times in some conditions. Yang et al. (2018) studied 10 modifications of T_{rml}, including $T^{(b)}_{rml}$, and the four proposed in Jiang and Yuan (2017). Using the average of the absolute deviations from the nominal level as a criterion, results in Yang et al. (2018) indicate that $T^{(b)}_{rml}$ performed the best across 604 conditions of N, p, and different population distributions. But still $T^{(b)}_{rml}$ rejected correct models from 0 to 96%. In particular, when N is small and p is relatively large, the rejection rate by $T^{(b)}_{rml}$ for correct models is close to 100% with normally distributed data, and 0% when data follow a population distribution with heavy tails. Thus, $T^{(b)}_{rml}$ is not a reliable test statistic for SEM when p is large and/or N is relatively small.

Comparing to T_{rml}, fewer studies for the adjusted statistic T_{aml} indicate that it tends to under-reject correct models, and perform rather well when p is relatively small (Fouladi, 2000). However, there is a noticeable lack of studies focusing on T_{aml} with respect to type I error control at relatively large p. Nevitt and Hancock (2004) noted that the resulting statistic of replacing the $(N-1)$ in the formulation of T_{aml} by N_b (Bartlett's formula for EFA) does not work well.

Note that T_{rml} and T_{aml} of Satorra and Bentler are derived from the principles of mean and mean-and-variance correction (e.g., Welch, 1938; Rao and Scott, 1984), and they are expected to work well in practice, especially when p or the degrees of freedom are large (e.g., Yuan and Bentler, 2010). Yang et al. (2018) recently examined the causes for T_{rml} and T_{aml} to fail to control type I errors, and found that neither T_{aml} nor T_{rml}

possess the properties to which they are entitled asymptotically. That is, their means and variances can be far from those of their reference distributions. Even for normally distributed data, the mean of T_{rml} can be hundreds of times greater than that of the nominal chi-square distribution in standardized units when p is large but N is not sufficiently large. For non-normally distributed data, the mean of T_{rml} can also be much smaller than that of the nominal chi-square distribution when both p and N are large. Also, there are many conditions under which the mean of T_{aml} is much greater than that of its reference chi-square distribution $\chi^2_{df_a}$ whereas the standard deviation of T_{aml} is much smaller than that of $\chi^2_{df_a}$. This is because the mean and mean-and-variance corrections for obtaining T_{rml} and T_{aml} are implemented via standard asymptotics, which fail when p is relatively large. Yang et al. (2018) noted that mean and mean-and-variance corrections are still expected to work well with big data but we have to use alternative methods instead of those based on asymptotics to implement them.

A recent development in correcting the behavior of T_{rml} is given by Yuan et al. (2017). Parallel to obtaining T_{mle}, they replaced the term $(N-1)$ in the formulation of T_{rml} with a scalar $N_c = N - \mathbf{c}'\boldsymbol{\beta}$. However, the elements of the vector $\mathbf{c}$ contain covariates that reflect the underlying population distribution of the sample in addition to various nonlinear functions of N, p, q, and df. The coefficients in $\boldsymbol{\beta}$ are estimated so that the corrected statistic $T_{rml}^{(c)} = N_c F_{rml}$ has a mean approximately equal to that of the nominal chi-square distribution, not according to asymptotics but according to empirical results. With many conditions of N, p, m, q, and population distribution, they evaluated different formulations of N_c, and recommended a statistic containing 20 predictors, denoted it as $T_{rml}^{(c20)}$. They further conducted an independent simulation study to evaluate $T_{rml}^{(c20)}$, and found that it performed substantially better than $T_{rml}^{(b)}$. In particular, for normally distributed data with p ranging from 20 to 80, the rejection rates of $T_{rml}^{(c20)}$ range from 2.4 to 7.6%, the rejection rates of $T_{rml}^{(b)}$ range from 2.2 to 57.8%, and those of T_{rml} range from 4.8 to 100%. For data that follow elliptical distributions, the rejection rates of $T_{rml}^{(c20)}$ range from 5.8 to 14%, those of $T_{rml}^{(b)}$ range from 0 to 5.4%, and those of T_{rml} range from 0 to 95.4%. We may think that $T_{rml}^{(b)}$ performed well for the condition of elliptical population distributions, however, its rejection is 0% for many of the conditions studied, and type I error rates do not tell how bad its performance is for these conditions once the rates are equal to zero.

The statistics we discussed so far, as listed in the first part of **Table 1**, are all derived from the normal-distribution-based maximum likelihood (NML), and the parameter estimates under these statistics are the same. In particular, unless data are normally distributed, NML does not account for the underlying population distribution in estimating model parameters. As we noted in the previous section, the GLS/ADF method uses the inverse of a consistent estimator of the covariance matrix of $N^{1/2}\mathbf{s}$ as the weight matrix, which directly accounts for the underlying population distribution. In addition to yielding asymptotically most efficient parameter estimates among all the methods of modeling $\mathbf{S}$, the corresponding statistic $T_{gls} = (N-1)F_{gls}$ also

TABLE 1 | Test statistics and their applicability.

Statistic	Source	Applicability		
		p	N	Distribution
T_{ml}	Likelihood ratio	S	L	NM
T_{mlb}	Nevitt and Hancock, 2004	S	L	NM
T_{mly}	Yuan, 2005	S	L	NM
T_{mls}	Swain, 1975	S to M	M to L	NM
$\underline{T_{mle}}$	Yuan et al., 2015	S to M	M to L	NM
T_{rml}	Satorra and Bentler, 1994	S	L	NM & NNM
T_{aml}	Satorra and Bentler, 1994	S	L	NM & NNM
$T_{rml}^{(b)}$	Nevitt and Hancock, 2004	S	L	NM
$\underline{T_{rml}^{(c20)}}$	Yuan et al., 2017	S to M	M to L	NM & NNM
T_{gls}	Browne, 1984	S	vL	NM & NNM
T_{cgls}	Yuan and Bentler, 1997a	S	M to L	NM & NNM
T_F	Yuan and Bentler, 1999	S	M to L	NM & NNM

S, small; M, medium; L, large; vL, very large; NM, normal; NNM, nonnormal
The two underlined statistics are recommended.

asymptotically follows the nominal chi-square distribution. With $p = 15$ manifest variables, results in Hu et al. (1992) suggested that T_{gls} performs as expected at $N = 5,000$, and it rejects the correct model 100% at $N = 150$. Thus, as p increases, the requirement for sample size by $T_{gls} \sim \chi^2_{df}$ to perform reasonably well is even more demanding than $T_{rml} \sim \chi^2_{df}$ or $T_{aml} \sim \chi^2_{df}$.

Statistics that are less demanding for sample size following the GLS/ADF estimation have also been developed over the years. One is a corrected GLS/ADF statistic $T_{cgls} = T_{gls}/(1+F_{gls})$, which was obtained by Yuan and Bentler (1997a) via estimating the covariance matrix of $N^{1/2}\mathbf{s}$ using the cross-product of residuals instead of using 4th-order sample moments. Another is an F-statistic (Yuan and Bentler, 1999)

$$T_F = \frac{(N-df)F_{gls}}{df}, \tag{3}$$

which is referred to an F-distribution with df and $N - df$ degrees of freedom. In addition, Yuan and Bentler (1998) also proposed statistics that are based on residuals $\mathbf{s} - \hat{\boldsymbol{\sigma}}$, where the parameter estimates $\hat{\boldsymbol{\theta}}$ in $\hat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}(\hat{\boldsymbol{\theta}})$ might be obtained by NML or least squares. In addition to being asymptotically distribution-free, these statistics perform reasonably well for medium to large sample sizes (Fouladi, 2000; Nevitt and Hancock, 2004). However, when N is relatively small, T_F tends to over-reject correct models and T_{cgls} tends to under-reject correct models (Bentler and Yuan, 1999). Also, we need to have $N > df$ for T_F and T_{cgls} to be properly defined, and a much larger N is needed for them to closely follow their nominal distributions. Because the value of df tends to increase fast with p, the statistics T_F and T_{cgls} are not solutions to inference issues for SEM when p is large.

4.3. Other GLS-Type Test Statistics

As described in the previous section, Arruda and Bentler (2017) studied a regularized normal-distribution-based GLS method and found that one of the resulting test statistics performed rather well at $p = 15$ and $N = 60$ for normally distributed data. While

this statistic is expected to work better than its unregularized version as well as T_{ml} with larger p, it is clear that further study on its performance with large p and small N as well as non-normally-distributed data is needed.

Following the NML estimation method, a GLS-type statistic is obtained when the weight matrix is obtained by the estimated covariance matrix instead of the sample covariance matrix. This statistic is called "the normal theory RLS chi-square" statistic in EQS (Bentler, 2006). A ridge version of this statistic obtained by fitting the polychoric correlation matrix with ordinal data was studied in Yuan et al. (2011). They found that the rescaled and adjusted versions of the ridge RLS type statistics performed much better than the rescaled and adjusted versions of the ridge ML type statistics. Because the formulation of the ridge RLS type statistic is very close to the regularized GLS statistic studied in Arruda and Bentler (2017), in general we would expect that the rescaled and adjusted ridge RLS type test statistics to perform better than the rescaled and adjusted ridge ML statistics with large p. Of course, additional research is clearly needed in this direction before any definitive conclusion can be drawn. As pointed by Huang and Bentler (2015) and noted in the previous section, the poor performance of the GLS/ADF statistic T_{gls} is closely related to the condition of the weight matrix $\mathbf{W} = \hat{\Gamma}^{-1}$ in Equation (2), especially if $\hat{\Gamma}$ is near singular when N is small relative to p. Using the idea of principal components, Chun et al. (2017) proposed to test the structural model by ignoring the directions corresponding to the smallest eigenvalues of $\hat{\Gamma}$. The resulting statistic is shown to control type I errors better. But the null hypothesis of the test is different from that of other test statistics reviewed in so far, because the test is unable to identify misspecifications in the direction represented by the vectors corresponding to the smallest eigenvalues of $\hat{\Gamma}$.

In summary, much research has focused on trying to obtain more reliable test statistics for overall model evaluation in cases with large p and/or relatively small N. Most of the obtained statistics are justified by either asymptotics or simple *ad-hoc* corrections to T_{ml} or T_{rml}. Currently, the most reliable statistic for normally distributed data with many variables is the statistic T_{mle}, developed in Yuan et al. (2015). The most reliable statistic for data with many variables that are possibly non-normally distributed is the statistic $T_{rml}^{(c20)}$, developed in Yuan et al. (2017). These two statistics are underlined in **Table 1**.

While ridge ML, ridge GLS and robust methods have been shown to yield more efficient parameter estimates, there is little development aimed at improving the performance of the rescaled and adjusted test statistics following these methods (Yuan and Chan, 2008, 2016; Tong et al., 2014).

5. STANDARD ERRORS OF PARAMETER ESTIMATES

Standard errors (SEs) for parameter estimates are also key elements in SEM although they are secondary compared to parameter estimates or test statistics for overall model fit evaluation. In particular, once a model is deemed adequate following a proper estimation method, the meaning of the estimated values of the parameters must then be properly

elaborated and explained. It is in this context that, accurate SE estimates are essential for proper interpretations. However, compared to test statistics or parameter estimation, much less research has focused on ways to improve the estimation of SEs. While a formula for computing SEs is typically provided with each method of parameter estimation in SEM, the formula is mainly justified by asymptotics, and may not work well when the number of variables is large. We review existing approaches to estimating SEs in this section, and point out their potential problems with small N and/or large p.

Coupled with the test statistic T_{ml}, standard errors following the NML method in SEM are computed by inverting the corresponding information matrix, as is given in the default output of most SEM software. Such obtained SEs are consistent when the normality assumption literally holds and the model is correctly specified. When either the normality assumption is violated or when the model is misspecified, SEs based on the information matrix are not consistent (Yuan and Hayashi, 2006). We are not aware of any study to date on their accuracy for large p or small N. While there is a general interest in the performance of the SEs based on the information matrix, because data commonly collected in social and behavioral science are typically non-normally distributed (Micceri, 1989), additional effort on improving the information-matrix-based SEs may not ultimately be worth the investment.

The NML method is widely used in practice regardless of the distribution of the data. This is because there are not many multivariate distributions to choose from, and we typically do not know the population distributions. Following NML estimation, SEs based on the so-called sandwich-type covariance matrices have been proposed to account for violations of normality (White, 1981; Bentler, 1983; Shapiro, 1983; Browne, 1984), and they have been implemented in most SEM software. Such SEs are also called robust SEs in the SEM literature, parallel to the rescaled statistic T_{rml}. However, the sandwich-type SEs as implemented in statistical packages are based on the assumption of a correctly specified model, because the formula becomes rather complicated otherwise (Yuan and Hayashi, 2006). While it is unlikely that a researcher can specify a model that is literally correct in practice, sandwich-type SEs are close to being consistent if the model is deemed adequate. However, consistency does not tell us how good the SE estimates are in a given application. While there are limited studies on the performance of sandwich-type SEs in SEM (Yuan and Bentler, 1997b; Yuan and Chan, 2016), there is a great deal of evidence that sandwich-type SEs are not reliable when p is large and N is not sufficiently large in other contexts (MacKinnon and White, 1985; Long and Ervin, 2000; Yang and Yuan, 2016). In particular, for regression models with heteroscedastic variances, various corrections to SEs have been proposed (see e.g., Cribari-Neto, 2004), however, because these may not be directly generalizable to the SEM context, it is evident that further research is needed on this topic.

In addition to yielding a statistic T_{gls} that asymptotically follows the nominal chi-square distribution, the GLS/ADF method also generates a formula that yields consistent SEs for the GLS estimates. However, like T_{gls}, it needs a rather large sample size for the formula-based SEs to match those of the empirical ones. When p is large while N is not sufficiently large, the SEs

computed by GLS/ADF formula are too small. Yuan and Bentler (1997b) proposed a correction to the formula of the covariance matrix of the GLS/ADF estimator. While the corresponding corrected SEs are much improved over the uncorrected ones, but they are still under-estimated, especially when N is small. Further improvement over the corrected SEs is possible. But the GLS/ADF estimator can be rather inefficient for small N, and the additional effort needed to improve the estimates of SEs for not efficient parameter estimates may not be worthwhile.

The bootstrap method has also been shown to yield reliable SEs, and is especially valuable when formula-based SEs are not available (Efron and Tibshirani, 1993). In particular, a model does not need to be literally correct in order for the bootstrap method to yield consistent SEs (Yuan and Hayashi, 2006). Indeed, since the bootstrap methodology is based on resampling that accounts for both the sample size and empirical distribution, we would expect it to work reliably regardless of the values of N and p. Currently, however, we are not aware of any study that verifies the validity of the bootstrap methodology for SEM with large p and small N.

In summary, few studies have focused on improving estimates of standard errors in SEM with large p. Although the bootstrap methodology appears promising, it is not a substitute for analytical formulas. This is because the bootstrap methodology is essentially Monte Carlo simulation with empirical data. It takes time to estimate the parameters in conducting the simulation with SEM models, especially when p is large. The issue of non-convergence with parameter estimation discussed earlier can be a serious problem for the bootstrap methodology since there exist systematic differences between converged and non-converged replications (Yuan and Hayashi, 2003), and SEs based on only the converged replications might under-estimate the true SEs. Since efficient parameter estimates are fundamental to statistical inference, future research should focus on developing more reliable SEs perhaps by focusing on the development of methods (such as ridge GLS and robust methods) that yield more efficient parameter estimates.

6. A REAL DATA EXAMPLE

In this section we present an empirical data example with a small N and a relatively large p. As discussed in the previous sections, small N and large p can cause many problems in estimating and evaluating SEM models. In the illustration we focus specifically on model evaluation with different test statistics, which is a key step for SEM analyses to provide reliable results. The data come from an intervention program for college students who had exhibited depression symptoms. Measurements for both pre- and post-interventions are obtained on $N = 57$ participants, all of whom are college students from universities located in Beijing, China. The data were collected by the first author, as part of a study examining the relationship between resilience and depression.

Resilience was measured by the Connor-Davidson Resilience Scale (CDRISC, Connor and Davidson, 2003). The CD-RISC contains 25 items, with each item rated on a 5-point scale

reflecting how a participant felt over the past month, where $1 =$ Not true at all, $2 =$ Rarely true, $3 =$ Sometimes true, $4 =$ Often true, $5 =$ True nearly all of the time. The CD-RISC has 3 subscales: toughness (13 items), powerful (8 items), and optimistic (4 items). The model of resilience and depression is formulated with the subscales, not the item scores.

For each participant, measures of depression using the Self-rating Depression Scale (SDS) were also collected. The SDS contains 20 items (Song and Liu, 2013; Zhang et al., 2015; Xu and Li, 2017), with each item rated on a 4-point scale according to how a participant has felt over the past week, where $1 =$ A little of the time, $2 =$ Some of the time, $3 =$ Good part of the time, $4 =$ Most of the time. Item 2, 5, 6, 11, 12, 14, 16, 17, 18, and 20 were reverse scored items. A higher score on the scale reflects higher levels of depression.

The illustrative data also included six items from "Forgiveness of Others subscale," which is part of the Heartland Forgiveness Scale (HFS, Thompson et al., 2005). All of the responses are on 7-point Likert scale, with 1–7 signifying responses from "Almost Always False of Me" to "Almost Always True of Me." Items 1, 3, and 5 were reverse scored items. A higher score indicates more willing to forgive others. While the item-level scores are ordinal variables, for purposes of the illustration, we treat them as continuous variables in the analysis, which leads to very little bias according to Li (2016) and Rhemtulla et al. (2012).

Each participant did the pre-test by filling out the questionnaire before the intervention started, and a post-test 3-months after the group intervention. Thus, we have $p = 20$ variables in total, 10 for the pre-test and 10 for the post-test.

Past literature on depression has indicated that a higher level of resilience generally corresponds to a lower level of depression (Kim and Yoo, 2004; Ding et al., 2017; Poole et al., 2017). This literature has also suggested that individuals with depression are expected to be mostly victims of negative events, and that forgiving others is positively correlated with resilience (Dai et al., 2016; Saffarinia et al., 2016). Additionally, it has been determined that a person having a high level of forgiveness is more likely to have a higher level of satisfaction with life, and thus is relatively less depressed (Yu and Zheng, 2008). In accordance with these past research findings, we hypothesize that forgiveness would play a mediating role. **Figure 1** is a hypothetical model for exploring the relationship between resilience, depression, and forgiveness of others, where for ease of presentation prediction and measurement errors are not included in the path diagram. We hypothesize that forgiving-others has a mediating effect between psychological resilience and depression. That is, resilience can predict the depression directly, and also can predict depression through forgiving-others. It is also hypothesized that resilience, depression, and forgiveness at time 1 (T1) will have a lasting effect after the group intervention (T2). In addition, the level of resilience at T1 also influences the level of depression and forgiving-others at T2, and the level of forgiving-others at T1 influences the level of depression at T2 as well.

The first line of **Table 2** contains the results of the test statistics T_{ml}, T_{rml} and $T_{rml}^{(c20)}$ for Model 1, following NML. With $df = 160$, $T_{ml} = 243.97$ and $T_{rml} = 242.60$ are noticeably statistically significant when compared to χ^2_{160}. In contrast, $T_{rml}^{(c20)} = 191.68$,
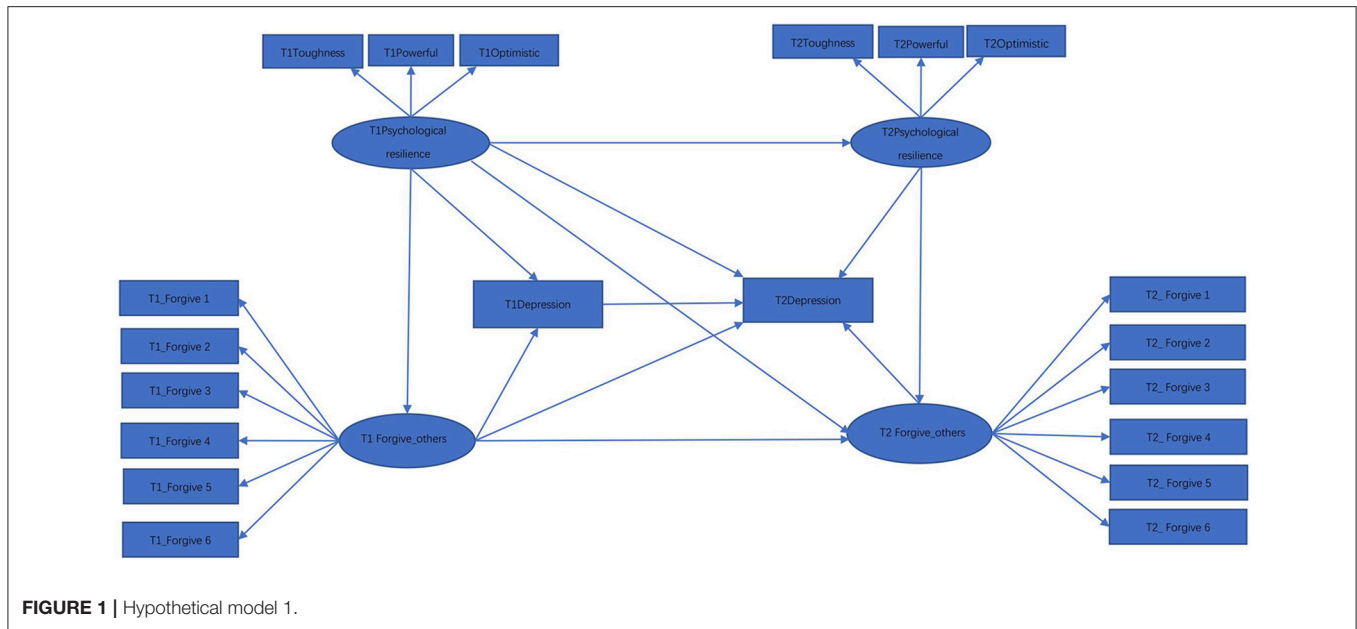
FIGURE 1 | Hypothetical model 1.

TABLE 2 | Test statistics T_{ml}, T_{rml} and $T_{rml}^{(c20)}$ for models 1–4.

Model	df	T_{ml}	p_{ml}	T_{rml}	p_{rml}	$T_{rml}^{(c20)}$	$p_{rml}^{(c20)}$
1	160	243.97	2.12×10^{-5}	242.60	2.72×10^{-5}	191.68	0.044
2	160	244.29	2.00×10^{-5}	240.38	4.05×10^{-5}	190.10	0.052
3	159	225.95	3.84×10^{-4}	225.33	4.24×10^{-4}	177.86	0.146
4	159	226.67	3.42×10^{-4}	224.34	4.96×10^{-4}	177.20	0.154

with a corresponding $p = 0.044$, would suggest that Model 1 is close to being adequate.

While many researchers have showed that psychological resilience can predict depression, numerous studies have also indicated that resilience is influenced by depression (Li et al., 2016; Song et al., 2017; Wang et al., 2017). In particular, these studies have found that resilience of depressed individuals tends to be significantly lower than that exhibited in healthy individuals. Such a hypothetical relationship is represented by **Figure 2**, where the relationship for the variables between the two time points are set to be identical to those in **Figure 1**. Test statistics for the model represented by **Figure 2** are in the 2nd line of **Table 2**. While both $T_{ml} = 244.29$ and $T_{rml} = 240.38$ reject the hypothetical model with $p < 0.001$, $T_{rml}^{(c20)} = 190.10$ suggests that Model 2 is not statistically significant at the nominal level 0.05.

It is important to note that our data are repeated measures, and it is known that such data involving the same measurement across time are likely to have additional correlation due to sharing specific traits. Accordingly, we used a model modification technique to identify the presence of such correlations. Following results obtained by model modification based on the score tests (Sarris et al., 1987), the two errors for indicators Forgive 4 were allowed to correlate in **Figure 1**, which yielded Model 3. In parallel, allowing the two errors of Forgive 4 in **Figure 2** to correlate yielded Model 4. The obtained results for Models

3 and 4 are displayed in the 3rd and 4th lines of **Table 2**, respectively. For each model, all the three test statistics become less significant. However, the conventional test statistics T_{ml} and T_{rml} still reject both models at the level of 0.001. In contrast, the p-value corresponding to $T_{rml}^{(c20)}$ is 0.146 for Model 3 and 0.154 for Model 4, indicating that both models fit the data reasonably well.

In summary, the considered example clearly showed that the most widely used test statistics are no longer reliable for data with large p and/or small N. By considering the conditions of N, p, and the empirical distribution, the empirically corrected statistic $T_{rml}^{(c20)}$ gives the models the appropriate credit they deserve.

7. DISCUSSION AND CONCLUSIONS

Data with a small sample size and many variables pose numerous challenges to conventional statistical methods. In this article, we reviewed the developments in SEM for dealing with the issues created by small N and/or large p. While there are many methods for parameter estimation and overall model evaluation in SEM, only a few can successfully account for the effect of large p. Although ridge ML and ridge GLS explicitly accounted for the effect of small N, their corresponding test statistics may not follow the nominal chi-square distribution. Among procedures of modeling the sample covariance matrix **S**, the test statistic $T_{rml}^{(c20)}$ has the mechanism to account for small N and the shape of the distribution of the sample. However, it is based on the NML method (which does not have the mechanism to account for small N), and may encounter estimation difficulties in practice when the sample covariance matrix is near singular. There is no doubt that additional developments are needed that focus on test statistics following the ridge estimation methods, along with formulas that can yield accurate SEs.

FIGURE 2 | Hypothetical model 2.
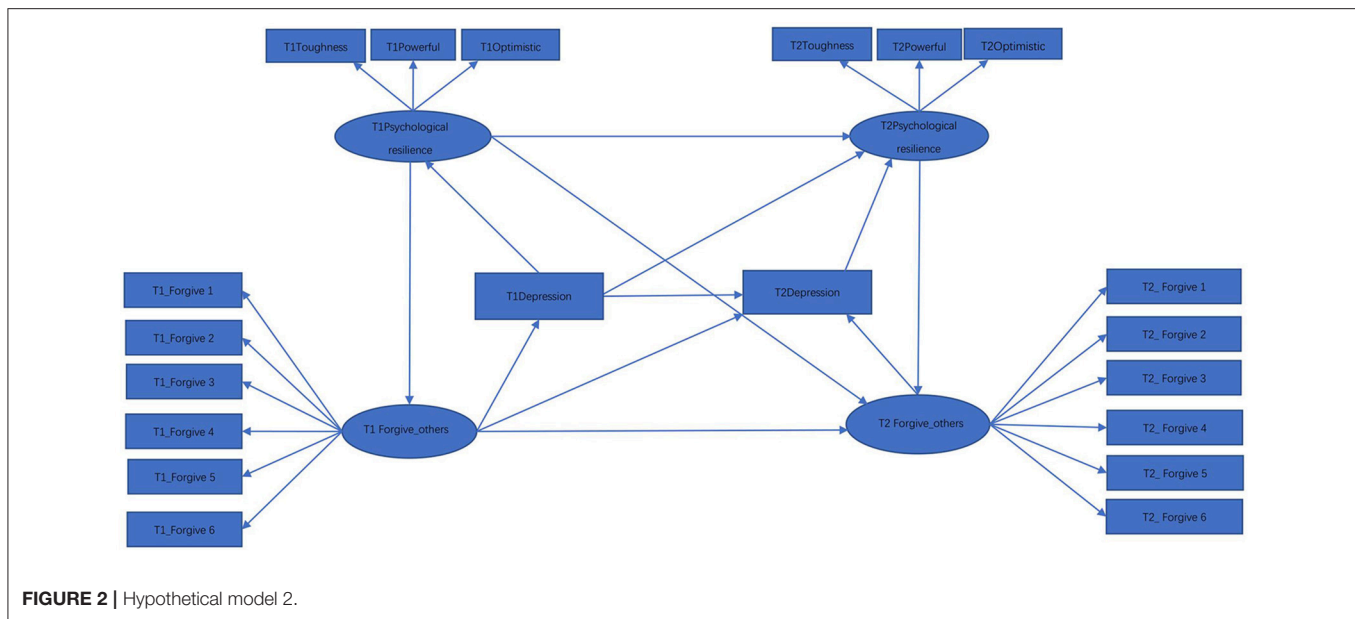
In this article, we mainly reviewed SEM methods based on modeling the sample covariance matrix **S**. When data are non-normally distributed, **S** is not an efficient estimate of Σ and the corresponding estimates for the structural parameters are not efficient either. Robust methods for SEM based on robust estimates of Σ have been developed (Yuan et al., 2004; Yuan and Zhong, 2008), and these methods are expected to yield more efficient parameter estimates than NML. However, additional developments with the robust methods are needed to deal with the issues of non-convergence and statistics not following the nominal chi-square distributions. Formulas to yield more reliable SEs of the robust estimates also need to be developed. In this article we did not describe methods for dealing with incomplete data, since existing methods for SEM with missing data do not have the mechanism to account for the effect of large p and/or small N yet (Yuan and Bentler, 2000; Savalei, 2010; Yuan and Zhang, 2012). Additional developments would appear to be needed for SEM with a large number of variables that contain missing data.

In addition to test statistics, fit indices are regularly used for overall model fit evaluation in applications of SEM. Since most popular fit indices are defined via test statistics (e.g., RMSEA, Steiger and Lind, 1980; CFI, Bentler, 1990), they too face the same issues with large p and small N (e.g., Jackson, 2001). Root mean squared residual (RMSR) is not defined via test statistics, but it needs a proper regulation to be an unbiased estimator of its population counterpart, and the construction of such an unbiased estimator can be especially challenging with small N and/or large p (Maydeu-Olivares, 2017). Some recent but limited

results have shown that the RMSEA and CFI defined via a statistic T_{mle} does perform much better than their counterparts defined using T_{ml} (Xing and Yuan, 2017). Based on this work, it is reasonable to expect that advances in test statistics will also improve the performances of other fit indices that are defined via these statistics.

In summary, data with large p and small N pose a big challenge for SEM methodology and many more new developments are still needed to tackle these issues, especially when the data are incomplete and/or non-normal.

AUTHOR CONTRIBUTIONS

LD: lead the writing and finalized the article; MY: contributed to computation of the example and the writing of the article; KM: contribute substantially to the writing of the whole article.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

Arrindell, W. A., and van der Ende. J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Appl. Psychol. Meas.* 9, 165–178. doi: 10.1177/014662168500900205

Arruda, E. H., and Bentler, P. M. (2017). A regularized GLS for structural equation modeling. *Struct. Equat. Model.* 24, 657–665. doi: 10.1080/10705511.2017.1318392

Baldwin, S. A., and Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex

variance structure. *Psychol. Methods* 18 151–164. doi: 10.1037/a00 30642

Barrett, P. T., and Kline. P. (1981). The observation to variable ratio in factor analysis. *Pers. Study Group Behav.* 1, 23–33.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. R. Soc. Lond. A* 160, 268–282. doi: 10.1098/rspa.1937.0109

Bartlett, M. S. (1951). The effect of standardization on a χ^2 approximation in factor analysis. *Biometrika* 38, 337–344. doi: 10.2307/2332580

Bartlett, M. S. (1954). A note on the multiplying factors in various χ^2 approximations. *J. R. Stat. Soc. B* 16, 296–298.

Bentler, P. M. (1983). Some contributions to efficient statistics in structural models: specification and estimation of moment structures. *Psychometrika* 48, 493–517. doi: 10.1007/BF02293875

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychol. Bull.* 107, 238–246. doi: 10.1037/0033-2909.107.2.238

Bentler, P. M. (2006). *EQS 6 Structural Equations Program Manual.* Encino, CA: Multivariate Software.

Bentler, P. M., and Chou, C.-P. (1987). Practical issues in structural modeling. *Sociol. Methods Res.* 16, 78–117. doi: 10.1177/0049124187016001004

Bentler, P. M., and Yuan, K.-H. (1999). Structural equation modeling with small samples: test statistics. *Multivariate Behav. Res.* 34, 181–197. doi: 10.1207/S15327906Mb340203

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika* 50, 229–242. doi: 10.1007/BF02294248

Box, G. E. P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika* 36, 317–346. doi: 10.1093/biomet/36.3-4.317

Browne, M. W. (1984). Asymptotic distribution-free methods for the analysis of covariance structures. *Br. J. Math. Stat. Psychol.* 37, 62–83. doi: 10.1111/j.2044-8317.1984.tb00789.x

Cattell, R. B. (1978). *The Scientific Use of Factor Analysis.* New York, NY: Plenum.

Chun, S. Y., Browne, M. W., and Shapiro, A. (2017). Modified distribution-free goodness-of-fit test statistic. *Psychometrika* 83, 48–66. doi: 10.1007/s11336-017-9574-9

Comrey, A. L., and Lee, H. B. (1992). *A First Course in Factor Analysis.* Hillsdale, NJ: Erlbaum.

Connor, K. M., and Davidson, J. R. (2003). Development of a new resilience scale: the Connor-Davidson resilience scale (CD-RISC). *Depress. Anxiety* 18, 76–82. doi: 10.1002/da.10113

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. *Comput. Stat. Data Anal.* 45, 215–233. doi: 10.1016/S0167-9473(02)00366-3

Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychol. Bull.* 105, 317–327. doi: 10.1037/0033-2909.105.2.317

Dai, Y., Gao, W., and Huang, Y. (2016). The influence of psychological resilience on subjective well-being of college freshmen: the function of interpersonal forgiveness and self-forgiveness. *J. Sichuan Univ. Sci. Eng.* 31, 11–19. doi: 10.11965/xbew20160102

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Struct. Equat. Modeli.* 21, 239–252. doi: 10.1080/10705511.2014.882686

de Winter, J. C. F., Dodou, D., and Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behav. Res.* 44, 147–181. doi: 10.1080/00273170902794206

Ding, H., Han, J., Zhang, M., Wang, K., Gong, J., and Yang, S. (2017). Moderating and mediating effects of resilience between childhood trauma and depressive symptoms in Chinese children. *J. Affect. Dis.* 21, 130–135. doi: 10.1016/j.jad.2016.12.056

Efron, B., and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* New York, NY: Chapman & Hall.

Everitt, B. S. (1975). Multivariate analysis: The need for data, and other problems. *Br. J. Psychiatry* 126, 237–240. doi: 10.1192/bjp.126.3.237

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Struct. Equat. Model.* 7, 356–410. doi: 10.1207/S15328007 SEM0703_2

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis,* 3rd. Edn. Boca Raton, FL: CRC Press.

Gerbing, D. W., and Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behav. Res.* 20, 255–271. doi: 10.1207/s15327906mbr2003_2

Gorsuch, R. L. (1983). *Factor Analysis,* 2nd Edn. Hillsdale, NJ: Erlbaum.

Guilford, J. P. (1954). *Psychometric Methods* (2nd Edn.) New York, NY: McGraw-Hill.

Herzog, W., and Boomsma, A. (2009). Small-sample robust estimators of noncentrality-based and incremental model fit. *Struct. Equat. Model.* 16, 1–27. doi: 10.1080/10705510802561279

Herzog, W., Boomsma, A., and Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Struct. Equat. Model.* 14, 361–390. doi: 10.1080/10705510701301602

Hu, L., Bentler, P. M., and Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychol. Bull.* 112, 351–362. doi: 10.1037/0033-2909.112.2.351

Huang, Y., and Bentler, P. M. (2015). Behavior of asymptotically distribution free test statistics in covariance versus correlation structure analysis. *Struct. Equat. Model.* 22, 489–503. doi: 10.1080/10705511.2014.954078

Huang, P.-H., Chen H., and Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika* 82, 329–354. doi: 10.1007/s11336-017-9566-9

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: a Monte Carlo investigation. *Struct. Equat. Model.* 8, 205–223. doi: 10.1207/S15328007SEM0802_3

Jacobucci, R., Grimm, K. J., and McArdle, J. J. (2016). Regularized structural equation modeling. *Struct. Equat. Model.* 23, 555–566. doi: 10.1080/10705511.2016.1154793

Jiang, G., and Yuan, K.-H. (2017). Four new corrected statistics for SEM with small samples and nonnormally distributed data. *Struct. Equat. Model.* 24, 479–494. doi: 10.1080/10705511.2016.1277726

Jöreskog, K. G., and Sörbom, D. (1993). *LISREL 8 User's Reference Guide.* Chicago: Scientific Software International.

Kim, D. H., and Yoo, I. Y. (2004). Relationship between depression and resilience among children with nephrotic syndrome. *J. Korean Acad. Nurs.* 34, 534–540. doi: 10.4040/jkan.2004.34.3.534

Kline, P. (1979). *Psychometrics and Psychology.* London: Academic Press.

Lawley, D. N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika* 43, 295–303. doi: 10.1093/biomet/43.3-4.295

Lawley, D. N., and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method* (2nd Edn.) New York, NY: American Elsevier.

Lee, S.-Y. (2007). *Structural Equation Modelling: A Bayesian Approach.* Chichester: Wiley.

Lee, S.-Y., and Jennrich, R. I. (1979). A study of algorithms for covariance structure analysis with specific comparisons using factor analysis. *Psychometrika* 44, 99–114. doi: 10.1007/BF02293789

Lee, S. Y., and Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behav. Res.* 39, 653–686. doi: 10.1207/s15327906mbr3904_4

Lei, P.-W. (2009). Evaluating estimation methods for ordinal data in structural equation modeling. *Qual. Quant.* 43, 495–507. doi: 10.1007/s11135-007-9133-z

Li, C.-H. (2016). The performance of ML, DWLS, and ULS estimation with robust corrections in structural equation models with ordinal variables. *Psychol. Methods* 21, 369–387. doi: 10.1037/met0000093

Li, J., Zhong, G. K., Liu, H. Li, G. P., and Li, P. (2016). Analysis of correlation between resilience, depressive mood and quality of life in patients with depression. *J. Qilu Nurs.* 22, 11–13. doi: 10.3969/j.issn.1006-7256.2016.03.005

Long, J. S., and Ervin L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *Am. Stat.* 54, 217–224. doi: 10.2307/2685594

MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong S. (1999). Sample size in factor analysis. *Psychol. Methods* 4, 84–99. doi: 10.1037/1082-989X.4.1.84

MacKinnon, J. G., and White, H. (1985). Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. *J. Econom.* 29, 305–325. doi: 10.1016/0304-4076(85)90158-7

Marcoulides, K. M., and Yuan, K.-H. (2017). New ways to evaluate goodness of fit: a note on using equivalence testing to assess structural equation models. *Struct. Equat. Model.* 24, 148–153. doi: 10.1080/10705511.2016.1225260

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika* 82, 533–558. doi: 10.1007/s11336-016-9552-7

McNeish, D. (2016). On using Bayesian methods to address small sample problems. *Struct. Equat. Model.* 23, 750–773. doi: 10.1080/10705511.2016.1186549

McNeish, D. M. (2016). Using data-dependent priors to mitigate small sample bias in latent growth models: a discussion and illustration using Mplus. *J. Educ. Behav. Stat.* 41, 27–56. doi: 10.3102/1076998615621299

McQuitty, S. (1997). Effects of employing ridge regression in structural equation models. *Struct. Equat. Model.* 4, 244–252. doi: 10.1080/10705519709540074

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychol. Bull.* 105, 156–166. doi: 10.1037/0033-2909.105.1.156

Mooijaart, A., and Bentler, P. B. (2011). The weight matrix in asymptotic distribution-free methods. *Br. J. Math. Stat. Psychol.* 38, 190–196. doi: 10.1111/j.2044-8317.1985.tb00833.x

Moore, T. M., Reise, S. P., Depaoli, S., and Haviland, M. G. (2015). Iteration of partially specified target matrices: applications in exploratory and Bayesian confirmatory factor analysis. *Multivariate Behav. Res.* 50, 149–161. doi: 10.1080/00273171.2014.973990

Moshagen, M. (2012). The model size effect in SEM: inflated goodness-of-fit statistics are due to the size of the covariance matrix. *Struct. Equat. Model.* 19, 86–98. doi: 10.1080/10705511.2012.634724

Mundfrom, D. J., Shaw, D. G., and Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *Int. J. Test.* 5, 159–168. doi: 10.1207/s15327574ijt0502_4

Nevitt, J., and Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behav. Res.* 39, 439–478. doi: 10.1207/S15327906MBR3903_3

Poole, J. C., Dobson, K. S., and Pusch, D. (2017). Childhood adversity and adult depression: the protective role of psychological resilience. *Child Abuse Neglect* 64, 89–100. doi: 10.1016/j.chiabu.2016.012

Preacher, K. J., and MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: factor recovery with small sample sizes. *Behav. Genet.* 32, 153–161. doi: 10.1023/A:1015210025234

Rao, J. N. K., and Scott, A. J. (1984). On chi-squared tests for multi-way contingency tables with cell proportions estimated from survey data. *Ann. Stat.* 12, 46–60. doi: 10.1214/aos/1176346391

Rhemtulla, M., Brosseau-Liard, P. E., and Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychol. Methods* 17, 354–373. doi: 10.1037/a0029315

Saffarinia, M., Mohammadi, N., and Afshar, H. (2016). The role of interpersonal forgiveness in resilience and severity of pain in chronic pain patients. *J. Fundam. Mental Health* 18, 212–219. Available online at: http://eprints.mums.ac.ir/id/eprint/1926

Sarris, W. E., Satorra, A., and Sörbom, D. (1987). The detection and correction of specification errors in structural equation models. *Sociol. Methodol.* 17, 105–129.

Satorra, A., and Bentler, P. M. (1994). "Corrections to test statistics and standard errors in covariance structure analysis," in *Latent Variables Analysis: Applications for Developmental Research,* eds A. von Eye and C. C. Clogg (Newbury Park, CA: Sage), 399–419.

Savalei, V. (2010). Small sample statistics for incomplete nonnormal data: extensions of complete data formulae and a Monte Carlo comparison. *Struct. Equat. Model.* 17, 245–268. doi: 10.1080/10705511003659375

Scheines, R., Hoijtink, H., and Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika* 64, 37–52. doi: 10.1007/BF02294318

Shapiro, A. (1983). Asymptotic distribution theory in the analysis of covariance structures (a unified approach). *South Afr. Stat. J.* 17, 33–81.

Shi, D., Lee, T., and Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Struct. Equat. Model.* 25, 21–40. doi: 10.1080/10705511.2017.1369088

Song, L. P., Bai, X. J., Wang, M. J., Li, S. P., Zhao, J., He, X. X., et al. (2017). The research of mental resilience, emotional intelligence and automatic thoughts

among the early adulthood of depression. *China J. Health Psychol.* 25, 188–191. doi: 10.13342/j.cnki.cjhp.2017.02.009

Song, R., and Liu, A. S. (2013). Childhood psychological maltreatment to depression: Mediating roles of automatic thoughts. *J. Psychol. Sci.* 36, 855–859. doi: 10.16719/j.cnki.1671-6981.2013.04.003

Steiger, J. H., and Lind, J. M. (1980). "Statistically based tests for the number of common factors," in *Paper Presented at the Annual Meeting of the psychometric Society* (Iowa City, IA).

Swain, A. J. (1975). *Analysis of Parametric Structures for Variance Matrices.* Doctoral dissertation, University of Adelaide.

Thompson, L. Y., Snyder, C. R., Hoffman, L., Michael, S. T., Rasmussen, H. N., Billings, L. S., et al. (2005). Dispositional forgiveness of self, others, and situations. *J. Pers.* 73, 313–359. doi: 10.1111/j.1467-6494.2005.00311.x

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. B* 58, 267–288.

Tong, X., Zhang, Z., and Yuan, K.-H. (2014). Evaluation of test statistics for robust structural equation modeling with nonnormal missing data. *Struct. Equat. Model.* 21, 553–565. doi: 10.1080/10705511.2014.919820

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., and van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *Eur. J. Psychotraumatol.* 6:25216. doi: 10.3402/ejpt.v6.25216

Wakaki, H., Eguchi, S., and Fujikoshi, Y. (1990). A class of tests for a general covariance structure. *J. Multivariate Anal.* 32, 313–325. doi: 10.1016/0047-259X(90)90089-Z

Wang, M. J., Song, L. P., Song, B. L., Du, Q. R., He, X. X., Bai, X. J., et al. (2017). Correlation Analysis of Resilience, Defense Mechanism and Depression in Patient with Depression. *China J. Health Psychol.* 25, 165–168. doi: 10.13342/j.cnki.cjhp.2017.02.002

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* 29, 350–361. doi: 10.1093/biomet/29.3-4.350

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *J. Am. Stat. Assoc.* 76, 419–433. doi: 10.1080/01621459.1981.10477663

Wothke, W. (1993). "Nonpositive definite matrices in structural modeling," in *Testing Structural Equation Models,* eds K. A. Bollen and J. S. Long (Newbury Park, CA: Sage), 256–293.

Xing, L., and Yuan, K.-H. (2017). "Model evaluation with small N and/ or large p," in *Poster Presented at the 2017 Annual Meeting of the Psychometric Society* (Zurich).

Xu, W., and Li, Y. (2017). Multiple mediation effects between physical exercise and depression of female college students. *China J. Health Psychol.* 25, 1231–1235. doi: 10.13342/j.cnki.cjhp.2017.08.031

Yang, M., Jiang, G., and Yuan, K.-H. (2018). The performance of ten modified rescaled statistics as the number of variables increases. *Struct. Equat. Model.* 25, 414–438. doi: 10.1080/10705511.2017.1389612

Yang, M., and Yuan, K.-H. (2016). Robust methods for moderation analysis with a two-level regression model. *Multivariate Behav. Res.* 51, 757–771. doi: 10.1080/00273171.2016.1235965

Yu, X. X., and Zheng, X. (2008). Healthy personality and its relationship to subjective well-being among college students. *Chin. J. Public Health* 24, 507–509. doi: 10.3321/j.issn:1001-0580.2008.04.067

Yuan, K.-H. (2005). Fit indices versus test statistics. *Multivariate Behav. Res.* 40, 115–148. doi: 10.1207/s15327906mbr4001_5

Yuan, K.-H., and Bentler, P. M. (1997a). Mean and covariance structure analysis: Theoretical and practical improvements. *J. Am. Stat. Assoc.* 92, 767–774. doi: 10.1080/01621459.1997.10474029

Yuan, K.-H., and Bentler, P. M. (1997b). Improving parameter tests in covariance structure analysis. *Comput. Stat. Data Anal.* 26, 177–198. doi: 10.1016/S0167-9473(97)00025-X

Yuan, K.-H., and Bentler, P. M. (1998). Normal theory based test statistics in structural equation modeling. *Br. J. Math. Stat. Psychol.* 51, 289–309. doi: 10.1111/j.2044-8317.1998.tb00682.x

Yuan, K.-H., and Bentler, P. M. (1999). *F*-tests for mean and covariance structure analysis. *J. Educ. Behav. Stat.* 24, 225–243. doi: 10.3102/10769986024003225

Yuan, K.-H., and Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociol. Methodol.* 30, 167–202. doi: 10.1111/0081-1750.00078

Yuan, K.-H., and Bentler, P. M. (2010). Two simple approximations to the distributions of quadratic forms. *Br. J. Math. Stat. Psychol.* 63, 273–291. doi: 10.1348/000711009X449771

Yuan, K.-H., and Bentler, P. M. (2017). Improving the convergence rate and speed of Fisher-scoring algorithm: ridge and anti-ridge methods in structural equation modeling. *Ann. Inst. Stat. Math.* 69, 571–597. doi: 10.1007/s10463-016-0552-2

Yuan, K.-H., Bentler, P. M., and Chan, W. (2004). Structural equation modeling with heavy tailed distributions. *Psychometrika* 69, 421–436. doi: 10.1007/BF02295644

Yuan, K.-H., and Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Comput. Stat. Data Anal.* 52, 4842–4858. doi: 10.1016/j.csda.2008.03.030

Yuan, K.-H., and Chan, W. (2016). Structural equation modeling with unknown population distributions: ridge generalized least squares. *Struct. Equat. Model.* 23, 163–179. doi: 10.1080/10705511.2015.1077335

Yuan, K.-H., Chan, W., and Bentler, P. M. (2000). Robust transformation with applications to structural equation modeling. *Br. J. Math. Stat. Psychol.* 53, 31–50. doi: 10.1348/000711000159169

Yuan, K.-H., Cheng, Y., and Zhang, W. (2010). Determinants of standard errors of MLEs in confirmatory factor analysis. *Psychometrika* 75, 633–648. doi: 10.1007/s11336-010-9169-1

Yuan, K.-H., and Hayashi, K. (2003). Bootstrap approach to inference and power analysis based on three statistics for covariance structure models. *Br. J. Math. Stat. Psychol.* 56, 93–110. doi: 10.1348/000711003321645368

Yuan, K.-H., and Hayashi, K. (2006). Standard errors in covariance structure models: asymptotics versus bootstrap. *Br. J. Math. Stat. Psychol.* 59, 397–417. doi: 10.1348/000711005X85896

Yuan, K.-H., Jiang, G., and Yang, M. (2017). Mean and mean-and-variance corrections with big data. *Struct. Equat. Model.* 25, 214–229. doi: 10.1080/10705511.2017.1379012

Yuan, K.-H., Tian, Y., and Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika* 80, 379–405. doi: 10.1007/s11336-013-9386-5

Yuan, K.-H., Wu, R., and Bentler, P. M. (2011). Ridge structural equation modeling with correlation matrices for ordinal and continuous data. *Br. J. Math. Stat. Psychol.* 64, 107–133. doi: 10.1348/000711010X497442

Yuan, K.-H., Yang, M., and Jiang, G. (2017). Empirically corrected rescaled statistics for SEM with small *N* and large *p*. *Multivariate Behav. Res.* 52, 673–698. doi: 10.1080/00273171.2017.1354759

Yuan, K.-H., and Zhang, Z. (2012). Robust structural equation modeling with missing data and auxiliary variables. *Psychometrika* 77, 803–826. doi: 10.1007/s11336-012-9282-4

Yuan, K.-H., and Zhong, X. (2008). Outliers, leverage observations and influential cases in factor analysis: minimizing their effect using robust procedures. *Sociol. Methodol.* 38, 329–368. doi: 10.1111/j.1467-9531.2008.00198.x

Zhang, Y., Liu, R., Li, G., Mao, S., and Yuan, Y. (2015). The reliability and validity of a Chinese-version short health anxiety inventory: an investigation of university students. *Neuropsychiatr. Dis. Treat.* 11, 1739–1747. doi: 10.2147/NDT.S83501

Zhang, Z., Hamagami, F., Wang, L. L., Nesselroade, J. R., and Grimm, K. J. (2007). Bayesian analysis of longitudinal data using growth curve models. *Int. J. Behav. Dev.* 31, 374–383. doi: 10.1177/0165025407077764