Political Science Research and Methods (2025), 1–21
doi:10.1017/psrm.2025.27

ORIGINAL ARTICLE

Improved LM test for robust model specification searches in covariance structure analysis: application in political science research

Bang Quan Zheng[1,2,3] and Peter M. Bentler[1,2]

[1]Annette Strauss Institute for Civic Life, Moody College of Communication, and Department of Government, University of Texas at Austin, Austin, TX, USA; [2]Departments of Psychology & Statistics, UCLA, Los Angeles, CA, USA and [3]School of Government & Public Policy, University of Arizona, Tucson, Arizona, US
Corresponding author: Bang Quan Zheng; Email: bangquan@ucla.edu

(Received 16 November 2023; revised 11 November 2024; accepted 17 November 2024)

Abstract
Covariance structure analysis or structural equation modeling is critical for political scientists measuring latent structural relationships, allowing for the simultaneous assessment of both latent and observed variables, alongside measurement error. Well-specified models are essential for theoretical support, balancing simplicity with optimal model fit. However, current approaches to improving model specification searches remain limited, making it challenging to capture all meaningful parameters and leaving models vulnerable to chance-based specification risks. To address this, we propose an improved Lagrange multiplier (LM) test incorporating stepwise bootstrapping in LM and Wald tests to detect omitted parameters. Monte Carlo simulations and empirical applications underscore its effectiveness, particularly in small samples and models with high degrees of freedom, thereby enhancing statistical fit.

Keywords: bootstrap method; covariance structure analysis; LM test; model specification

1. Introduction

Political scientists often work with complicated and abstract concepts like democracy, political efficacy, values, ideology, identity, trust, among others, which are difficult to measure directly (Acock *et al.*, 1985; Feldman, 1988; Goren, 2005; Davidov, 2009; Pietryka and MacIntosh, 2013). To measure these concepts, political scientists often use covariance structure analysis (CSA) or structural equation modeling (SEM) with latent variables to build statistical models that combine multiple observed indicators, enabling the assessment of underlying theoretical constructs. The versatility of CSA or SEM is crucial in political science, especially in political psychology. It offers a rigorous statistical framework for modeling structural relationships, testing mediation, analyzing latent constructs, and evaluating measurement validity. Additionally, it allows researchers to test multiple hypotheses about the relationships between latent and observable variables while simultaneously accounting for measurement error (Blackwell *et al.*, 2017; Yuan and Liu, 2021; Zheng and Bentler, 2024). For instance, SEM has been successfully employed in studying diverse relationships, including the connection between party identification and core values (Goren, 2005), political conceptualization (Zheng, 2023), political tolerance and democracy theory (Sullivan *et al.*, 1981), values and support for immigration (Davidov, 2009), the underlying dimensions of racial

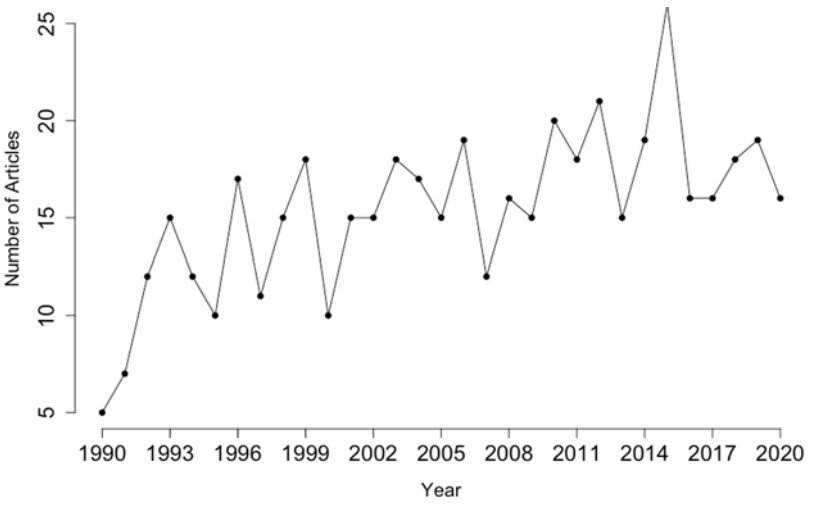

Figure 1. Number of articles published in selected PS journals using SEM.
Note: The data are based on a Google Scholar advanced search covering the years 1990–2020, focusing on publications in *The American Political Science Review, American Journal of Political Science, The Journal of Politics, Political Psychology, Political Behavior*, and *Public Opinion Quarterly.*

attitudes (DeSante and Smith, 2022), and measurement invariance analysis (Davidov, 2009; Pietryka and MacIntosh, 2013; Oberski, 2014).

Due to its versatility, CSA or SEM has shown a modest yet consistent trend in usage within political science research over the past decades. Figure 1 shows the frequency of articles involving SEM across six political science journals. Data collection, conducted through Google Scholar advanced search, spans from 1990 to 2020. The keywords used were "structural equation modeling," "covariance structure analysis," and "factor analysis." Figure 1 illustrates that between 1990 and 2020, the number of articles utilizing SEM increased in a nearly linear fashion, starting at approximately 10 articles per year and rising to around 20 articles per year. This trend underscores a sustained interest in applying this methodology, particularly within political psychology.

As with any modeling technique, the adequate specification of CSA or SEM models is critical for making sound decisions and drawing valid inferences (Zheng and Bentler, 2024). Identifying suitable parameters to fit complex models becomes particularly challenging when dealing with a large number of observed variables relative to small sample sizes. In such instances, there's a heightened chance-based model risks (a.k.a. capitalizing on chance in psychometric literature), which can compromise the reliability of the findings (Bentler 2006; Sörbom, 1989; Yuan and Liu, 2021). This pervasive issue highlights the necessity for robust techniques that can enhance the stability of these models. Unlike regression models, where the focus is primarily on the relationship between the dependent and key independent variables, with coefficients holding the most weight, SEM can handle intricate models with numerous interrelated variables and pathways. It facilitates the examination of complex theoretical frameworks. Therefore, researchers' arguments rely on the underlying structural relationships, rendering both model specification and fit equally crucial.

This study proposes a novel method, the improved Lagrange multiplier (LM) test, for model specification searches, addressing the challenge of noise interference. Our data-driven specification search method, using the stepwise bootstrap approach in both LM and Wald tests, effectively identifies potential omitted parameters, improving the precision of parameter identification. Through a series of simulation studies and two empirical applications in political science, our results demonstrate that the improved LM test is particularly reliable when dealing with small sample sizes in models with

high degrees of freedom. The improved LM test enhances the reliability, validity, and statistical fit of model specifications while mitigating the risk of being misled by noise, enabling researchers to draw sound conclusions grounded in solid statistical evidence.

As we will demonstrate later with empirical examples from Huddy and Khatib (2007), Davidov (2009), and Oberski (2014), the theoretical arguments in these studies are grounded in structural relationships among sets of latent and observed variables. Inadequate model specification could undermine these arguments, whereas a well-specified model with robust goodness-of-fit can reinforce them. For instance, Huddy and Khatib argue that national identity is distinct from other forms of national attachment, such as symbolic, constructive, uncritical patriotism, and nationalism, and that a strong American identity promotes civic involvement. However, the weak χ^2 test statistic reported in their research may call this claim into question. Moreover, the new parameter identified through the improved LM test not only reinforces Huddy and Khatib's original argument but also underscores the significant role of national identity in driving emotional reactions, such as anger—an aspect that was overlooked in their original model. In Davidov's (2009) and Oberski's (2014) models, we focused exclusively on the German sample and identified two omitted variables that indicate potential model misspecification. While this misspecification may not be substantial enough to alter the substantive conclusions, including these variables strengthens the authors' arguments by improving model fit. If untested, confounding measurement inequivalence with structural differences could lead to specification issues.

2. Challenges and existing approaches in model specification searches

In this section, we review existing approaches for model specification searches, covering major tests, their procedures, and model fit evaluations. To evaluate the model fit between the theoretical model and sample data, researchers must assess the model's adequacy using goodness-of-fit tests. However, before trusting the χ^2 test statistics and other fit indices, adequate model modification and specification are necessary. In CSA or SEM, a desirable model fit involves striking a balance between simplifying the model without compromising the overall fit and improving the model fit without making it more complicated (Bentler and Chou, 1992; MacCallum *et al.*, 1992). Several critical factors can affect overall model fit, such as poorly specified models. Typically, researchers specify a model based on a priori knowledge and fit it to sample data by estimating parameters. To modify the model, researchers determine the number of parameters to add or remove from the existing model and then refit it with the same dataset. If the initial model fit is inadequate, a common practice is to free parameter restrictions to enhance the model's fit to the data (Leamer, 1978; Kaplan, 1988; Sörbom, 1989; Bentler and Chou, 1992; MacCallum *et al.*, 1992). This process is referred to as model specification search.

The goal of model specification searches and modifications is to develop a generalizable model that demonstrates stability. Stability refers to the consistency of model results across repeated samples (MacCallum *et al.*, 1992). While it is challenging to achieve a perfect model in practice, an acceptable model specification should consist of a set of parameters supported by substantive theories that also has an adequate statistical fit (Bentler and Chou, 1992; MacCallum *et al.*, 1992; Yuan *et al.*, 2007; Chou and Huh, 2012).

The process of modifying and specifying any statistical models can be influenced by idiosyncratic characteristics of the data, meaning that modifications and specifications that improve the fit of one model may not necessarily apply to another random sample from the same population. This challenge, often referred to as the chance-based model risks, becomes particularly pronounced in large models with high degrees of freedom but relatively small sample sizes (MacCallum *et al.*, 1992). In such cases, increased sampling variability in the sample covariance can significantly impact the results of CSA or SEM analyses, leading to inconsistencies across different samples. Despite the significance of this issue, there are currently no systematic approaches to enhance model modification and specification

in CSA or SEM, making it challenging to include all statistically meaningful parameters in the model without the interference of noise.

Two key considerations for assessing model adequacy are model parsimony and model fit. Model parsimony refers to the number of free parameters in the model, while model fit is evaluated using empirical fit indices. Poor model fit can occur in two scenarios: if a model inadequately fits the data (under-specified), requiring modifications by releasing constraints on fixed parameters in a "forward search," or if a model fits the data well but has excessive parameters (overfitting), necessitating simplification through constraints on free parameters in a "backward search."

2.1. The LM test

In multivariate analysis of CSA or SEM, two commonly used test statistics are the LM test and the Wald test. The LM test only requires estimating the restricted model, while the Wald test requires a more comprehensive model. Notably, the statistical theory for the LM test is more complex than for the Wald test. This study focuses on estimating the restricted model under various constraints. The LM test is particularly useful for guiding model modifications to improve fit, as it identifies the effects of freeing initially fixed parameters (Lee and Bentler, 1980; Bentler, 1986; Satorra, 1989; Sörbom, 1989; Yuan and Liu, 2021).

Standard LM tests rely on a single snapshot of the initial model, which may not accurately identify missing parameters in population data. Consequently, model misspecifications can occur, leading to poor generalization to new samples. This limitation is particularly evident with small sample sizes, as the effectiveness of model modification using the LM test becomes compromised and susceptible to random variations (MacCallum *et al.*, 1992; Yuan and Liu, 2021).

A model is deemed acceptable when its parameters align with theories and show good statistical fit to the data (Chou and Huh, 2012). If a model has q free parameters and an additional nondependent variable, where $r < q$, the LM test can be used to identify which fixed parameters should be freed for better model fit. This is done by computing the LM test statistic T_{LM}. The LM test employs forward specification searching, where a constraint in the initial model is proposed to be freed based on how much it would enhance model fit.

A model consists of both free and fixed parameters, with the latter included to specify the model. Let $\widehat{\boldsymbol{\theta}}$ be a vector of constrained estimators of $\boldsymbol{\theta}$ that satisfies the $r < q$ constraints $h(\boldsymbol{\theta}) = 0$ when minimizing the fit function $F(\boldsymbol{\theta})$ for a given model. This is equivalent to minimizing the function of a constrained model while assuming $h(\boldsymbol{\theta}) = \theta_r = 0$. With r constraints, there exists an $r \times 1$ constraint vector $h(\boldsymbol{\theta})' = (h_1, \ldots, h_r)$. When minimizing the fit function $F(\boldsymbol{\theta})$ with constraints of $h(\boldsymbol{\theta}) = 0$, matrices of derivatives $\boldsymbol{g} = \left(\frac{\partial F}{\partial \boldsymbol{\theta}}\right)$ and $\boldsymbol{L}' = \left(\frac{\partial h}{\partial \boldsymbol{\theta}}\right)$ exist for the "forward search," and there will be a vector of LMs, $\boldsymbol{\lambda}$, such that

$$\hat{\boldsymbol{g}} + \hat{\boldsymbol{L}}'\hat{\boldsymbol{\lambda}} = 0 \text{ and } h\left(\widehat{\boldsymbol{\theta}}\right) = 0. \tag{1}$$

For the LM test to be applicable, several technical regularity conditions must be met, including the continuity of $\partial h/\partial \boldsymbol{\theta}$, model identification, positive definiteness of $\boldsymbol{\Sigma}$, linearly independent constraints, and full rank matrices of derivatives $\boldsymbol{g}$ and $\boldsymbol{L}'$. In the context of constraints, the asymptotic covariance matrix can be derived from an information matrix $\boldsymbol{H}$, augmented by the matrix of derivatives $\boldsymbol{L}$ and a null matrix $\mathbf{O}$. Thus, the sample variance covariance matrix of the estimated parameter, $\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)$, is given by the inverse of the Fisher information matrix of $\boldsymbol{H}(\boldsymbol{\theta})$, associated with q free parameters in $\boldsymbol{\theta}$ in the case of maximum likelihood (ML) estimation, and $\boldsymbol{R}$ gives the covariance matrix of the LMs $\sqrt{n}\left(\hat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}\right)$, which is derived from the inverse of the information matrix $\boldsymbol{L}$. Therefore, we can define

$$\begin{bmatrix} \boldsymbol{H} & \boldsymbol{L}' \\ \boldsymbol{L} & \boldsymbol{O} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{L}'(\boldsymbol{L}\boldsymbol{H}^{-1}\boldsymbol{L}')^{-1} & \boldsymbol{H}^{-1}\boldsymbol{L}'(\boldsymbol{L}\boldsymbol{H}^{-1}\boldsymbol{L}')^{-1} \\ (\boldsymbol{L}\boldsymbol{H}^{-1}\boldsymbol{L}')^{-1}\boldsymbol{L}\boldsymbol{H}^{-1} & -(\boldsymbol{L}\boldsymbol{H}^{-1}\boldsymbol{L}')^{-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{M} & \boldsymbol{T}' \\ \boldsymbol{T} & -\boldsymbol{R} \end{bmatrix}. \tag{2}$$

Under regularity conditions and the null hypothesis $\boldsymbol{H}(\boldsymbol{\theta})$, the LM test is available in two versions, and the multivariate LM statistics are asymptotically distributed as a χ^2 variate with r degrees of freedom. They compute all constraints simultaneously:

$$T_{LM} = n\hat{\boldsymbol{\lambda}}'\widehat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{\lambda}} \sim \chi^2_r. \tag{3}$$

A univariate LM statistic is used to test a single constraint and is distributed as a χ^2 variate with 1 *df*. This test is particularly useful for evaluating whether a specific parameter in the $\boldsymbol{\theta}_r$ vector is equal to 0:

$$T_{LMi} = n\hat{\boldsymbol{\lambda}}_i^2\widehat{\boldsymbol{R}}^{-1}\hat{\boldsymbol{\lambda}}_{\boldsymbol{i}} \sim \chi^2_{r1}. \tag{4}$$

This is also known as a modification index in the LISREL program (Jöreskog and Sörbom, 1988). For Equations (3) and (4), r is the number of nondependent constraints, and the matrices $\boldsymbol{H}$ and $\boldsymbol{L}'$ have dimensions $q \times q$ and $q \times r$, respectively:

$$\hat{\boldsymbol{\lambda}} = \left(\hat{\boldsymbol{L}}\widehat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{L}}'\right)^{-1}\hat{\boldsymbol{L}}\widehat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{g}}\,. \tag{5}$$

The LM test performance depends on factors like sample size, degrees of freedom, and the number of variables and parameters. Degrees of freedom are influenced by the number of parameters and variables. Higher p and lower q result in more degrees of freedom. The LM test considers all possible paths based on the degrees of freedom. With more degrees of freedom, the number of possible paths increases, increasing the risk of falsely identifying nonexistent paths in the true model, especially when there are few missing paths. However, this risk decreases with larger sample sizes, reducing reliance on chance occurrences.

2.2. The Wald test

The Wald test follows a backward stepwise procedure to identify which free parameter, starting with the one with the smallest Wald test statistic, should be removed from the model. This is done by including candidate parameters and performing univariate Wald tests on each. Parameters with statistically significant Wald test values are retained, and the process continues until no further parameters can be added. The Wald test statistic is calculated as follows:

$$W = n\widehat{\boldsymbol{\theta}}_{\boldsymbol{r}}'\left(\hat{\boldsymbol{L}}\widehat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{L}}'\right)\widehat{\boldsymbol{\theta}}_{\boldsymbol{r}} \sim \chi^2_r, \tag{6}$$

where $\hat{\boldsymbol{L}}$ is a quadratic form. The closer $\hat{\boldsymbol{L}}$ is to 0, the more likely it is that the null hypothesis equals 0 will be rejected. The univariate Wald statistics $\hat{\theta}_i$ for each of the parameters in $\widehat{\boldsymbol{\theta}}_{\boldsymbol{r}}$ can be expressed as

$$W_i = n\hat{\theta}_i\widehat{H}_{ii}^{-1}, \hat{\theta}_i = n\theta_i^2/\widehat{H}_{ii} \sim \chi^2_i, \tag{7}$$

where $\hat{\theta}_i$ is one of the parameters in $\widehat{\boldsymbol{\theta}}_{\boldsymbol{r}}$ and $\widehat{H}_{ii}$ is the ith parameter in the diagonal of the $\boldsymbol{H}$ matrix. The computational complexity of the LM and Wald tests depends on the number of free and fixed parameters in a model. As the number of parameters increases, estimating and comparing their effects becomes more computationally demanding. The $\boldsymbol{H}$ matrix, which represents the covariance matrix of the independent variables, reflects the number of parameters that can be either free or fixed. For a model with k independent variables, the $\boldsymbol{H}$ matrix will have k^2 elements (Chou and Huh, 2012).

2.3. Evaluation of model fit

This study uses ML estimation, the standard method for deriving goodness-of-fit statistics and parameter estimates in CSA under normal theory. In CSA, a random sample, $x \in\{x_1, \ldots., x_n\}$ is assumed to be independently and identically distributed, following a multivariate normal distribution $\mathcal{N}[\boldsymbol{\mu}, \boldsymbol{\Sigma}_0]$. Here, $\boldsymbol{\mu}$ represents a vector of sample means, and the covariance matrix $\boldsymbol{\Sigma}_0$ is assumed

positive definite with an unknown population parameter vector $\boldsymbol{\theta}_0$ of dimension $q \times 1$, where $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$. The sample covariance matrix is:

$$\boldsymbol{S} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' \tag{8}$$

where the sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}(x_1, \ldots, x_n)$. $\boldsymbol{S}$ serves as an unbiased estimator of the population covariance $\boldsymbol{\Sigma}_0$.

In confirmatory factor analysis (CFA), a model can be represented as:

$$\boldsymbol{x_i} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\xi_i} + \boldsymbol{\epsilon_i}, \ \ i = 1, \ldots, n \tag{9}$$

where $\boldsymbol{x_i}$ is a random sample, $\boldsymbol{\mu}$ is a sample mean vector, $\boldsymbol{\Lambda}$ is a matrix of factor loadings, $\boldsymbol{\xi_i}$ is a vector of latent factors, and $\boldsymbol{\epsilon_i}$ is a vector of residuals. Here, the parameters involved in a model are contained in the covariance matrix $\boldsymbol{\Sigma}$ of the observed variables. $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$, where $\boldsymbol{\Lambda}$ again is a factor loading matrix, and $\boldsymbol{\Phi}$ is a covariance matrix of the latent factors, and $\boldsymbol{\Psi}$ is a covariance matrix of unique scores.

The population covariance matrix $\boldsymbol{\Sigma}$ is modeled as $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ contains free parameters $\boldsymbol{\Lambda}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$. The sample covariance matrix $\boldsymbol{S}$ serves as an unbiased estimator of $\boldsymbol{\Sigma}$, with the null hypothesis $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$. An objective function $F[\boldsymbol{\Sigma}(\boldsymbol{\theta}), \ \boldsymbol{S}]$ measures the discrepancy between $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ and $\boldsymbol{S}$.

This study derives the goodness-of-fit statistic T_{ML} using the ML discrepancy function (Jöreskog, 1969), fitting the model-implied covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ to the sample covariance matrix $\boldsymbol{S}$, as shown in Equation (10):

$$F_{ML}(\theta) = \log|\boldsymbol{\Sigma}(\boldsymbol{\theta}) - \log|\boldsymbol{S}| + tr\left(\boldsymbol{S}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\right) - p \tag{10}$$

$$\widehat{\boldsymbol{\theta}}_{ML} = argmin\, F_{ML}(\boldsymbol{\theta}) \tag{11}$$

The ML fit function $F_{ML}(\boldsymbol{\theta})$ derives parameter estimates in $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ that minimize the test statistic. At its minimum, as shown in Equation (11), $\widehat{\boldsymbol{\theta}}_{ML}$ contains parameter estimates $\widehat{\boldsymbol{\Lambda}}$, $\widehat{\boldsymbol{\Phi}}$, and $\widehat{\boldsymbol{\Psi}}$. Using these, we can reconstruct the sample covariance matrix to align with the model-implied covariance $\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right) = \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Phi}}\widehat{\boldsymbol{\Lambda}}' + \widehat{\boldsymbol{\Psi}}$, assuming the sample-implied matrix matches the population matrix. If $\boldsymbol{S} \approx \boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)$ with p-value > 0.05, the model is considered plausible.

The ML goodness-of-fit test statistic is calculated as:

$$T_{ML} = (N-1)\, F_{ML}\left(\widehat{\boldsymbol{\theta}}\right). \tag{12}$$

This statistic is the product of $F_{ML}(\widehat{\boldsymbol{\theta}})$ and $(N-1)$, where N is sample size. As N increases, T_{ML} is expected to asymptotically follow a χ^2 distribution with corresponding degrees of freedom.

3. Improved LM test

To overcome the limitations in existing model specification searches, we propose a novel approach leveraging bootstrap methods for data-driven model specification searches, integrating the LM and Wald tests. It involves generating multivariate random samples through bootstrap resampling based on the initial model. We start with a forward stepwise bootstrap resampling method in the standard LM test. Following this, using the statistically significant results from the bootstrap LM, we apply a backward stepwise bootstrap Wald test to mitigate overfitting by identifying potential paths that may not be needed. This iterative workflow strikes a balance between maximizing model fit, which the LM test emphasizes, and maintaining parsimony, as the Wald test tends to emphasize. We term this approach the "improved LM test," offering a valuable tool for enhancing model fit and reducing chance-based model risks in applied CSA or SEM.

The improved LM approach for specification searches involves a multi-stage process. Initially, we create a hypothetical CFA population model and generate multivariate normal data with varying sample sizes using the Monte Carlo method. We then construct a misspecified analysis model by omitting several true parameters from the population model and fit it to the simulated data. To identify missing parameters in the measurement model, we conduct a univariate LM test to detect potential omissions. The parameters are ranked by their χ^2 statistics, and we select a series of them as the testing parameters. Next, we perform a multivariate stepwise LM test. This forward stepwise procedure follows a *general-to-specific* approach in specification searches within spatial econometrics (Florax *et al.*, 2003; Mur and Angulo, 2009). To enhance the reliability of LM test results, each sequence in the forward stepwise LM procedure is based on bootstrap resampling of the initial data. We calculate the means of the bootstrap LM test statistics and p-values, selecting parameters with p-values < 0.05 for further testing using bootstrap Wald tests to assess their stability.

In the Wald test procedure, we conduct a backward stepwise search by initially including all LM-based parameters in the model and then sequentially fixing one parameter at a time. This *specific-to-general* approach employs bootstrap resampling to compute the mean χ^2 test statistics and p-values. However, applying the Wald test to all missing parameters is generally impractical for midsize to large models, as the number of possible omitted parameters may exceed the sample size, resulting in a singular matrix that cannot be inverted. Limiting the Wald test to bootstrap LM-based selected parameters resolves this issue. We include the bootstrap LM-based parameters in the model and perform univariate Wald tests on each. By focusing on these parameters, we significantly reduce the number of elements in the $\boldsymbol{H}$ matrix in Equation (6) related to variances and covariances, as compared to a full model with all potential missing parameters. This reduction in $\boldsymbol{H}$ matrix elements facilitates efficient matrix inversion, leading to faster computations and more stable estimates. Thus, unlike other specification search methods in spatial econometrics, the improved LM test integrates LM and Wald tests with bootstrap resampling to reduce noise, enhance reliability, and distinguish between meaningful results and random outcomes. A detailed illustration is provided in the next section.

4. Simulation and multi-stage process for model specification searches

In this section, we illustrate the improved LM test for model specification, describing the setup of hypothetical population and analysis models, the simulation procedure, and a multi-stage process for conducting model specification searches that integrates bootstrap sampling within the LM and Wald tests.

4.1. Population model and analysis model

The simulation begins with a hypothetical population model consisting of a three-factor structure, with each factor measured by eight manifest variables, as illustrated in Figure 2. An analysis model, shown in Figure 3, consists of 3 factors, each linked to 8 indicators, with all factors freely correlated, resulting in a total of 24 variables. The four dashed lines in Figure 2 represent the paths not included in the analysis model. Omitting four parameters aims to reduce the chance of the initial model closely resembling the true model. Including many missing parameters can result in a poorly specified model, potentially rendering LM test results meaningless or falsely indicating statistical significance by chance (Yuan *et al.*, 2003).

4.2. Monte Carlo simulations

The simulated data are generated using a standard confirmatory factor model, given by Equation (13):

$$\boldsymbol{x_i} = \boldsymbol{\Lambda \xi_i} + \boldsymbol{\epsilon_i} \tag{13}$$

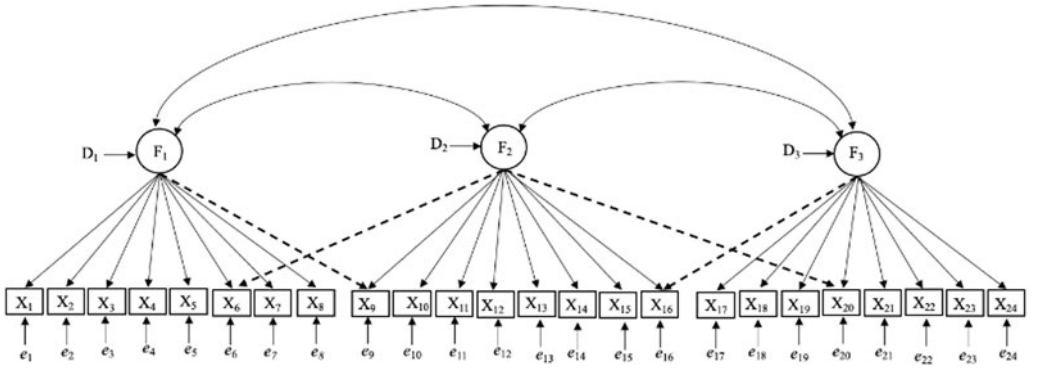

Figure 2. Path diagram of the population model.

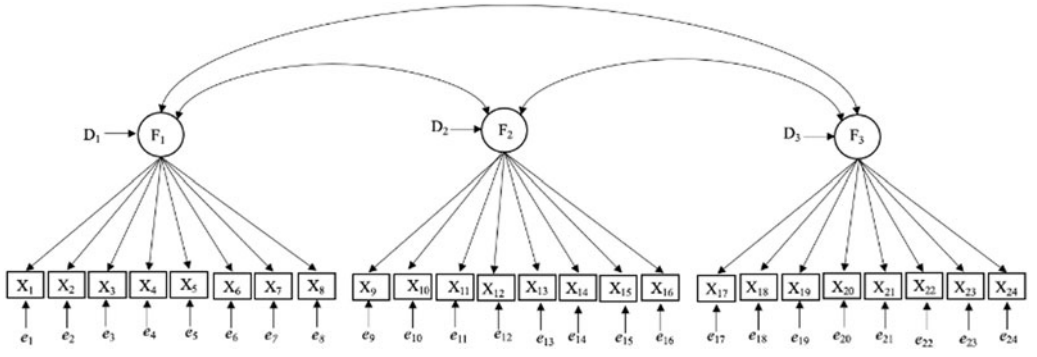

Figure 3. Path diagram of the misspecified analysis model.

where $\boldsymbol{x_i} = \left(x_{i1}, x_{i2}, \ldots x_{ip}\right)'$ is a vector of p observations on person i in a population, and $i = 1, 2, \ldots n$. $\boldsymbol{\Lambda}$ is a matrix of factor loadings, and $\boldsymbol{\epsilon_i} = \left(\epsilon_{i1}, \epsilon_{i2}, \ldots \epsilon_{ip}\right)'$ is a vector of error terms, and var $(\varepsilon) = \boldsymbol{\Psi}$. $\boldsymbol{\xi_i} = (\xi_{i1}, \xi_{i2,\ldots}\xi_{im})'$ is a vector of latent factors, and var$(\boldsymbol{\xi}) = \boldsymbol{\Phi}$. Each latent factor $\boldsymbol{\xi_i}$ has a mean and a variance and may correlate with other latent factors $\boldsymbol{\xi_i}$; whereas $\boldsymbol{\xi_i}$ and $\boldsymbol{\epsilon_i}$ are uncorrelated, so that $E\left(\boldsymbol{\xi}\right) = \boldsymbol{\mu}_{\xi}$, which is the mean of the factors.

With the data generation scheme and population model described above, we simulate a population and draw samples using Monte Carlo simulation, based on the predefined matrices $\boldsymbol{\Lambda}'$ and $\boldsymbol{\Phi}$:

$$\boldsymbol{\Lambda}' = \begin{bmatrix} 0.65 & 0.65 & 0.7 & 0.7 & 0.7 & 0.7 & 0.6 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0 & 0.6 & 0.6 & 0.6 & 0.7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 0.5 & 0.5 & 0.65 & 0 & 0 & 0 & 0.65 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.45 & 0.5 & 0.5 & 0.5 & 0.6 & 0.6 & 0.6 & 0.7 & 0.7 \end{bmatrix},$$

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & & \\ 0.3 & 1 & \\ 0.4 & 0.5 & 1 \end{bmatrix}$$

When $\text{diag}(\boldsymbol{\Sigma}) = \mathbf{I}$, which is an identity matrix, the unique variances can be determined by $\boldsymbol{\Psi} = \mathbf{I}_{24} - \text{diag}\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}'\right)$. Since we are not interested in the mean structure, we set the factor means $\mu's = (0,\ 0,\ 0)$. The data generating process consists of two steps. (1) We draw from a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Phi}$. Unique factors $\boldsymbol{\epsilon_i}$ are drawn from a multivariate normal distribution with zero mean and covariance $\boldsymbol{\Psi}$. Utilizing Equation (13), this procedure generates multivariate normal observations characterized by a covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

The data generation and all analyses for this research are conducted using the "lavaan" package (Version 4.2.3.) (Rosseel, 2012) in R, based on the previously specified population and assuming multivariate normality. The simulation studies involved sample sizes of $N = 100$ to 10,000. Our testing models consisted of 24 observed variables ($p = 24$) and 3 latent factors, resulting in a covariance component of $p^{\star} = 24(24 + 1)/2 = 300$, with 55 free parameters to estimate, and 245 degrees of freedom. This model size is a good representation of most SEM research.

To assess the stability of the model specifications, we conduct a series of Monte Carlo simulations using the population model (Figure 2) across different sample sizes and fit the analysis model (Figure 3). The study includes 12 sample sizes—100, 150, 200, 250, 300, 350, 400, 500, 1,000, and 2,000—that are selected to reveal important phenomena related to the issues under study. To evaluate the goodness-of-fit, we employ various methods, including χ^2 test statistics, standard deviations of the χ^2 test, and the rejection rate. Additionally, we utilize alternative fit measures such as the comparative fit index (CFI), normed fit index (NFI), and root mean square error of approximation (RMSEA). However, as the analysis model does not fit the population model, we expect the χ^2 test statistic to be larger than the degrees of freedom and the p-value < 0.05.

4.3. Selection of testing parameters

To commence the specification search, an initial set of parameters is required. The process of selecting these parameters begins with an exploratory univariate LM test via model modification indices (Sörbom, 1989). However, as the data are generated from Monte Carlo simulations, each sample drawn from the population model will be different, leading to variability in T_{LMi}. For instance, if we draw 500 samples of the same sample size, we will obtain 500 unique sets of initial testing parameters. Nonetheless, there should be a set of common parameters that frequently appear across all samples, including the true missing parameters. To obtain a more representative set of initial parameters, after we simulate the data for 500 trials, we calculate the mean T_{LMi} of each parameter and sort them in descending order. We then select the top 12 parameters with the largest T_{LMi} to test the proposed improved LM test. The LM test is designed to enhance the fit of the existing model, assuming it is reasonably well-fitted. Consequently, a sensible model should expect only a few significant omitted parameters. If the count exceeds 12, the model may suffer from severe misspecification issues, necessitating a new formulation.

4.4. Bootstrap simulation

The bootstrap method efficiently approximates population covariance structures in simulations, providing a practical alternative when the distribution of the sample is unknown. This approach tackles numerous challenges that conventional statistical methods encounter. For instance, the bootstrap approach does not assume normally distributed data. Even in cases where the data are normally distributed, at a given sample size, the bootstrap often provides more accurate results than those based on standard asymptotic methods (Yuan *et al.*, 2007).

Let $x_1, x_2, \cdots, x_n$ denote a sample with a covariance matrix represented as $\boldsymbol{S}$ where its population counterpart is $\boldsymbol{\Sigma}_0$. The bootstrap method iteratively draws samples from a known empirical distribution function, effectively substituting it for the population in the bootstrap samples. However, since

the population covariance matrix $\boldsymbol{\Sigma}_0$ is unknown, an alternative matrix $\dot{\boldsymbol{S}}$ must be found to serve as a surrogate for $\boldsymbol{\Sigma}_0$. Consequently, each x_i can be transformed into x_i' by:

$$x_i' = \dot{\boldsymbol{S}}^{1/2}\boldsymbol{S}^{-1/2}x_i, i = 1, 2, \cdots, n \tag{14}$$

where $\dot{\boldsymbol{S}}^{1/2}\ \dot{\boldsymbol{S}}^{1/2}$ is a $p \times p$ matrix satisfying $\left(\boldsymbol{S}^{-1/2}\right)\left(\boldsymbol{S}^{-1/2}\right)' = \dot{\boldsymbol{S}}$. The subsequent steps involve generating the bootstrap samples by sampling with replacement from $(x_1',\ x_2', \cdots,\ x_n')$, thereby computing the sample covariance matrix denoted as $\boldsymbol{S}^*$ for these bootstrap samples.

4.5. Bootstrap LM test

Chance-based model risks occur when different model fits arise from different samples of the same population, influenced by factors such as sample size, model complexity, and degrees of freedom. Significant chance-based model risks imply a higher likelihood of overlooking pathways within the model. To illustrate, Figure 4 visualizes the relationship between univariate LM tests, parameters, and various sample sizes. The x-axis represents the distribution of testing parameters, derived from 500 Monte Carlo simulations based on the population and analysis models depicted in Figures 2 and 3. On the other hand, the y-axis illustrates the univariate LM test statistics. Figure 4 highlights the true missing parameters. As observed, when the sample size is small, the distributions of LM tests for all parameters exhibit relatively large variations, including the true missing parameters. This phenomenon arises due to the small sample size relative to the model size and degrees of freedom. Consequently, the standard LM test faces challenges in distinguishing the true missing parameters from other parameters or effectively identifying any potential missing paths, presenting an issue due to its vulnerability to chance-based interpretations. However, as the sample size increases, the variation in the LM tests for the true missing parameters diminishes. This leads to a clearer distinction between the true missing parameters and other parameters, reducing the likelihood of being misled by noise.

In the third stage of our analysis, we employ a forward stepwise approach along with bootstrap resampling to identify the optimal specifications. While statistically significant large LM test values are observed, they don't necessarily imply that the suggested parameters accurately reflect the "true" values. This is because the data are generally a sample drawn from the larger population. Furthermore, in real-world data analysis, researchers frequently contend with finite sample sizes and unknown data distributions, giving rise to sample-specific errors and characteristics that may impact the model's accuracy in reflecting the population. Consequently, the parameters recommended by the standard LM test might not generalize effectively to different samples. In this regard, bootstrap resampling can handle the issue of unknown distribution and provide more accurate results than those based on standard asymptotic properties.

In our approach, we use bootstrap resampling in every forward stepwise procedure. The LM test for a set of omitted parameters can be broken down into a series of 1-*df* tests. Bentler and Dijkstra (1985) developed a forward stepwise LM procedure, where at each step, the parameter is chosen that will maximally increase the LM χ^2. We will perform the bootstrap LM test by examining two parameters at a time. At each step, we randomly draw 500 samples with replacements and compute the mean LM test statistic and its p-value. We will repeat this process by adding another pair of parameters until we have tested all possible omitted parameters in the analysis model.

It is crucial to perform multiple repetitions of the test during this process as the LM χ^2 value can vary depending on the model parameters and their correlations with each other. Adding or removing parameters will change the covariance structure for the LM test, and hence, the LM test statistic at each step will provide more accurate information about the remaining missing parameters. We designate the parameters with p-values less than 0.05 as statistically significant and refer to them as bootstrap LM-based parameters for the Wald test selection.

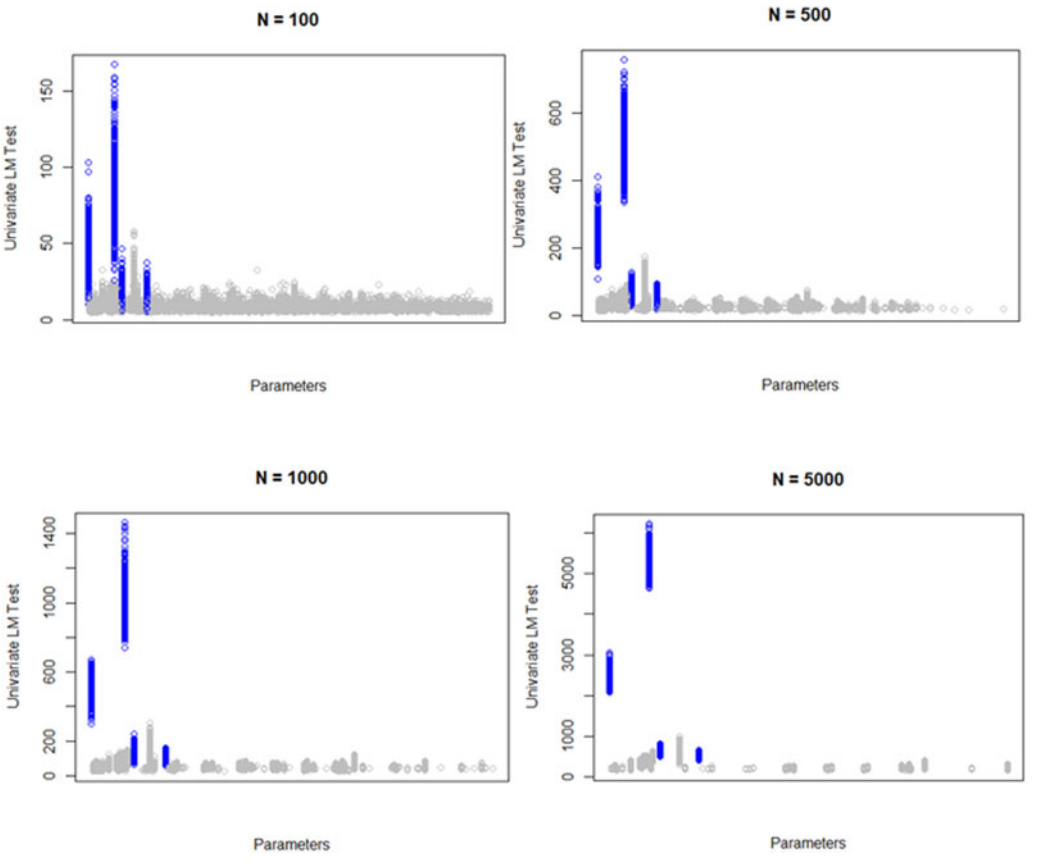

Figure 4. Univariate LM test statistics across varying sample sizes.

4.6. Bootstrap Wald test

The bootstrap LM tests establish a set of parameters for subsequent validation, while the Wald tests incorporate these recommended parameters and conduct a series of backward stepwise bootstrap Wald tests. The Wald tests assess whether each initially treated-as-free parameter can be collectively set to zero without a significant loss in model fit. This simplifies the model by removing nonsignificant parameters and provides further validation. Parameters that are truly missing exhibit p-values < 0.05, confirming their significance and justifying their inclusion in the model. Conversely, parameters that should be excluded from the model yield p-values ≥ 0.05. This integrated approach effectively addresses the problem of false positives.

5. Simulation results

To evaluate the performance and reliability of the improved LM test, we compare the results across various sample sizes while maintaining consistent degrees of freedom. The likelihood ratio test (LRT) is widely regarded as one of the most commonly used methods for assessing the performance of nested models. In this study, we compare the performances of the improved LM test to LRT for each sample size (see the Appendix for details on LRT calculation). Table 1 shows consistent performance for all models using bootstrap Wald tests across different sample sizes, affirming the combined methodology's effectiveness. The first column displays the top 12 possible parameters based on univariate LM tests, with the highlighted gray parameters representing known omitted parameters. For

brevity, the middle two columns only present the bootstrap Wald test (B-Wald) χ^2 statistics and their associated p-values. The last two columns show the LRTs and their associated p-values. Statistically significant values are highlighted in gray, indicating that their corresponding parameters should be included in the modified model. For detailed test results, please refer to Table A1 in the Appendix. As shown in Table 1, the improved LM tests accurately identify omitted parameters across all sample sizes, consistently outperforming LRTs, especially when sample sizes are small.

5.1. Model specification stability and model fit validity

In this section, we aim to assess the stability of the improved LM test-suggested model fit over repeated samples of different sample sizes. A model that fits the data well should follow a standard χ^2 distribution, $T_{ML} \xrightarrow{\mathcal{L}} \chi^2_{df}$, as N grows larger, demonstrating asymptotic properties (Browne, 1984; Bentler and Dijkstra, 1985; Jöreskog and Sörbom, 1988). Based on this reasoning, we fit the improved LM test-suggested model to the simulated data drawn from the population model (Figure 2). If the T_{ML} are close to the degrees of freedom, it provides strong empirical evidence that the improved LM test-suggested model fits better. To ensure its generalizability, we randomly draw samples of different sizes from the population model and fit the improved LM test-suggested model. If consistency is maintained, we are confident that the improved LM test-suggested model is adequate for general use.

The Monte Carlo simulations in this study are based on Equation (13). We conduct 1,000 trials and calculate the average statistics, which are reported in Table 2. We examine the performance of the analysis model suggested by the improved LM test by varying the sample sizes from 100 to 10,000. Since the simulated data are normally distributed, the ML estimator is sufficient to examine the basic statistical performance and asymptotic properties.

Table 2 presents the mean χ^2 test statistics, their mean standard deviations, mean p-values, mean rejection rates, and the 2.5th and 97.5th percentiles of fit indices (NFI, CFI, and RMSEA) by sample size. As shown in Table 2, as the sample size increases, all mean statistics test statistics get closer to the expected values: $\chi^2 = 269$, SD $= 23.195$, p-value $= 0.50$, and empirical rejection rate is 0.05. Note that when sample sizes are less than 500, the χ^2 test statistics are increasingly inflated, deviating from the expected value of 245. As documented by previous studies, ML estimator is biased against small sample sizes (Arruda and Bentler, 2017; Hayakawa, 2019; Zheng and Bentler, 2021, 2023).

In addition, NFI and CFI become closer to 1, and RMSEA is about 0 when the sample sizes are greater than 200. The collective statistical indicators provide compelling evidence that the improved LM test effectively detected the omitted parameters in the analysis model and provided a satisfactory fit to the simulated data. Moreover, the modified model, derived from the specification search results using the improved LM test, exhibited a robust statistical fit across various sample sizes.

5.2. Robustness of the improved LM test

To test the robustness of the improved LM test, we vary the magnitudes of factor correlations, the number of indicators per factor from low to high, and factor loadings. First, we find that with more indicators per factor, it becomes easier to detect omitted parameters. When the number of indicators per factor is fewer, the statistical power of the improved LM test is weakened, particularly with smaller sample sizes. Nevertheless, the improved LM test still outperforms the LRT across all sample sizes. When the number of indicators per factor increases, both the improved LM test and LRT perform similarly.

Second, when factor correlations are low, the improved LM test becomes more efficient at detecting omitted parameters. The performances of the improved LM test and the LRT become similar when sample sizes exceed 100. However, with smaller sample sizes, the improved LM test consistently outperforms the LRT. In contrast, when factor correlations are high, increasing potential

https://doi.org/10.1017/psrm.2025.27 Published online by Cambridge University Press

Table 1. Test statistics by different sample sizes

	$N = 100$					$N = 150$					$N = 250$				
	Para.	B-Wald	*P*-value	LRT	*P*-value	Para.	B-Wald	*P*-value	LRT	*P*-value	Para.	B-Wald	*P*-value	LRT	*P*-value
1	**F2, x6**	**63.877**	**0.000**	**363.190**	**0.000**	**F2, x6**	137.789	**0.000**	**375.610**	**0.000**	**F2, x6**	**128.070**	**0.000**	**474.880**	**0.000**
2	**F3, 16**	**19.996**	**0.000**	**337.110**	**0.000**	**F1, x9**	69.773	**0.000**	**309.720**	**0.000**	**F1, x9**	**71.814**	**0.000**	**387.540**	**0.000**
3	**F2, x20**	**27.681**	**0.000**	**313.420**	**0.000**	**F3, x16**	28.847	**0.006**	**285.180**	**0.000**	F3, x6	6.920	0.066	384.990	0.111
4	**F1, x9**	**19.313**	**0.001**	**283.980**	**0.000**	**F2, x20**	24.366	**0.000**	**241.480**	**0.000**	**F2, x20**	**63.686**	**0.000**	**312.610**	**0.000**
5	x4, x8			283.320	0.416	F3, x9	19.485	0.474	241.480	0.931	**F3, x16**	**29.380**	**0.000**	**258.990**	**0.000**
6	F2, x2	6.206	0.094	**277.230**	**0.014**	F2, x2	17.325	0.483	240.800	0.412	F2, x8	1.606	0.385	258.120	0.351
7	x4, x12	5.722	0.099	**272.450**	**0.029**	F1, x16	18.947	0.176	237.060	0.053	x6, x20	1.386	0.434	257.880	0.622
8	x4, x7			**267.480**	**0.026**	x6, x4	13.988	0.106	**232.230**	**0.028**	F2, x1	1.450	0.411	256.610	0.260
9	x5, x8			264.640	0.092	x6, x12	15.180	0.405	231.400	0.363	F2, x3	0.983	0.495	256.610	0.999
10	x12, x15	5.020	0.133	**259.810**	**0.028**	F2, x5	13.907	0.495	230.880	0.469	F1, x13	4.793	0.149	253.160	0.063
11	x6, x4	7.359	0.081	**254.870**	**0.026**	F2, x8	12.154	0.230	228.680	0.138	x3, x12	6.919	0.075	**246.850**	**0.012**
12	F2, x3			254.620	0.620	F3, x6	19.202	0.343	227.400	0.258	F2, x5	0.969	0.503	246.780	0.795
	$N = 350$					$N = 400$					$N = 500$				
	Para.	B-Wald	*P*-value	LRT	*P*-value	Para.	B-Wald	*P*-value	LRT	*P*-value	Para.	B-Wald	*P*-value	LRT	*P*-value
1	**F2, x6**	**203.534**	**0.000**	**584.870**	**0.000**	**F2, x6**	**214.186**	**0.000**	**537.270**	**0.000**	**F2, x6**	**302.512**	**0.000**	**720.280**	**0.000**
2	**F1, x9**	**133.946**	**0.000**	**419.970**	**0.000**	**F1, x9**	**115.641**	**0.000**	**365.330**	**0.000**	**F1, x9**	**189.343**	**0.000**	**466.390**	**0.000**
3	**F2, x20**	**88.283**	**0.000**	**291.160**	**0.000**	**F2, x20**	**80.792**	**0.000**	**252.840**	**0.000**	**F2, x20**	**120.855**	**0.000**	**321.340**	**0.000**
4	F3, x6	3.770	0.176	**281.600**	**0.002**	**F3, x16**	**80.792**	**0.000**	**208.280**	**0.000**	F3, x6	1.113	0.486	321.020	0.570
5	x6, x20	3.544	0.207	279.110	0.115	F2, x5	2.363	0.290	205.690	0.108	**F3, x16**	**46.686**	**0.000**	**255.710**	**0.000**
6	**F3, x16**	**31.355**	**0.000**	**231.680**	**0.000**	F1, x20	3.365	0.233	203.600	0.148	F2, x5	2.420	0.275	254.170	0.215
7	F2, x4	1.589	0.383	231.600	0.781	F2, x4	2.504	0.284	201.820	0.182	F3, x9	3.392	0.238	251.400	0.096
8	x11, x15	**11.806**	**0.012**	**219.580**	**0.001**	x20, x11	2.073	0.360	199.950	0.171	F2, x8	0.898	0.500	251.340	0.813
9	F2, x3	2.146	0.306	218.730	0.358	x5, x10	6.893	0.098	**194.350**	**0.018**	F2, x1	1.162	0.449	251.250	0.763
10	F2, x5	1.797	0.367	217.980	0.387	x5, x8	2.779	0.279	192.380	0.160	F2, x3	1.476	0.414	250.680	0.451
11	F2, x8	1.161	0.483	217.680	0.581	F1, x11	2.092	0.338	190.940	0.229	x6, x20	1.840	0.378	249.740	0.332
12	F2, x2	2.280	0.287	215.920	0.185	F2, x8	1.052	0.490	190.930	0.925	F1, x10	5.693	0.106	**244.640**	**0.024**

(*Continued*)

https://doi.org/10.1017/psrm.2025.27 Published online by Cambridge University Press

Table 1. (*Continued.*)

	***N* = 700**					***N* = 1000**					***N* = 2000**				
	Para.	B-Wald	*P*-value	LRT	*P*-value	Para.	B-Wald	*P*-value	LRT	*P*-value	Para.	B-Wald	*P*-value	LRT	*P*-value
1	**F2, x6**	**437.878**	**0.000**	**794.400**	**0.000**	**F2, x6**	**622.506**	**0.000**	**1231.510**	**0.000**	**F2, x6**	**1212.599**	**0.000**	**2087.210**	**0.000**
2	**F1, x9**	**183.251**	**0.000**	**561.740**	**0.000**	**F1, x9**	**312.393**	**0.000**	**752.360**	**0.000**	**F1, x9**	**598.461**	**0.000**	**1240.160**	**0.000**
3	**F2, x20**	**187.452**	**0.000**	**321.070**	**0.000**	**F2, x20**	**265.379**	**0.000**	**407.060**	**0.000**	**F2, x20**	**528.742**	**0.000**	**538.860**	**0.000**
4	F3, x6	1.177	0.484	319.520	0.214	F3, x6	1.257	0.455	404.750	0.129	F1, x20	1.541	0.427	538.500	0.552
5	F1, x20	1.102	0.470	319.460	0.800	**F3, x16**	**97.400**	**0.000**	**266.690**	**0.000**	**F3, x16**	**232.153**	**0.000**	**224.930**	**0.000**
6	F2, x8	3.063	0.240	**315.180**	**0.039**	F2, x8	1.369	0.460	266.590	0.761	F2, x8	1.190	0.460	224.840	0.755
7	**F3, x16**	**60.078**	**0.000**	**240.780**	**0.000**	F1, x20	1.125	0.477	266.590	0.929	F3, x6	1.196	0.479	224.750	0.766
8	x6, x20	2.224	0.326	239.580	0.273	x6, x20	3.020	0.270	264.710	0.171	x6, x20	1.003	0.495	224.650	0.751
9	F2, x4	1.181	0.454	239.500	0.787	F2, x5	1.188	0.471	264.710	0.940	F2, x3	1.028	0.487	224.050	0.438
10	F2, x3	1.630	0.416	238.640	0.353	F2, x3	1.042	0.484	264.330	0.537	F2, x5	1.995	0.374	224.020	0.887
11	F2, x5	1.055	0.480	238.400	0.624	F2, x4	1.862	0.402	263.740	0.445	F2, x1	0.946	0.510	222.330	0.192
12	x6, x8	1.829	0.401	237.680	0.397	F2, x1	1.403	0.447	263.600	0.702	F2, x4	9.140	0.059	**215.660**	**0.010**

Table 2. Monte Carlo simulation results for asymptotic properties

					NFI		CFI		RMSEA	
N	χ^2	SD	*P*-value	Rej. rate	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
100	275.650	25.824	0.184	0.386	0.748	0.836	0.922	1.000	0.000	0.059
150	263.874	22.901	0.277	0.205	0.829	0.885	0.960	1.000	0.000	0.042
200	259.340	24.503	0.332	0.167	0.866	0.913	0.970	1.000	0.000	0.036
250	255.329	22.978	0.371	0.115	0.894	0.929	0.978	1.000	0.000	0.031
300	253.984	23.378	0.394	0.109	0.911	0.939	0.982	1.000	0.000	0.029
400	251.213	22.787	0.424	0.095	0.933	0.954	0.988	1.000	0.000	0.023
500	249.332	22.811	0.442	0.078	0.946	0.963	0.991	1.000	0.000	0.020
800	247.594	22.136	0.467	0.063	0.966	0.976	0.994	1.000	0.000	0.015
1,000	247.450	22.108	0.466	0.052	0.973	0.981	0.996	1.000	0.000	0.014
2,000	247.235	22.054	0.468	0.062	0.986	0.990	0.998	1.000	0.000	0.010
5,000	246.822	22.195	0.477	0.062	0.994	0.996	0.999	1.000	0.000	0.006
10,000	245.462	22.410	0.495	0.054	0.997	0.998	1.000	1.000	0.000	0.004

relationships among factor loadings and residuals, the improved LM test continues to deliver outstanding performance.

Third, the magnitudes of factor loadings influence the performance of the improved LM test, and this is dependent on the sample size. When $N \geq 400$, the improved LM test delivers efficient and robust performance compared to the LRT. However, low factor loadings in smaller sample sizes tend to have a stronger impact on the detection of omitted variables and convergence. We found that when $N < 400$, the models encounter convergence issues, mainly because the covariance matrix becomes not positive definite. In contrast, with high factor loadings, both the improved LM test and LRT perform well across all sample sizes in this study. Nonetheless, the improved LM test consistently outperforms the LRT in detecting correct parameters. For the results of the simulation tests, please refer to the Appendix.

6. Empirical examples

6.1. Example 1: national identity and patriotism

In this study, we evaluate the effectiveness of the improved LM test using a covariance structure model constructed from Huddy and Khatib's (2007) student data, gathered in 2002. For detailed data collection and student sample information, please refer to page 66 in Huddy and Khatib (2007). This dataset comprises 341 respondents. The survey questions use a 4-point Likert scale, with response options ranging from "strongly approve" to "strongly disapprove." We employ the *diagonally weighted least squares* (DWLS) estimator to handle the ordered categorical variables. For brevity, the indicators and factors are unlabeled here; please refer to the appendix for the survey questions.

The path diagram for the three-factor model is depicted in Figure 5. This model involves a small sample size ($N = 341$) relative to a larger number of degrees of freedom ($df = 39$). The national identity and patriotism model consists of three latent factors, F_1, F_2, and F_3. These factors represent the constructs of *national identity, symbolic patriotism*, and *uncritical patriotism*, respectively. Each of these constructs is assessed by a series of indicators X_i. Factor loadings are denoted by λ_i, residuals by ε_i, and the residuals of factor by D_i. The coefficients β_1, β_2, and β_3 measure the correlations between the three factors. To evaluate the performance of the improved LM test, we follow the same procedure as employed in the simulated data. The dashed lines in Figure 5 represent the recommended parameters suggested by the improved LM test.

Table 3 displays the outcomes of three distinct tests: the univariate LM test, bootstrap LM test, and bootstrap Wald test. The bootstrap LM test, executed through a forward stepwise approach, identified five missing parameters that were statistically significant. The highlighted items indicate the actual

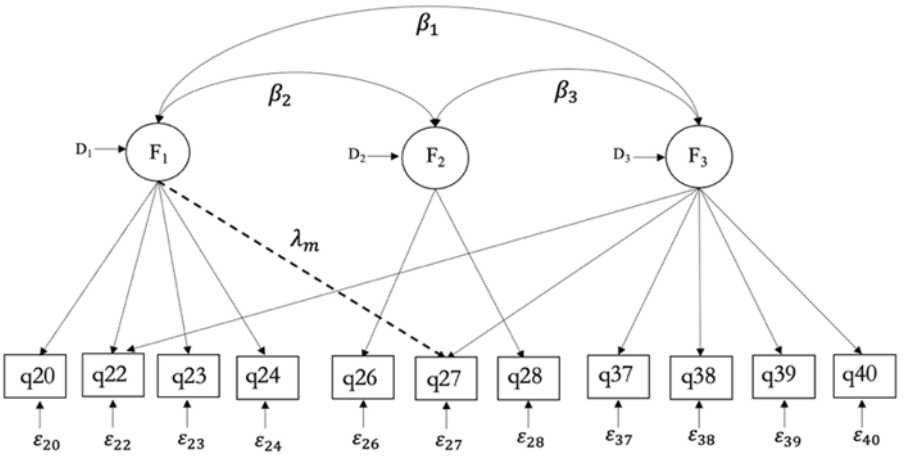

Figure 5. Path diagram of national identity and patriotism (Huddy and Khatib, 2007).

Table 3. Summary of example 1 test statistics

		LM test		Bootstrap LM test		Bootstrap W test	
	Parameters	LM χ^2	*P*-values	LM χ^2	*P*-values	Chi-square	*P*-values
1	F2, q27	22.291	0.000	**14.221**	**0.004**	5.948	0.081
2	**F1, q27**	19.856	0.000	**13.102**	**0.005**	**7.628**	**0.041**
3	q27, q38	10.340	0.001	**6.761**	**0.023**	1.554	0.392
4	q37, q38	10.311	0.001	**6.670**	**0.035**	0.497	0.601
5	F1, q40	9.749	0.002	**6.829**	**0.038**	2.206	0.225
6	q40, q38	7.495	0.006	4.982	0.068		
7	q40, q39	6.701	0.010	5.166	0.106		
8	F2, q37	6.247	0.012	4.240	0.077		
9	q37, q39	6.231	0.013	4.423	0.109		
10	q27, q39	5.663	0.017	3.808	0.088		

missing parameters, and the items in bold in the bootstrap LM and Wald tests are statistically significant. The bootstrap Wald test concurred that the parameter λ_m (the factor loading linking F_1 and q27) should be included in the original model to improve the model fit. q27 inquires, "How angry does it make you feel, if at all, when you hear someone criticizing the United States?" Response options range from extremely angry to not at all. Furthermore, the standardized factor loading of λ_m is 0.47 and statistically significant. This indicates that differing levels of national identity (F_1) and uncritical patriotism (F_3) are likely to influence feelings of anger. While the addition of this parameter may not alter the overall substantive conclusion, it provides new insights into the nuances of these latent structural relationships. However, without including this parameter, Huddy and Khatib's (2007) original model suffers from some degree of misspecification.

Table 4 presents the results of our replication study based on Huddy and Khatib's (2007) research using the DWLS estimator. In the original study, the χ^2 statistic is 65.176 with 40 degrees of freedom, resulting in a *p*-value of 0.007. The CFI is 0.997, NFI is 0.992, TLI is 0.996, and the RMSEA is 0.043. The χ^2 test statistic provides limited support for the substantive argument. However, upon introducing the omitted parameter, λ_m, as suggested by the improved LM test, the χ^2 test statistic reduces to 27.534, resulting in a *p*-value of 0.532. The improvement in the χ^2 test statistic is crucial as it suggests that the model-implied covariance structure is highly consistent with the sample covariance structure. Moreover, the NFI increases to 0.996, the CFI and TLI increase to 1.0, and the RMSEA decreases to 0. The final column in Table 4 indicates the differences in test statistics and fit indices between the

Table 4. Comparison of test statistics and model fit in example 1

	Original model	Improved LM test	Differences
Chi-square	65.176	37.642	27.534
Degrees of freedom	40	39	1
P-value	0.007	0.532	−0.525
NFI	0.992	0.996	−0.004
CFI	0.997	1.000	−0.003
TLI	0.996	1.000	−0.004
RMSEA	0.043	0.000	0.043

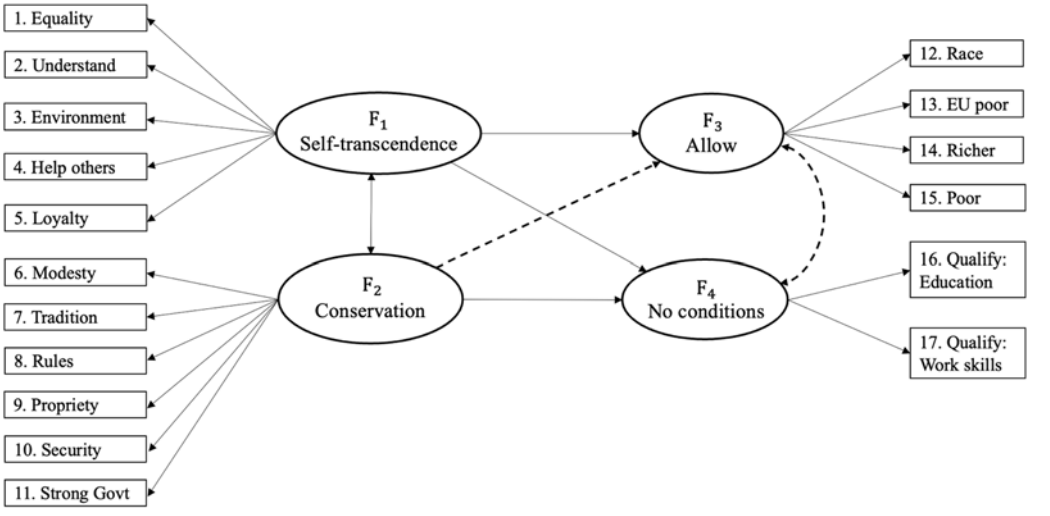

Figure 6. SEM of human value priorities (Davidov, 2009; Oberski, 2014).
Note: Error and factor variances are not shown in the path diagram.

two models, showcasing a significant enhancement in model fit and strengthening the theoretical argument based on this structural relationship.

6.2. Example 2: SEM of relationship of human value priorities

In another empirical application, we conducted an analysis of the German sample ($N = 2919$) from the 2002 European Social Survey, which is a cross-national probability survey. To illustrate the effectiveness of our improved LM test in small sample sizes, we randomly selected 500 observations from the German sample. Detailed information on data collection procedures and original survey questions can be found on the ESS website.

To analyze the data, we employed a four-factor SEM model, as depicted in Figure 6, following standard practice. The model comprises four latent factors represented by ovals, each measured by multiple indicators. In the original model, factor F_3 is predicted by factors F_1 and F_2, while factor F_4 is predicted by factors F_1 and F_2. Furthermore, there are correlations between factors F_1 and F_2, as well as F_3 and F_4. To evaluate the effectiveness of our proposed improved LM test approach, we removed the coefficient parameters between F_2 and F_3, and between F_3 and F_4, as indicated by the dash lines in Figure 6.

Table 5 presents the results of three tests conducted on the model: the LM test, bootstrap LM test, and bootstrap Wald test. The LM test identified the top 10 omitted parameters based on LM test statistics, out of which 2 parameters (F_2, F_3) and (F_3, F_4) were the actual missing parameters. The improved LM test confirmed the validity of these two parameters and identified four additional parameters

Table 5. Summary of example 2 test statistics

		LM test		Bootstrap LM test		Bootstrap Wald test	
	Parameters	LM χ^2	*P*-values	LM χ^2	*P*-values	Chi-square	*P*-values
1	**F1, F3**	64.563	0.000	**63.523**	**0.000**	**33.833**	**0.000**
2	**F2, F3**	64.563	0.000	**63.523**	**0.000**	**14.780**	**0.004**
3	**Rules, Propriety**	25.415	0.000	**25.920**	**0.002**	**23.352**	**0.001**
4	**F3, F4**	23.890	0.000	**23.690**	**0.000**	**10.550**	**0.038**
5	**F3, Understand**	19.611	0.000	**20.327**	**0.003**	**13.041**	**0.023**
6	**Equality, Tradition**	14.447	0.000	**15.342**	**0.007**	**13.599**	**0.012**
7	**F2, Richer**	13.483	0.000	**13.698**	**0.012**	**10.895**	**0.020**
8	F1, Rules	12.185	0.000	**13.063**	**0.015**	1.180	0.461
9	Environment, Tradition	11.458	0.001	**12.283**	**0.026**	6.874	0.103
10	F2, Help others	11.245	0.001	**13.235**	**0.015**	6.112	0.103

Table 6. Comparisons of test statistics and fit indices

	Original model	Modified model	Difference
Chi-square	324.224	167.902	156.322
DF	115	109	6
P-value	0.000	0.000	0
NFI	0.844	0.919	−0.075
TLI	0.872	0.962	−0.090
CFI	0.892	0.970	−0.078
RMSEA	0.064	0.035	0.029

(highlighted in bold in the first column in Table 5) from the ten tested in the model. Note that (F_1, F_3) is meaningless here because the original research aims to use a unidirectional arrow to indicate the effect of F_1 on F_3 based on their theoretical argument, while the improved LM test suggests a correlation instead. Thus, we disregard this suggested omitted parameter. The suggested parameter (F3, Understand) suggests that attitudes toward immigration may also be influenced by the universalism value, which emphasizes understanding and concern for the welfare of all people, as well as the influence of self-transcendence. Additionally, the suggested parameter (Equality, Tradition) indicates that the belief in treating every person equally is correlated with the value placed on tradition. Naturally, this relationship may vary significantly across different countries and cultures. Such variations could introduce measurement invariance and model misspecification issues when comparing the effects of values on attitudes toward immigration without accounting for these structural relationships.

To determine whether adding these suggested parameters could improve the model fit, we compared the test statistics and fit indices. Table 6 reports that the original model's χ^2 test is 324.224 with 115 degrees of freedom. However, when we added all the suggested parameters to the model, the χ^2 test decreased to 167.902, a reduction of 156.322, with a loss of only 6 degrees of freedom. Additionally, the NFI increased from 0.844 to 0.919, TLI increased from 0.872 to 0.962, and the CFI increased from 0.892 to 0.970, and the RMSEA decreased from 0.064 to 0.035. All these statistics confirmed that the improved LM test method yielded favorable results for this model by incorporating the additional parameters it recommended.

In summary, the improved LM test effectively identified omitted parameters in empirical examples 1 and 2. Incorporating these parameters, as shown in Tables 4 and 6, substantially improves the overall model fit in the χ^2 test statistics and fit indices. While these additions may not alter the substantive conclusions, they enhance confidence in the authors' arguments and provide new insights into their theoretical claims. It is important to note, however, that although the improved LM test is a valuable data-driven method for uncovering hidden parameters, the decision to include suggested parameters should be guided by strong theoretical justification.

7. Conclusion

CSA and SEM stand as formidable tools that enjoy wide adoption in the behavioral and social sciences, facilitating the understanding of latent structural relationships among variables. Nevertheless, the creation of an accurate model proves challenging, and conventional practices occasionally engender the perils of chance-based model risks occur when different model fits arise from different samples of the same population, influenced by factors such as sample size, model complexity, and degrees of freedom. Significant chance-based model risks imply a higher likelihood of overlooking pathways within the model and undue rejection of the null hypothesis. To surmount these predicaments, the present study proposed an improved LM test, designed to rectify instances of falsely statistically significant parameters and to effectively pinpoint omitted parameters, particularly in scenarios featuring modest sample sizes. The improved LM test integrates bootstrap LM and Wald tests, enhancing model specification searches by accurately identifying missing parameters. This robust framework advances the field, enabling researchers to effectively model complex phenomena and make well-informed decisions based on well-specified models.

Though our investigation predominantly centers on a model boasting a substantial number of degrees of freedom, it is reasonable to anticipate that our approach will likewise prove efficacious for models featuring fewer degrees of freedom. This expectation stems from the recognition that a model with fewer degrees of freedom reduces the likelihood of succumbing to the perils of chance-based model risks, which occur when different model fits arise from different samples of the same population, influenced by factors such as sample size, model complexity, and degrees of freedom. Significant chance-based model risks imply a higher likelihood of overlooking pathways within the model. Our confidence in the applicability of this approach is reinforced by replicating empirical examples, such as Huddy and Khatib's (2007) model of national identity and patriotism, which involved a small sample size relative to a larger number of degrees of freedom. Similarly, Davidov's (2009) and Oberski's (2014) SEM models examining the relationship of human value priorities exhibit a moderate sample size accompanied by a comparatively greater number of degrees of freedom.

Our simulations, which varied the magnitudes of factor loadings, factor correlations, and the number of indicators per factor, consistently showed that our proposed improved LM test performs noticeably better than the LRT. These extensive evaluations under various realistic scenarios provided further insight into the application and effectiveness of the improved LM test. Nevertheless, the simulation study conducted should not be considered as a comprehensive evaluation encompassing a broad spectrum of realistic conditions. Similarly, conducting a systematic comparison of the practical utility of the improved LM test with other testing methods across diverse topics, such as nonnormal data, varying levels of model complexity, degrees of misspecification, and so on, was beyond the scope of this specific study. However, we remain confident in the contributions made by this research. Future studies could further explore the potential applications and comparative effectiveness of the improved LM test. Additionally, this test is not limited to CSA and SEM; future research could expand its use to regression and other domains, offering broader applicability for applied researchers.

Lastly, as numerous scholars have rightly emphasized, researchers bear the crucial responsibility of interpreting the results yielded by any proposed approach with caution, ensuring they remain aligned with substantive theory (MacCallum *et al.*, 1992; Bentler 2006). The improved LM test is no exception. It is vital to recognize that the results of the improved LM test should be considered merely as a suggestion for including statistically indispensable parameters. The decision to integrate these parameters into a modified model should be guided by a solid theoretical foundation.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psrm.2025.27. To obtain replication material for this article, https://doi.org/10.7910/DVN/W77NEA.

Statements and declarations. This manuscript has not been submitted for review elsewhere. The research was conducted in adherence to the Ethical Principles of *Political Science Research and Methods* and Code of Conduct, following all appropriate

ethical guidelines. No funding was received for the implementation of this study. The authors have no relevant financial or non-financial interests to declare.

References

Acock A, Clarke HD and Stewart MC (1985) A new model for old measures: A covariance structure analysis of political efficacy. *Journal of Politics* **47**(4), 1062–1084. https://doi.org/10.2307/2130807.

Arruda EH and Bentler P (2017) A regularized GLS for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal* **24**, 657–665.

Bentler P (2006) *EQS 6 structural equations program manual. Multivariate Software, Inc.*

Bentler PM (1986) *Lagrange Multiplier and Wald Tests for EQS and EQS/PC. BMDP Statistical Software: Los Angeles. EQS 6 Structural Equation Program Manual*. Encino, CA: Multivariate Software, Inc.

Bentler PM and Chou C-P (1992) Some new covariance structure model improvement statistics. *Sociological Methods & Research* **21**(2), 259–282. https://doi.org/10.1177/0049124192021002006.

Bentler PM and Dijkstra T (1985) Efficient estimation via linearization in structural models. In Kirshnaiah PR (ed.), *Multivariate Analysis VI*. Amsterdam: North-Holland, pp. 9–42.

Blackwell M, Honaker J and King G (2017) A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods & Research* **46**(3), 303–341. https://doi.org/10.1177/0049124115585360.

Browne M (1984) Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology* **37**(1), 62–83. https://doi.org/10.1111/bmsp.1984.37.

Chou C-P and Huh J (2012) Model modification in structural equation modeling. In Hoyle RH (ed.), *Handbook of Structural Equation Modeling*. New York, NY: The Guilford Press, pp. 232–246.

Davidov E (2009) Measurement equivalence of nationalism and constructive patriotism in the ISSP: 34 countries in a comparative perspectives. *Political Analysis* **17**, 64–82.

DeSante CD and Smith CW (2022) Fear, institutionalized racism, and empathy: The underlying dimensions of whites' racial attitudes. *PS: Political Science and Politics* **53**(4), 639–645. https://doi.org/10.1017/S1049096520000414.

Feldman S (1988) Structure and consistency in public opinion: The role of core beliefs and values. *American Journal of Political Science* **32**(2), 416–440.

Florax RJGM, Folmer H and Rey SJ (2003) Specification searches in spatial econometrics: The relevance of Hendry's methodology. *Regional Science and Urban Economics* **33**, 557–579.

Goren P (2005) Party identification and core political values. *American Journal of Political Science* **49**(4), 882–897.

Hayakawa K (2019) Corrected goodness-of-fit test in covariance structure analysis. *Psychological Methods* **24**(3), 371–389.

Huddy L and Khatib N (2007) American patriotism, national identity, and political involvement. *American Journal of Political Science* **51**(1), 63–77.

Jöreskog KG (1969) Some contribution to maximum likelihood factor analysis. *Psychometrika* **34**, 183–202.

Jöreskog KG and Sörbom D (1988) *LISREL 7, A Guide to the Program and Applications*. SPSS: Chicago.

Kaplan D (1988) The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research* **23**, 69–86.

Leamer EE (1978) *Specification Searches: Ad Hoc Inference with Non-experimental Data*. Wiley: New York.

Lee S-Y and Bentler P (1980) Some asymptotic properties of constrained generalized least squares estimation in covariance structure models. *South African Statistical Journal* **14**, 121–136.

MacCallum RC, Roznowski M and Nectowitz LB (1992) Model modifications covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin* **111**(3), 490–504.

Mur J and Angulo A (2009) Model selection strategies in a spatial setting: Some additional results. *Regional Science and Urban Economics* **39**, 200–213.

Oberski DL (2014) Evaluating sensitivity of parameters of interest to measurement invariance in latent variable models. *Political Analysis* **22**, 45–60. https://doi.org/10.1093/pan/mpt014.

Pietryka MT and MacIntosh RC (2013) An analysis of ANES items and their use in the construction of political knowledge scales. *Political Analysis* **21**, 407–429.

Rosseel Y (2012) Lavaan: An R package for structural equation modeling. *Journal of Statistical Software* **48**(2), 1–36.

Satorra A (1989) Alternative test criteria in covariance structure analysis: A unified approach. *Psychometrika* **54**, 131–151.

Sörbom D (1989) Model modification. *Psychometrika* **54**, 371–384.

Sullivan J, Marcus G, Feldman S and Piereson J (1981) The source of political tolerance: A multivariate analysis. *The American Political Science Review* **75**(1), 92–106. https://doi.org/10.2307/1962161.

Yuan KH, Hayashi K and Yanagihara H (2007) A class of population covariance matrices in the bootstrap approach to covariance structure analysis. *Multivariate Behavioral Research* **42**(2), 261–281. https://doi.org/10.1080/00273170701360662.

Yuan KH and Liu F (2021) Which method is more reliable in performing model modification: Lasso regularization or Lagrange multiplier test? *Structural Equation Modeling: A Multidisciplinary Journal* **28**(1), 69–81.

Yuan KH, Marshall LL and Peter MB (2003) Assessing the effect of model misspecifications on parameter estimates in structural equation models. *Sociological Methodology* **33**, 241–265.

Zheng BQ (2023) Asian and Latino American political conceptualization: A dual-concept model. *American Politics Research* **51**(2), 182–196. https://doi.org/10.1177/1532673X221132479.

Zheng BQ and Bentler PM (2021) Testing mean and covariance structures with reweighted least squares. *Structural Equation Modeling: A Multidisciplinary Journal*. https://doi.org/10.1080/10705511.2021.1977649.

Zheng BQ and Bentler PM (2023) RGLS and RLS in covariance structure analysis. *Structural Equation Modeling: A Multidisciplinary Journal* **30**(2), 234–244. https://doi.org/10.1080/10705511.2022.2117182.

Zheng BQ and Bentler PM (2024) Enhancing model fit evaluation in SEM: Practical tips for optimizing chi-square tests. *Structural Equation Modeling: A Multidisciplinary Journal* Advance online publication. https://doi.org/10.1080/10705511.2024.2354802.

Cite this article: Zheng BQ and Bentler PM (2025) Improved LM test for robust model specification searches in covariance structure analysis: application in political science research. *Political Science Research and Methods*, 1–21. https://doi.org/10.1017/psrm.2025.27