Behaviormetrika (2020) 47:243–272
https://doi.org/10.1007/s41237-019-00098-0

ORIGINAL PAPER

An empirical comparison of generalized structured component analysis and partial least squares path modeling under variance-based structural equation models

Gyeongcheol Cho[1] · Ji Yeh Choi[2]

Received: 26 April 2019 / Accepted: 15 October 2019 / Published online: 9 November 2019

Abstract

Generalized structured component analysis (GSCA) and partial least squares path modeling (PLSPM) are component-based, or also called variance-based, structural equation modeling (SEM). They define latent variables as components or weighted composites of indicators, attempting to maximize the explained variances of indicators or endogenous components or both. Despite this common conceptualization of latent variables, GSCA and PLSPM involve distinct model specifications and estimation procedures. This paper focuses on comparing four modeling approaches—GSCA *with reflective indicators*, GSCA *with formative indicators*, PLSPM *with mode A*, and PLSPM *with mode B*—regarding their capability of parameter recovery and statistical power via Monte Carlo simulation. For comparison, we propose a new data generating process for variance-based SEM, appropriate to handle all possible modeling approaches for both GSCA and PLSPM. It was found that although every approach produced consistent estimators, GSCA *with reflective indicators* yielded the most efficient estimators under variance-based structural equation models.

Keywords Generalized structured component analysis · Partial least squares path modeling · Data generating process for variance-based structural equation models · Comparison · Simulation

Communicated by Heungsun Hwang.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s41237-019-00098-0) contains supplementary material, which is available to authorized users.

✉ Gyeongcheol Cho
gyeongcheol.cho@mail.mcgill.ca

[1] Department of Psychology, McGill University, 2001 McGill College Avenue, Montreal, QC H3A 1G1, Canada

[2] Department of Psychology, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada

1 Introduction

Generalized structured component analysis (GSCA) and partial least squares path modeling (PLSPM) are two full-fledged approaches to component-based structural equation modeling (SEM) (Hwang and Takane 2014; Tenenhaus 2008). In component-based SEM, latent variables are conceptualized as components or weighted composites of indicators. These components are constructed to maximize the explained variances of either their indicators or endogenous components, as in principal component or canonical correlation analysis, or both. This is a key difference from factor-based SEM (i.e. covariance structural analysis (CSA) proposed by Jöreskog 1970, 1978), where latent variables are defined as factors to best explain the covariances of their indicators. Accordingly, component-based SEM are also called variance-based SEM, whereas factor-based SEM are called covariance-based SEM at times (Reinartz et al. 2009; Roldán and Sánchez-Franco 2012).

Although both GSCA and PLSPM fall within variance-based SEM, they involve different model specifications and estimation procedures. GSCA specifies three sub-models—weighted relation, structural, and measurement models—and derives a single optimization function unifying the sub-models. It allows for constructing two different modeling approaches, (i.e., GSCA *with reflective indicators* and GSCA with *formative indicators*) within one general modeling framework. On the other hand, PLSPM does not utilize a single objective function: each of the possible two modeling approaches (i.e., PLSPM *with mode A* and *mode B*) just modifies the specification of measurement model at a time.[1] Accordingly, unlike GSCA that uses a full information method with a global optimization function, PLSPM employs a limited information estimation method (Tenenhaus 2008). Despite such differences in model specification and estimation procedure, conceptually, PLSPM *with mode A* is compatible with GSCA *with reflective indicators*, in which the weight parameters are estimated to maximize the explained variances of indicators, as in principal component analysis (Hwang et al. 2015; Reinartz et al. 2009), as well as those of endogenous components. PLSPM *with mode B* is regarded as a counterpart of GSCA *with formative indicators*, explaining the variances of endogenous components only as much as possible, like canonical correlation analysis (Dijkstra 2017; Hwang et al. 2015).

In the literature, several simulation studies have assessed relative performances of GSCA and PLSPM under various simulation conditions. In Hwang et al. (2010), a simulation study was conducted to investigate the performance of GSCA *with reflective indicators* and PLSPM *with mode A*, varying sample sizes, data distributions, and model specifications. They found that the performance between the two approaches was similar when a model was specified without any cross-loadings. In other conditions where cross-loadings were specified, on average, GSCA

[1] Both GSCA and PLSPM may take different modeling approaches in which their two modeling approaches are combined (i.e. GSCA *with both reflective indicators and formative indicators* and PLSPM *with mode C*), but, for simplicity, we do not handle those mixed types of modeling approaches in this study.

with reflective indicators outperformed PLSPM *with mode A* (also see Hwang and Takane 2014, Chapter 2). However, the data used for this simulation study were generated under the assumption of covariance-based structural equation models, rendering it difficult to evaluate the comparative performance of variance-based SEM approaches (Hwang et al. 2010; Reinartz et al. 2009). Hence, it becomes necessary to develop a new data generating process (DGP) appropriate for variance-based SEM approaches.

Recently, a team of PLSPM researchers suggested a DGP for variance-based SEM (Becker et al. 2013; Dijkstra 2017) and evaluated the relative performance of GSCA *with formative indicators*, PLSPM *with mode B*, and sum-scores regression (i.e. a component's scores are calculated by simply summing scores of its indicators) in terms of parameter recovery and statistical power (Hair et al. 2017). In Hair et al. (2017)'s study, GSCA *with formative indicators* and PLSPM *with mode B* turned out to produce consistent estimators under their DGP and performed better than the sum-scores regression. Between the two variance-based SEM approaches, GSCA *with formative indicators* provided more accurate estimates for the weight parameters than PLSPM *with mode B*, whereas they performed similarly in estimating path coefficients except for small sample sizes ($N = 100$). For small sample sizes, PLSPM *with mode B* was slightly better than GSCA *with formative indicators*, and the power of PLSPM *with mode B* was larger than that of GSCA *with formative indicators*.

Hair et al. (2017) made a meaningful contribution to variance-based SEM in that they initially evaluated the comparative performance of GSCA and PLSPM under structural equation models with components. Nevertheless, their study had limitations in (1) the range of the modeling approaches of GSCA and PLSPM considered and (2) the DGP they used for their simulation. First, Hair et al. (2017), compared GSCA *with formative indicators* and PLSPM *with mode B* only, although both GSCA and PLSPM could take other modeling approaches, as discussed earlier. Second, as will be explicated in Sect. 3, the components under their DGP were constructed with a set of arbitrarily chosen values for the weight parameters, rather than deriving the weight parameter values while considering the covariances of their indicators. Accordingly, these components may not capture the variances of both indicators and endogenous components well, and thus their DGP is hard to serve as a standard for evaluating all the variance-based SEM approaches. It is, therefore, required to develop a new DGP suited for variance-based SEM and to further assess the relative performance of all possible modeling approaches for both GSCA and PLSPM.

In this paper, we propose a new DGP for variance-based structural equation models with components that maximize the explained variances of their indicators and endogenous components. Under these structural equation models, all of the representatives for variance-based SEM—GSCA *with reflective indicators*, GSCA *with formative indicators*, PLSPM *with mode A*, and PLSPM *with mode B*—are evaluated using a Monte Carlo simulation.

The remainder of this article is organized as follows: In Sect. 2, we briefly review GSCA and PLSPM with respect to model specification and estimation process. In Sect. 3, a new data generating process for variance-based structural equation models is proposed. Its characteristics relative to that of the previous DGP are also discussed in detail. In Sect. 4, we report the design and results of our simulation. In the final section, the simulation results are summarized and their implications are discussed.

2 Overview of GSCA and PLSPM

GSCA and PLSPM have distinct model specifications. GSCA specifies three sub-models—weighted relation, structural, and measurement models. Let $\mathbf{z} = \left[\mathrm{z}_j\right] \in \mathbb{R}^{J\times 1}$ denote the vector of observed variables or indicators where z_j is the jth indicator and J is the number of indicators. Let $\boldsymbol{\gamma} = \left[\gamma_p\right] \in \mathbb{R}^{P\times 1}$ denote the vector of latent variable or component, where γ_p is the pth component and P is the number of components. Both indicators and components are assumed to be standardized (i.e. $\mathrm{Var}(\mathrm{z}_j)=\mathrm{Var}(\gamma_p)=1$). Let $\mathbf{W} = \left[\mathrm{w}_{j,p}\right] \in \mathbb{R}^{J\times P}$ denote the component weight matrix, where $\mathrm{w}_{j,p}$ is the weight assigned to the jth indicator to construct the pth component. Let $\mathbf{C} = \left[\mathrm{c}_{p,j}\right] \in \mathbb{R}^{P\times J}$ denote the loading matrix where $\mathrm{c}_{p,j}$ relates the pth component to the jth indicator. Let $\mathbf{B} = \left[\mathrm{b}_{p*,p}\right] \in \mathbb{R}^{P\times P}$ denote the path coefficient matrix where $\mathrm{b}_{p*,p}$ denotes the effect of the p*th component on the pth component. Let $\boldsymbol{\varepsilon} = [\varepsilon_p] \in \mathbb{R}^{P\times 1}$ denote a residual vector where ε_p is the residual for the pth component. Let $\mathbf{e} = \left[\mathrm{e}_j\right] \in \mathbb{R}^{J\times 1}$ denote a residual vector where e_j is the residual for the jth indicator. The weighted relation, measurement, and structural models can be generally expressed as

$$\boldsymbol{\gamma} = \mathbf{W}'\mathbf{z} \tag{1}$$

$$\boldsymbol{\gamma} = \mathbf{B}'\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{2}$$

$$\mathbf{z} = \mathbf{C}'\boldsymbol{\gamma} + \mathbf{e}. \tag{3}$$

In the weighted relation model (1), each latent variable or component is defined as a weighted composite of some observed variables, and the observed variables are considered indicators of the component. This sub-model identifies GSCA as a component-based SEM. The weight assigned to each indicator is determined for its component(s) to well explain the relations among the variables specified in the other two sub-models, (2) and (3). The structural model (2) specifies a series of directional relations among the components, whereas the relations between the components and their indicators are specified in the measurement model (3). Specifically, the measurement model (3) sets some indicators to be explained by their components, which, in effect, may allow the components to capture variances of the indicators better in estimation process of weight parameter. This type of indicator is called ‘reflective indicator’, whereas the indicator whose relation with its component is specified in the weighted relation model (1) only is called ‘formative indicator’. We named the GSCA modeling approach with reflective indicators only GSCA *with reflective indicators* and the one with formative indicators only GSCA *with formative indicators* for the sake of comparison between PLSPM *with mode A* and PLSPM *with mode B*. Note that GSCA *with formative indicators* virtually specifies two sub-models—(1) the weighted relation model and (2) the structural model.

Unlike GSCA, PLSPM specifies two sub-models only—structural (or inner) and measurement (or outer) model, the latter of which are differently specified in PLSPM *with mode A* and PLSPM *with mode B* (Tenenhaus et al. 2005). In the PLSPM with *mode A*, two sub-models can be written as follows:

$$\gamma_p = \sum_{p^*=1,p^*\neq p}^{P_p} \mathrm{b}_{p^*,p}\gamma_{p^*,p} + \varepsilon_p \quad \forall p \in \{1, 2, \ldots, P\} \tag{4}$$

$$\mathrm{z}_j = \mathrm{c}_{p,j}\gamma_p + \mathrm{e}_j \quad \forall j \in \{1, 2, \ldots, J\}, \tag{5}$$

where P_p is the number of the independent components for the pth component, γ_p is the pth component, and $\gamma_{p^*,p}$ is the p*th independent component for the pth components. The measurement model used in PLSPM *with mode A* is called 'reflective measurement model' or 'outwards directed model'. As in the GSCA sub-models, the structural and reflective measurement models of PLSPM *with mode A* specify the relations among components and between components and their indicators, respectively. They can also be re-expressed with the matrix notations for GSCA structural and measurement models in the same way. In this case, however, the path coefficient matrix, **B**, is a triangular matrix, implying that PLSPM does not allow reciprocal relations among components unless an additional statistical technique such as instrumental variables are utilized. For the loading matrix, **C**, zero constraints are imposed on all the off-diagonal blocks of its entries, which means that indicators can be explained by one component only in PLSPM *with mode A*. When cross-loadings and reciprocal relations are not specified, the structural and measurement models of PLSPM *with mode A* are equivalent to those of GSCA *with reflective indicators*.

On the other hand, while specifying the same structural model, PLSPM *with mode B* specifies different measurement model, called 'formative measurement model' or 'inwards directed model.' The formative measurement model can be expressed as

$$\gamma_p = \sum_{j=1}^{J_p} \mathrm{w}_{j,p^*}\mathrm{z}_j + \zeta_p \quad \forall p \in \{1, 2, \ldots, P\}, \tag{6}$$

where J_p is the number of indicators for the pth component, w_{j,p^*} is the formative weight assigned to the jth indicator for the pth component, and ζ_p denote a residual for the pth component in formative measurement model. It implies that each component is defined by the weighted sum of its indicators as in (1) but with additional formative measurement errors: ζ_p.

PLSPM is distinct from GSCA in that, regardless of its mode, PLSPM does not define latent variables as weighted composites of their indicators in model specification (Hwang et al. 2019; Lohmöller 1989). In the estimation process, however, PLSPM always computes component scores as if they specified the weight relation model (3) of GSCA. More specifically, in PLSPM *with mode A*, a score for a component is computed as a weighted sum of scores for the indicators which are specified to be affected by the component in the reflective measurement model (5). PLSPM *with mode B* computes component scores based on the specified formative measurement model (6) but with the values of formative measurement errors excluded. Consequently, it is reasonable to think that both PLSPM *with mode A* and PLSPM *with mode B* implicitly assume the same type of sub-model as the weight relation model (3) of GSCA, as follows:

$$\gamma_p = \sum_{j=1}^{J_p} \mathrm{w}_{j,p}\mathrm{z}_j \quad \forall p \in \{1, 2, \ldots, P\}, \tag{7}$$

where $\mathrm{w}_{j,p}$ corresponds to $\mathrm{c}_{p,j}$ for PLSPM *with mode A* and to w_{j,p^*} for PLSPM *with mode B*. In this respect, PLSPM has been classified as a component-based SEM with GSCA, and the models of PLSPM *with mode A* corresponds to those of GSCA *with reflective indicators*, while those of PLSPM *with mode B* does to those of GSCA *with formative indicators*.

Another distinction lies in the unification of model equations. Combining models from (1) to (3), GSCA builds a unified model as follows:

$$\begin{aligned}
\begin{bmatrix}\mathbf{z}\\ \boldsymbol{\gamma}\end{bmatrix} &= \begin{bmatrix}\mathbf{C}'\\ \mathbf{B}'\end{bmatrix}\boldsymbol{\gamma} + \begin{bmatrix}\mathbf{e}\\ \boldsymbol{\varepsilon}\end{bmatrix}\\
\begin{bmatrix}\mathbf{I}\\ \mathbf{W}'\end{bmatrix}\mathbf{z} &= \begin{bmatrix}\mathbf{C}'\\ \mathbf{B}'\end{bmatrix}\mathbf{W}'\mathbf{z} + \begin{bmatrix}\mathbf{e}\\ \boldsymbol{\varepsilon}\end{bmatrix}\\
\mathbf{V}'\mathbf{z} &= \mathbf{A}'\mathbf{W}'\mathbf{z} + \mathbf{r}
\end{aligned} \tag{8}$$

where $\mathbf{V}' = \begin{bmatrix}\mathbf{I}\\ \mathbf{W}'\end{bmatrix}$, $\mathbf{A}' = \begin{bmatrix}\mathbf{C}'\\ \mathbf{B}'\end{bmatrix}$, and $\mathbf{r} = \begin{bmatrix}\mathbf{e}\\ \boldsymbol{\varepsilon}\end{bmatrix}$. In contrast, PLSPM does not integrate its sub-model equations (i.e. (4), (5) for PLSPM *with mode A* and (4), (6) for PLSPM *with mode B*) into one single equation and just leaves them as they are specified for each dependent variable. As will be explained below, this difference leads to the selection of different estimation process for each approach.

For parameter estimation, GSCA employs the full information estimation method owing to its global optimization function. The unification of sub-model Eqs. (8) allows GSCA to use the following global optimization function:

$$\varphi = \sum_{i=1}^{N} \mathbf{r}_i'\mathbf{r}_i = \sum_{i=1}^{N} (\mathbf{V}'\mathbf{z}_i - \mathbf{A}'\mathbf{W}'\mathbf{z}_i)'(\mathbf{V}'\mathbf{z}_i - \mathbf{A}'\mathbf{W}'\mathbf{z}_i), \tag{9}$$

where $\mathbf{z}_i$ denote a J by 1 vector of indicators for the ith observations ($i = 1, 2, \ldots, N$), and $\mathbf{r}_i$ denote the residuals for the ith observations. This function is equivalent to the sum of squared residuals for all the equations in GSCA model. Finding the values that minimize this function amounts to estimating parameters that maximize the explained variances of indicators and endogenous components. GSCA estimates the entire entries of **A** (i.e. loadings and path coefficients for GSCA *with reflective indicators* and path coefficients only for GSCA *with formative indicators*) and **W** (i.e. weights) alternatively and concurrently to minimize the optimization function using alternative least squares (ALS) algorithm. The alternating procedure continues until the value of the optimization function does not decrease more than the pre-determined convergence criterion. A detailed description of this algorithm can be found in Hwang and Takane (2014). In the full information method, estimation proceeds for the entire system of equations, thereby utilizing all the information from every equation. Accordingly, estimators of full information methods can be more efficient under correct model specification with sufficient sample (Bollen 1996; Fomby et al. 2011; Gerbing and Hamilton 1994).

Conversely, PLSPM does not have a single optimization criterion to be minimized and consequently relies on the limited information estimation method whereby parameters for each equation are estimated separately based solely on the information specific to the equation (Fomby et al. 2011, Chapter 22). In addition, the estimation process of PLSPM is utterly segregated and sequential. At the first stage, weights are estimated by the iterative process of two steps. Given the random initial values of component scores, component scores are updated at the first step as the weighted sum of the other components specified in (4), which is called inner estimates for components. With these inner estimates for components, weights are estimated at the second step in two different manners—*mode A* and *mode B* (Lohmöller 1989). Under *mode A*, weights are estimated to regress indicators on a component with the specified relations in (5), while the loading relating the *p*th component to the *j*th indicator, $c_{p,j}$, is considered equivalent to the weight of the *j*th indicator to construct the *p*th component, $w_{j,p}$. Because of the basic design setting that each indictor is explained by one component only, weights become the correlation between a component and its indicators. On the other hand, *Mode B* estimates weights by regressing a component on its indicators using the formative relations specified in (6). Then, using the weight estimates obtained from either *mode A* or m*ode B*, component scores are computed as weighted sums of each set of indicators and standardized so that the variance of each component is equivalent to 1. These component score estimates are called outer estimates. Given the outer estimates for components, the 1st step proceeds again. These two steps iterate and stop when the estimates do not alter more than the convergence criterion. With the final component score estimates and specified relations in (4) and (5), path coefficients and loadings are estimated by ordinary least squares at the second stage. Note that, even in PLSPM *with mode B*, loadings are estimated at this stage by regressing indicators on their components as if they assumed the reflective measurement model (5). You may see more detailed explanation on this algorithm in Tenenhaus et al. (2005). It is known that limited information methods may render estimators more robust to model misspecification in general (Bollen 1996; Fomby et al. 2011; Gerbing and Hamilton 1994).

There is no distributional assumption on the data for both GSCA and PLSPM, so they estimate standard error or confidence interval of the estimates via the bootstrap method (Efron 1979).

3 Data generating process for variance-based structural equation models

We first explain how data have been generated from a structural equation model with several layers of endogenous components in a recursive structural model but without cross-loadings and cross-weights in the measurement and weight relation models. This extant DGP is an expansion of the ones proposed by Becker et al. (2013) and Cho et al. (2019) and corresponds to the one specified in Dijkstra (2017), which has been used in Sarstedt et al. (2016)'s and Hair et al. (2017)'s simulation study. In the DGP, weight parameter values are arbitrarily manipulated by an experimenter, thereby not reflecting information on covariances of indicators. We discuss the intrinsic limitations of this DGP as a standard for evaluating all the variance-based SEM techniques and propose a new DGP tailored to variance-based SEM.

The general variance-based structural equation model can be defined as a class of (1) the weighted relation model, (2) the structural model and (3) the measurement model. Without cross-loadings and cross-weights specified, the variance-based structural equation model can be seen as a set of (7) the weight relation model, (4) the structural model and (5) the reflective measurement models as well, from which we generated data for our simulation study. To facilitate the explanations on the DGP for this model, we initially re-express the general variance-based structural equation model while splitting components into the exogenous and endogenous components, and specify additional restrictions imposed on the model where cross-loadings and cross-weights are not specified. The notations used in Sect. 2 still retain. In addition, let $\mathbf{z} = \begin{bmatrix} \mathbf{z}_{\mathrm{X}} \\ \mathbf{z}_{\mathrm{Y}} \end{bmatrix}$, where $\mathbf{z}_{\mathrm{X}}$ is a J_{X} by 1 vector of indicators for exogenous components, J_{X} is the number of the indicators for exogenous components, $\mathbf{z}_{\mathrm{Y}}$ is a J_{Y} by 1 vector of indicators for endogenous components, and J_{Y} is the number of the indicators for endogenous components. Let $\boldsymbol{\Sigma}\mathrm{z} = \begin{bmatrix} \boldsymbol{\Sigma}\mathrm{z}_{\mathrm{X}} & \boldsymbol{\Sigma}\mathrm{z}_{\mathrm{X}}\mathrm{z}_{\mathrm{Y}} \\ \boldsymbol{\Sigma}\mathrm{z}_{\mathrm{X}}\mathrm{z}'_{\mathrm{Y}} & \boldsymbol{\Sigma}\mathrm{z}_{\mathrm{Y}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}\mathrm{z}_1 & \cdots & \boldsymbol{\Sigma}\mathrm{z}_1\mathrm{z}_p & \cdots & \boldsymbol{\Sigma}\mathrm{z}_1\mathrm{z}_P \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{\Sigma}\mathrm{z}_1\mathrm{z}'_p & \ddots & \boldsymbol{\Sigma}\mathrm{z}_p & \ddots & \boldsymbol{\Sigma}\mathrm{z}_p\mathrm{z}_P \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{\Sigma}\mathrm{z}_1\mathrm{z}'_P & \cdots & \boldsymbol{\Sigma}\mathrm{z}_p\mathrm{z}'_P & \cdots & \boldsymbol{\Sigma}\mathrm{z}_P \end{bmatrix} \in \mathbb{R}^{J \times J}$ denote a J by J covariance matrix of indicators, where $\boldsymbol{\Sigma}\mathrm{z}_{\mathrm{X}}$ is a J_{X} by J_{X} covariance matrix of indicators for exogenous components, $\boldsymbol{\Sigma}\mathrm{z}_{\mathrm{Y}}$ is a J_{Y} by J_{Y} covariance matrix of indicators for endogenous components, $\boldsymbol{\Sigma}\mathrm{z}_{\mathrm{X}}\mathrm{z}_{\mathrm{Y}}$ is a J_{X} by J_{Y} cross-covariance matrix of indicators for exogenous and endogenous components, $\boldsymbol{\Sigma}\mathrm{z}_p$ is a J_p by J_p covariance matrix of indicators for the pth component, J_p is the number of the indicators for the pth component and $\boldsymbol{\Sigma}\mathrm{z}_p\mathrm{z}_{p*}$ is a cross-covariance matrix of indicators for the pth and $p*$th components. $\boldsymbol{\Sigma}\mathrm{z}$ is assumed to be positive definite, implying that all the indicators are linearly independent. Let $\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\gamma}_{\mathrm{X}} \\ \boldsymbol{\gamma}_{\mathrm{Y}} \end{bmatrix}$, where $\boldsymbol{\gamma}_{\mathrm{X}}$ is a P_{X} by 1 vector of exogenous components and $\boldsymbol{\gamma}_{\mathrm{Y}}$ is a P_{Y} by 1 vector of endogenous components. Let $\boldsymbol{\Sigma}\gamma_{\mathrm{X}}$ denote a P_{X} by P_{X} covariance matrix of exogenous components and $\boldsymbol{\Sigma}\gamma_{\mathrm{Y}}$ denote a P_{Y} by P_{Y} covariance matrix of endogenous components. Let $\mathbf{e} = \begin{bmatrix} \mathbf{e}_{\mathrm{X}} \\ \mathbf{e}_{\mathrm{Y}} \end{bmatrix} = [\mathbf{e}'_1, \ldots, \mathbf{e}'_p, \ldots \mathbf{e}'_P]'$, where $\mathbf{e}_{\mathrm{X}}$ is a J_{X} by 1 vector of residuals for the indicators forming exogenous components, $\mathbf{e}_{\mathrm{Y}}$ is a J_{Y} by 1 vector of residuals for the indicators forming endogenous components, and $\mathbf{e}_p$ denote a J_p by 1 vector of residuals for the indicators forming the pth components. Let $\boldsymbol{\Sigma}\mathrm{e} = \begin{bmatrix} \boldsymbol{\Sigma}\mathrm{e}_{\mathrm{X}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}\mathrm{e}_{\mathrm{Y}} \end{bmatrix} \in \mathbb{R}^{J \times J}$ denote a J by J covariance matrix of residuals, where $\boldsymbol{\Sigma}\mathrm{e}_{\mathrm{X}}$ is a J_{X} by J_{X} covariance matrix of residuals for the indicators related to exogenous components and $\boldsymbol{\Sigma}\mathrm{e}_{\mathrm{Y}}$ is a J_{Y} by J_{Y} covariance matrix of residuals for the indicators related to exogenous components. Let $\boldsymbol{\varepsilon} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\varepsilon}^* \end{bmatrix}$, where $\boldsymbol{\varepsilon}^*$ is a P_{Y} by 1 vector of errors for the endogenous components. Let $\boldsymbol{\Sigma}\varepsilon^* = \mathrm{diag}\left([\delta_1, \ldots, \delta_k \ldots, \delta_{P_Y}]\right)$ denote the P_{Y} by P_{Y} covariance matrix of errors for the endogenous components, where δ_k is the variance of the error for the kth endogenous component and diag() is an operator to convert a vector or matrix argument into a block-diagonal matrix. Let $\mathbf{W} = \begin{bmatrix} \mathbf{W}_{\mathrm{X}} & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_{\mathrm{Y}} \end{bmatrix}$, where $\mathbf{W}_{\mathrm{X}}$ is a J_{X} *by* P_{X} matrix of weights of the indicators for exogenous components, and $\mathbf{W}_{\mathrm{Y}}$ is a J_{Y} by P_{Y} matrix of weights of the indicators for endogenous components. Let $\mathbf{B} = \begin{bmatrix} \mathbf{0} & \mathbf{B}_{\mathrm{X}} \\ \mathbf{0} & \mathbf{B}_{\mathrm{Y}} \end{bmatrix}$, where $\mathbf{B}_{\mathrm{X}}$ is a P_{X} by P_{X} matrix of path coefficients relating exogenous components to endogenous components and $\mathbf{B}_{\mathrm{Y}}$ is a P_{Y} by P_{Y} upper triangular matrix of path coefficients relating endogenous components among themselves. Let $\mathbf{C} = \begin{bmatrix} \mathbf{C}_{\mathrm{X}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}_{\mathrm{Y}} \end{bmatrix}$, where $\mathbf{C}_{\mathrm{X}}$ is a P_{X} by J_{X} matrix of loading relating exogenous components to their indicators, and $\mathbf{C}_{\mathrm{Y}}$ is a P_{Y} by J_{Y} matrix of loadings relating endogenous components to their indicators. Then, the general variance-based structural equation model is re-written as

$$\boldsymbol{\gamma}_{\mathrm{X}} = \mathbf{W}'_{\mathrm{X}}\mathbf{z}_{\mathrm{X}} \tag{10}$$

$$\boldsymbol{\gamma}_{\mathrm{Y}} = \mathbf{W}'_{\mathrm{Y}}\mathbf{z}_{\mathrm{Y}}, \tag{11}$$

$$\boldsymbol{\gamma}_{\mathrm{Y}} = \mathbf{B}'_{\mathrm{X}}\boldsymbol{\gamma}_{\mathrm{X}} + \mathbf{B}'_{\mathrm{Y}}\boldsymbol{\gamma}_{\mathrm{Y}} + \boldsymbol{\varepsilon}^*. \tag{12}$$

$$\mathbf{z}_{\mathrm{X}} = \mathbf{C}'_{\mathrm{X}}\boldsymbol{\gamma}_{\mathrm{X}} + \mathbf{e}_{\mathrm{X}} \tag{13}$$

$$\mathbf{z}_{\mathrm{Y}} = \mathbf{C}'_{\mathrm{Y}}\boldsymbol{\gamma}_{\mathrm{Y}} + \mathbf{e}_{\mathrm{Y}} \tag{14}$$

where $\mathrm{Cov}(\boldsymbol{\gamma}_p, \mathbf{e}_p) = 0$, and $\mathrm{Cov}(\boldsymbol{\gamma}_{\mathrm{X}}, \boldsymbol{\varepsilon}^*) = 0$. When the model does not include cross-loadings and cross-weights, we can impose additional zero constraints on the off-diagonal submatrix of $\mathbf{W}$, $\mathbf{C}$ and $\boldsymbol{\Sigma}\mathrm{e}$, so that $\mathbf{C} = \mathrm{diag}\left([\mathbf{c}_1, \ldots, \mathbf{c}_p \ldots, \mathbf{c}_P]\right)$, $\mathbf{W} = \mathrm{diag}([\mathbf{w}_1, \ldots, \mathbf{w}_p, \ldots, \mathbf{w}_P])$, and $\boldsymbol{\Sigma}\mathrm{e} = \mathrm{diag}\left([\boldsymbol{\Sigma}\mathrm{e}_1, \ldots, \boldsymbol{\Sigma}\mathrm{e}_p, \ldots, \boldsymbol{\Sigma}\mathrm{e}_P]\right)$, where $\mathbf{c}_p$ is a vector of loadings relating the pth component to its indicators, $\mathbf{w}_p$ is a vector of weights of the indicators for the pth component, and $\boldsymbol{\Sigma}\mathrm{e}_p$ is a J_p by J_p covariance matrix of residuals for the indicators forming the pth component. Also, it can be additionally postulated that $\mathrm{Cov}(\boldsymbol{\gamma}_p, \mathbf{e}_{p^*}) = 0\, \forall p, p^* \in \{1, 2, \ldots, P\}$, $p \neq p^*$, followed by $\mathrm{Cov}(\boldsymbol{\gamma}, \mathbf{e}) = \mathbf{0}$.

A covariance matrix of $\boldsymbol{\Sigma}\mathrm{z}$, a matrix needed for data generation, would be obtained by the following steps:

Step 1: For each p, prescribe the values of $\boldsymbol{\Sigma}\mathrm{z}_p$, $\mathbf{B}_{\mathrm{X}}$, $\mathbf{B}_{\mathrm{Y}}$, $\boldsymbol{\Sigma}\gamma_{\mathrm{X}}$, and unstandardized weight vectors for the pth components, denoted as $\tilde{\mathbf{w}}_p$. With the pre-determined values of $\tilde{\mathbf{w}}_p$, $\mathbf{w}_p$ would be re-calculated as $\frac{\tilde{\mathbf{w}}_p}{\sqrt{\tilde{\mathbf{w}}_p' \boldsymbol{\Sigma}\mathrm{z}_p \tilde{\mathbf{w}}_p}}$, because the variance of γ_p is expressed as

$$\mathrm{Var}(\gamma_p) = \frac{1}{N}\mathrm{E}(\mathbf{w}_p'\mathbf{z}_p'\mathbf{z}_p\mathbf{w}_p) = \mathbf{w}_p'\frac{1}{N}\mathrm{E}(\mathbf{z}_p'\mathbf{z}_p)\mathbf{w}_p = \mathbf{w}_p'\boldsymbol{\Sigma}\mathrm{z}_p\mathbf{w}_p, \tag{15}$$

where E(X) is the expectation of a random variable, X. As an alternative to prescribing $\tilde{\mathbf{w}}_p$, $\mathbf{w}_p$ satisfying $\mathbf{w}_p'\boldsymbol{\Sigma}\mathrm{z}_p\mathbf{w}_p = 1$ can be directly chosen as well.

Step 2: Given the values of either $\tilde{\mathbf{w}}_p$ and $\mathbf{w}_p$, calculate $\mathbf{c}_p$ as follows:

$$\mathbf{c}_p = \mathbf{w}_p'\boldsymbol{\Sigma}\mathrm{z}_p \tag{16}$$

This Eq. (16) is derived from $\mathrm{Cov}(\gamma_p, \mathbf{e}_p) = \mathbf{0}$, because

$$\begin{aligned}
&\mathrm{Cov}(\gamma_p, \mathbf{e}_p)\\
&\quad = \mathrm{C}ov(\gamma_p, \mathbf{z}_p - \mathbf{c}_p'\mathbf{w}_p'\mathbf{z}_p) = \mathrm{C}ov(\mathbf{w}_p'\mathbf{z}_p, (\mathbf{I} - \mathbf{c}_p'\mathbf{w}_p')\mathbf{z}_p)\\
&\quad = \mathbf{w}_p'\boldsymbol{\Sigma}\mathrm{z}_p(\mathbf{I} - \mathbf{c}_p'\mathbf{w}_p')' = \mathbf{w}_p'\boldsymbol{\Sigma}\mathrm{z}_p - \mathbf{w}_p'\boldsymbol{\Sigma}\mathrm{z}_p\mathbf{w}_p\mathbf{c}_p\\
&\quad = \mathbf{w}_p'\boldsymbol{\Sigma}\mathrm{z}_p - \mathbf{c}_p.
\end{aligned}$$

Step 3: Calculate $\boldsymbol{\Sigma}\mathrm{e}_p$. As $\mathrm{Var}(\mathbf{e}_p)$= $\mathrm{Var}(\mathbf{z}_p - \mathbf{c}_p'\mathbf{w}_p'\mathbf{z}_p)$, this becomes equivalent to $\boldsymbol{\Sigma}\mathrm{e}_p = (\mathbf{I} - \mathbf{c}_p'\mathbf{w}_p')\boldsymbol{\Sigma}\mathrm{z}_p(\mathbf{I} - \mathbf{w}_p\mathbf{c}_p)$,

Step 4: Construct matrices of $\mathbf{C}_{\mathrm{X}}$, $\mathbf{C}_{\mathrm{Y}}$, $\mathbf{W}_{\mathrm{X}}$, $\mathbf{W}_{\mathrm{Y}}$, $\boldsymbol{\Sigma}\mathrm{e}_{\mathrm{X}}$ and $\boldsymbol{\Sigma}\mathrm{e}_{\mathrm{Y}}$ from the determined values of $\mathbf{c}_p$, $\mathbf{w}_p$ and $\boldsymbol{\Sigma}\mathrm{e}_p$ in earlier steps.

Step 5: Determine $\boldsymbol{\Sigma}\gamma_{\mathrm{Y}}$. From (12), γ_{Y} becomes $\boldsymbol{\gamma}_{\mathrm{Y}} = (\mathbf{I} - \mathbf{B}_{\mathrm{Y}}')^{-1}\mathbf{B}_{\mathrm{X}}'\boldsymbol{\gamma}_{\mathrm{X}} + (\mathbf{I} - \mathbf{B}_{\mathrm{Y}}')^{-1}\mathrm{e}^*$. Thus, the covariance matrix of γ_{Y} can be expressed as

$$\boldsymbol{\Sigma}\gamma_{\mathrm{Y}} = (\mathbf{I} - \mathbf{B}_{\mathrm{Y}}')^{-1}(\mathbf{B}_{\mathrm{X}}'\boldsymbol{\Sigma}\gamma_{\mathrm{X}}\mathbf{B}_{\mathrm{X}} + \boldsymbol{\Sigma}\varepsilon^*)(\mathbf{I} - \mathbf{B}_{\mathrm{Y}})^{-1}. \tag{17}$$

From (17), the diagonal entries of $\boldsymbol{\Sigma}\varepsilon^*$ are numerically determined such that the diagonal entries of $\boldsymbol{\Sigma}\gamma_{\mathrm{Y}}$ are equal to one, since $\boldsymbol{\Sigma}\varepsilon^*$ cannot be expressed as a function of other matrices like (17) for $\boldsymbol{\Sigma}\gamma_{\mathrm{Y}}$. A nonlinear optimization function or package developed for various programming software can be utilized to find a numerical solution for $\boldsymbol{\Sigma}\varepsilon^*$, for instance, *fminsearch* function in MATLAB or *optimr* package in R.

Step 6: Determine $\boldsymbol{\Sigma}\mathrm{z}_{\mathrm{X}}$, $\boldsymbol{\Sigma}\mathrm{z}_{\mathrm{Y}}$ and $\boldsymbol{\Sigma}\mathrm{z}_{\mathrm{X}}\mathrm{z}_{\mathrm{Y}}$. Inserting the prescribed or determined values in earlier steps, $\boldsymbol{\Sigma}\mathrm{z}_{\mathrm{X}} = \mathbf{C}_{\mathrm{X}}'\boldsymbol{\Sigma}\gamma_{\mathrm{X}}\mathbf{C}_{\mathrm{X}} + \boldsymbol{\Sigma}\mathrm{e}_{\mathrm{X}}$ and $\boldsymbol{\Sigma}\mathrm{z}_{\mathrm{Y}} = \mathbf{C}_{\mathrm{Y}}'\boldsymbol{\Sigma}\gamma_{\mathrm{Y}}\mathbf{C}_{\mathrm{Y}} + \boldsymbol{\Sigma}\mathrm{e}_{\mathrm{Y}}$. Afterwards, $\boldsymbol{\Sigma}\mathrm{z}_{\mathrm{X}}\mathrm{z}_{\mathrm{Y}}$ is obtained by $\boldsymbol{\Sigma}\mathrm{z}_{\mathrm{X}}\mathrm{z}_{\mathrm{Y}} = \mathbf{C}_{\mathrm{X}}'\boldsymbol{\Sigma}\gamma_{\mathrm{X}}\mathbf{B}_{\mathrm{X}}(\mathbf{I} - \mathbf{B}_{\mathrm{Y}})^{-1}\mathbf{C}_{\mathrm{Y}}$. It follows from

$$\begin{aligned}
&\mathrm{Cov}(\mathbf{Z}_{\mathrm{X}}, \mathbf{Z}_{\mathrm{Y}})\\
&\quad = \mathrm{Cov}(\mathbf{C}_{\mathrm{X}}'\boldsymbol{\gamma}_{\mathrm{X}} + \mathbf{e}_{\mathrm{X}}, \mathbf{C}_{\mathrm{Y}}'\boldsymbol{\gamma}_{\mathrm{Y}} + \mathbf{e}_{\mathrm{Y}})\\
&\quad = \mathbf{C}_{\mathrm{X}}'\mathrm{Cov}(\boldsymbol{\gamma}_{\mathrm{X}}, \boldsymbol{\gamma}_{\mathrm{Y}})\mathbf{C}_{\mathrm{Y}}\\
&\quad = \mathbf{C}_{\mathrm{X}}'\mathrm{Cov}(\boldsymbol{\gamma}_{\mathrm{X}}, (\mathbf{I} - \mathbf{B}_{\mathrm{Y}}')^{-1}\mathbf{B}_{\mathrm{X}}'\boldsymbol{\gamma}_{\mathrm{X}} + (\mathbf{I} - \mathbf{B}_{\mathrm{Y}}')^{-1}\boldsymbol{\varepsilon}^*)\mathbf{C}_{\mathrm{Y}}\\
&\quad = \mathbf{C}_{\mathrm{X}}'\boldsymbol{\Sigma}\gamma_{\mathrm{X}}\mathbf{B}_{\mathrm{X}}(\mathbf{I} - \mathbf{B}_{\mathrm{Y}})^{-1}\mathbf{C}_{\mathrm{Y}}.
\end{aligned}$$

Now, we have $\mathbf{\Sigma}\mathrm{z} = \begin{bmatrix} \mathbf{\Sigma}\mathrm{z}_\mathrm{X} & \mathbf{\Sigma}\mathrm{z}_\mathrm{X}\mathrm{z}_\mathrm{Y} \\ \mathbf{\Sigma}\mathrm{z}_\mathrm{X}\mathrm{z}_\mathrm{Y}' & \mathbf{\Sigma}\mathrm{z}_\mathrm{Y} \end{bmatrix}$ and can generate data from a multivariate distribution with the zero mean vector and the $\mathbf{\Sigma}\mathrm{z}$. By generating data based on the above steps of DGP with the structural equation model specified in Hair et al. (2017), the equivalent results on the relative performance of GSCA *with formative indicators* and PLSPM *with mode B* were obtained (see the Table 1 in the Supplementary material).

This extant DGP is of significance as an initial proposal for variance-based structural equation models. Dijkstra (2017) analytically showed that the components in this DGP are canonical variables, implying that the components are constructed to maximize the explained variances of the endogenous components, and it was empirically verified in Hair et al. (2017)'s simulation study: the estimators of GSCA *with formative indictors* and PLS *with mode B* were consistent. In the abovementioned DGP, however, values of the weight parameters were arbitrarily chosen regardless of covariances of indicators in Step 1. Except for scaling constraint of weights, no functional relations between weights and covariances of indicators are considered in the DGP. Simply put, the extant DGP has no mechanism of accounting for the variances of indicators when forming the components. This result may be against many researchers' expectation that the components would also reflect the information about indicators and adequately explain their variances. Consequently, the extant DGP is not concordant with GSCA *with reflective indicators* and PLSPM *with mode A*, which aim to form components that explain the variances of indicators as well as those of endogenous components. When applied to the data generated from the extant DGP, GSCA *with reflective indicators* and PLSPM *with mode A* are expected to produce biased estimates (see the Table 2 in Supplementary material).

Addressing this concern, we propose a new DGP specifying the functional relation between weights of indicators for components and their covariance matrix. In this DGP, weights of indicators for a component are initially determined to well explain the variances of the indicators given the covariances of the indicators. Set the values of $\mathbf{\Sigma}\mathrm{z}_p$ for $p = 1, 2, \ldots, P$. Then, $\mathbf{w}_p$ is obtained by,

$$\mathbf{w}_p = (\mathbf{\Sigma}\mathrm{z}_p)^{-\frac{1}{2}}\mathbf{u}_{1,p} \tag{18}$$

where $(\mathbf{\Sigma}\mathrm{z}_p)^{-\frac{1}{2}} = \mathbf{U}_p(\mathbf{D}_p)^{-\frac{1}{2}}\mathbf{U}_p'$, $\mathbf{D}_p = \mathrm{diag}\left([\mathrm{d}_1, \mathrm{d}_2, \ldots, \mathrm{d}_{Jp}]\right)$ is a J_p by J_p diagonal matrix of eigenvalues of $\mathbf{\Sigma}\mathrm{z}_p$ arranged in a descending order, $\mathbf{U}_p = [\mathbf{u}_{1,p}, \mathbf{u}_{2,p}, \ldots, \mathbf{u}_{Jp,p}]$ is a J_p by J_p matrix of eigenvectors corresponding to the eigenvalues, and $\mathbf{u}_{1,p}$ is the eigenvector corresponding to the largest eigenvalue, d_1. The rest of the procedures are the same as in the previous DGP.

We delineate the procedure to derive (18). The first step is to find the deterministic relation between the weights of indicators and the amount of explained variances of the indicators by their weighted composites. Let $\mathbf{R}_p^2$ denote a J_p by 1 vector of the explained variances of indicators forming the pth components relative to the entire variances of indicators. The average $\mathbf{R}_p^2$ is the mean of the elements of $\mathbf{R}_p^2$. Since the model does not involve cross-loadings and cross-weights and every indicator and component is standardized, those relations can be expressed as follows:

$$\begin{aligned}&\text{average } \mathbf{R}_p^2\\ &\quad = J_p^{-1}\text{trace}(\text{Cov}(\mathbf{c}_p'\gamma_p))\\ &\quad = J_p^{-1}\text{trace}(\mathbf{c}_p'\boldsymbol{\Sigma}\gamma_p\mathbf{c}_p)\\ &\quad = J_p^{-1}\text{trace}((\mathbf{w}_p'\boldsymbol{\Sigma}\text{z}_p)'\mathbf{w}_p'\boldsymbol{\Sigma}\text{z}_p)\\ &\quad = J_p^{-1}\text{trace}(\mathbf{w}_p'(\boldsymbol{\Sigma}\text{z}_p)^2\mathbf{w}_p)\\ &\quad = J_p^{-1}\mathbf{w}_p'(\boldsymbol{\Sigma}\text{z}_p)^2\mathbf{w}_p\end{aligned} \tag{19}$$

Then, finding the weights with which the composite of indicators maximizes its capability to explain the variances of the indicators amounts to solving the following optimization problem,

$$\underset{\mathbf{w}_p}{\text{Max}}\ \mathbf{w}_p'(\boldsymbol{\Sigma}\text{z}_p)^2\mathbf{w}_p \text{ subject to } \mathbf{w}_p'\boldsymbol{\Sigma}\text{z}_p\mathbf{w}_p = 1. \tag{20}$$

This is the constrained quadratic optimization problem on the ellipsoid (Gallier and Quaintance 2019, Chapter 37.3). Since $\boldsymbol{\Sigma}\text{z}_p$ is positive definite, it can be orthogonally diagonalized as

$$\boldsymbol{\Sigma}\text{z}_p = \mathbf{U}_p\mathbf{D}_p\mathbf{U}_p'. \tag{21}$$

Let $(\boldsymbol{\Sigma}\text{z}_p)^{\frac{1}{2}}$ and $(\boldsymbol{\Sigma}\text{z}_p)^{-\frac{1}{2}}$ be defined by

$$\begin{aligned}(\boldsymbol{\Sigma}\text{z}_p)^{\frac{1}{2}} &= \mathbf{U}_p(\mathbf{D}_p)^{\frac{1}{2}}\mathbf{U}_p'\\ (\boldsymbol{\Sigma}\text{z}_p)^{-\frac{1}{2}} &= \mathbf{U}_p(\mathbf{D}_p)^{-\frac{1}{2}}\mathbf{U}_p'.\end{aligned} \tag{22}$$

They are the inverse matrix of each other so that

$$(\boldsymbol{\Sigma}\text{z}_p)^{\frac{1}{2}}(\boldsymbol{\Sigma}\text{z}_p)^{-\frac{1}{2}} = \mathbf{U}_p(\mathbf{D}_p)^{\frac{1}{2}}\mathbf{U}_p'\mathbf{U}_p(\mathbf{D}_p)^{-\frac{1}{2}}\mathbf{U}_p' = \mathbf{I} \tag{23}$$

By (21) and (22), the reparameterization of $\mathbf{w}_p$ as $\mathbf{w}_p = (\boldsymbol{\Sigma}\text{z}_p)^{-\frac{1}{2}}\widehat{\mathbf{w}}_p$ transforms (20) into well-known constrained quadratic optimization problem on the unit sphere,

$$\underset{\widehat{\mathbf{w}}_p}{\text{Max}}\ \widehat{\mathbf{w}}_p'\boldsymbol{\Sigma}\text{z}_p\widehat{\mathbf{w}}_p \text{ subject to } \widehat{\mathbf{w}}_p'\widehat{\mathbf{w}}_p = 1. \tag{24}$$

The solution for (24) is $\widehat{\mathbf{w}}_p = \mathbf{u}_{1,p}$, that is, $\mathbf{w}_p = (\boldsymbol{\Sigma}\text{z}_p)^{-\frac{1}{2}}\mathbf{u}_{1,p}$, at which the largest value of the objective function, $\mathbf{w}_p'(\boldsymbol{\Sigma}\text{z}_p)^2\mathbf{w}_p$, is attained as d_1. You may see the detailed explanation for (24) from Chapter 7 in Lay et al. (2015) or Chapter 18.4 in Gallier and Quaintance (2019).

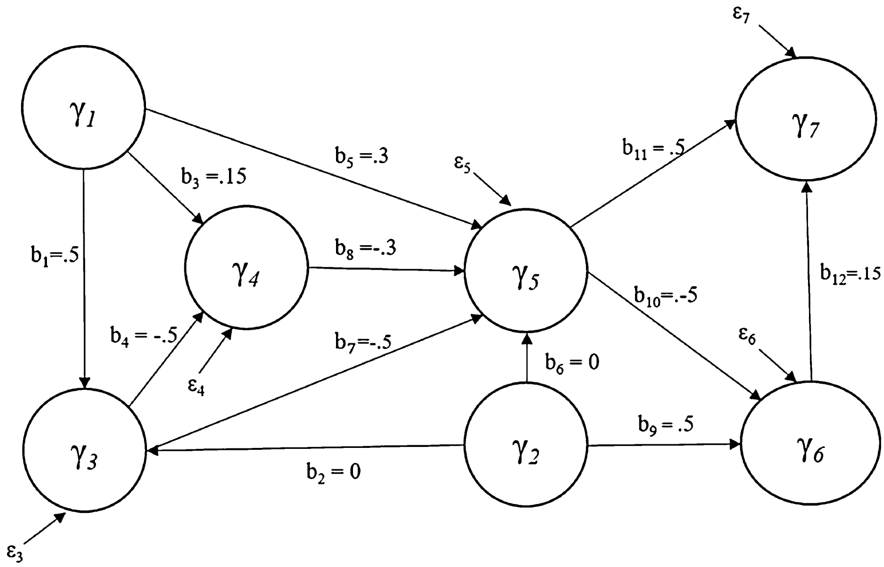

Fig. 1 A variance-based structural equation model with two exogenous components and five endogenous components specified for the simulation study. Note that measurement models are omitted for a simpler depiction

The components constructed from this new DGP can be interpreted in three different ways. Firstly, they can be seen as principal components because the vector of weights for each component is determined in the same manner as for the first principal component in the principal component analysis. Secondly, those components can be also classified as canonical components, as they still satisfy all the relations among the parameters specified in the previous DGP. The new DGP just additionally specifies the relations between the weights of indicators and their covariances. Lastly, the components in the new DGP can be regarded as the one constructed to explain the variances of all the dependent variables specified in both measurement and structural models concurrently as much as possible. We name this type of components nomological components in that they can correspond to the concepts defined by the entire nomological network including both observed and latent variables (Cronbach and Meehl 1955). Accordingly, all the variance-based SEM techniques, whether to consider maximizing explained variances of either indicators or endogenous components only or both in constructing components (i.e. whether to construct principal, canonical or nomological components), can be adopted to analyze data from this new DGP, and thus, we can call this DGP a DGP for variance-based structural equation models. This is the condition where PLSPM *with mode A* and PLSPM *with mode B* are perfectly matched with each other and may work well asymptotically according to Dijkstra (2017). We employed this DGP for empirically evaluating the relative performance of both PLSPM *with mode A* and PLSPM *with mode B* and their counterparts, GSCA *with reflective indicators* and GSCA *with formative indicators* in our simulation study.

4 Simulation

We undertook a comprehensive examination of four SEM approaches—GSCA *with reflective*, GSCA *with formative indicators*, PLSPM *with mode A*, and PLSPM *with mode B* in parameter recovery and hypothesis testing under variance-based structural equation models. For simplicity, GSCA *with reflective indicators* and GSCA *with formative indicators* are abbreviated to $\mathrm{GSCA_R}$ and $\mathrm{GSCA_F}$, while PLSPM with *mode A* and PLSPM *with mode B* to $\mathrm{PLS_A}$ and $\mathrm{PLS_B}$, respectively. We considered three simulation design factors: sample size ($N=100$, 250, 500, 1000), the number of indicators per component ($N_{\mathrm{ind}}=2$, 4, 6, 8), and the average correlation within the indicators for a component ($r=0.2$, 0.4, 0.6). In total, our experiment was comprised of 48 simulation conditions (4 sample sizes × 4 indicator numbers × 3 average correlations).

We specified a variance-based structural equation model with two exogenous and five endogenous components, as in Hair et al. (2017) (Fig. 1). This model reflects the American Customer Satisfaction Index model (ACSI; Fornell et al. 1996) which is one of the most influential variance-based structural equation models in studying the behavior of consumer satisfaction (e.g. Anderson and Fornell 2002; Eklöf and Westlund 2002; Rego et al. 2013). To mirror the reality, we assigned various values of different signs to path coefficients: two null values (i.e. $\mathrm{b}_2=\mathrm{b}_6=0$), two small values (i.e. $\mathrm{b}_3=\mathrm{b}_{12}=0.15$), two medium values (i.e. $\mathrm{b}_5=0.3$, $\mathrm{b}_8=-0.3$), and six large values (i.e. $\mathrm{b}_1=\mathrm{b}_9=\mathrm{b}_{11}=0.5$, $\mathrm{b}_4=\mathrm{b}_7=\mathrm{b}_{10}=-0.5$). Two exogenous components were correlated to each other by 0.3 (i.e. $\boldsymbol{\Sigma}\gamma_{\mathrm{X}}=\begin{bmatrix} 1 & .3 \\ .3 & 1 \end{bmatrix}$).

Given the number of indicators per component and the value of the average correlation among the indicators for a component, individual correlations among the indicators for a component are randomly chosen to construct their covariance matrix, $\boldsymbol{\Sigma}\mathrm{z}_p$. Once $\boldsymbol{\Sigma}\mathrm{z}_p$ is determined, we applied this matrix to every block of indicators. The range of individual correlations for each condition was [0.1, 0.3] for $r=0.2$, [0.2, 0.6] for $r=0.4$, [0.4 0.8] for $r=0.6$. For each experimental condition, we have $\boldsymbol{\Sigma}\mathrm{z}_p$, $\mathbf{B}_{\mathrm{X}}$, $\mathbf{B}_{\mathrm{Y}}$, and $\boldsymbol{\Sigma}\gamma_{\mathrm{X}}$, from which we derived the covariance matrix of entire indicators with all the true parameter values via the DGP for variance-based structural equation models (see Tables 4–7 in Supplementary materials). We generated 500 random samples from the multivariate normal distribution with a zero vector of mean and $\boldsymbol{\Sigma}\mathrm{z}$ obtained under each condition, to which, in turn, the four variance-based SEM approaches were applied. Based on their estimates, we evaluated three properties of estimators of each approach—bias, consistency, and relative efficiency—, and their performance in hypothesis testing—type I error and statistical power. Note that, in $\mathrm{GSCA_F}$, loadings were additionally estimated as in $\mathrm{PLS_B}$, though $\mathrm{GSCA_F}$ does not estimate loading parameters in general. As explained in Sect. 2, the post examination on directional relations between components and their indicators are always conducted in $\mathrm{PLS_B}$. With the component scores computed on the final weight estimates, indicators are regressed on their components by OLS to obtain loading estimates. We applied the same procedure to $\mathrm{GSCA_F}$ and computed loading estimates. These loading estimates have the same meaning as those in $\mathrm{GSCA_R}$ and $\mathrm{PLS_F}$—how strongly correlated the components are with its indicators, but their absolute values may be smaller on average since variances of indicators are not considered in the estimation of weight parameters.

For bias and consistency, we calculated the relative bias (RB) of an estimator ($\hat{\theta}$):

$$\mathrm{RB}(\hat{\theta}) = \frac{\mathrm{E}(\hat{\theta}) - \theta}{\theta} \approx \frac{\frac{1}{Nrep}\sum_{i=1}^{Nrep}(\hat{\theta}_i) - \theta}{\theta}, \tag{25}$$

where θ is the parameter to be estimated by $\hat{\theta}$, *Nrep* is the number of replications in an experiment, and $\hat{\theta}_i$ is an estimate for θ given the *i*th sample. Estimators whose absolute value of the relative bias was larger than 10% were treated as unacceptably biased ones (e.g. Bollen et al. 2007; Hwang et al. 2010). If the relative bias of an estimator becomes close to zero with larger sample size and the value is below 10% with the largest sample size, the estimator was regarded as being empirically consistent. On the other hand, to assess the relative efficiency of the estimators, we computed root mean squared error (RMSE). Root mean squared error (RMSE),

$$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)} \approx \sqrt{\frac{1}{Nrep}\sum_{i=1}^{Nrep}(\hat{\theta}_i - \theta)^2} \tag{26}$$

is a metric to quantify errors of an estimator. An estimator with lower RMSE can be said to be more efficient than the others with higher RMSE. RMSE may serve as a better criterion in estimating expected errors of an estimator than mean absolute error (MAE) when errors are expected to follow normal distribution rather than uniform distribution (Chai and Draxler 2014).

Table 1 depicts the average RB values of the estimators for each sub-model (i.e. averaged over the entire weights in the weighted relation model), for each approach, given the simulation condition. The average RBs were calculated with the absolute RB values of estimators and did not consider the estimators for the parameter of zero value. As shown in Table 1, $\mathrm{GSCA_R}$ and $\mathrm{PLS_A}$ provided unbiased and consistent estimators across all the simulation conditions. Their average RBs were less than 10% across the sample sizes, irrespective of the level of r and N_{ind}, and became close to zero as N increased. $\mathrm{GSCA_F}$ and $\mathrm{PLS_B}$ also produced unbiased and consistent estimators in general. Their average RBs were less than 10% except for that N was small (i.e. 100) and N_{ind} was large (i.e. 8), and tended to zero value as N increased. In those exceptional cases, average RBs of $\mathrm{GSCA_F}$ and $\mathrm{PLS_B}$ estimators for weights and loadings were more than 10% (e.g. 12.39 and 10.18 for weights and loadings of $\mathrm{GSCA_F}$, and 15.60 and 13.42 for weights and loadings of $\mathrm{PLS_B}$ when $N=100$, $N_{\mathrm{ind}}=8$, and $r=0.6$), whereas only $\mathrm{GSCA_F}$ estimators had the average RBs greater than 10% for path coefficients (e.g. 28.46 for path coefficients of $\mathrm{GSCA_F}$, and 8.99 for path coefficients of $\mathrm{PLS_B}$ when $N=100$, $N_{\mathrm{ind}}=8$, and $r=0.6$). Overall, $\mathrm{GSCA_R}$ and $\mathrm{PLS_A}$ estimators yielded smaller RBs than $\mathrm{GSCA_F}$ and $\mathrm{PLS_B}$, and the difference was enlarged as r and N_{ind} got larger and N got smaller.

Table 2 presents the average RMSE values of the estimators for each sub-model, for each approach, given the simulation condition. The average RMSE was

Table 1 Relative bias expressed as a percentage (RB) of the estimators of each approach per experimental condition in the simulation study

N_{ind}		$r=0.2$				$r=0.4$				$r=0.6$			
		GSCA_R	PLS_A	GSCA_F	PLS_B	GSCA_R	PLS_A	GSCA_F	PLS_B	GSCA_R	PLS_A	GSCA_F	PLS_B
Weight													
2	$N=100$	0.45	0.82	1.35	1.60	0.28	0.38	1.33	1.59	0.29	0.32	1.52	2.00
	$N=250$	0.34	0.60	0.81	0.92	0.14	0.14	0.42	0.42	0.10	0.17	0.51	0.72
	$N=500$	0.18	0.21	0.39	0.33	0.09	0.16	0.32	0.40	0.12	0.11	0.50	0.45
	$N=1000$	0.10	0.13	0.19	0.18	0.13	0.19	0.46	0.43	0.09	0.07	0.31	0.26
4	$N=100$	1.03	1.48	3.52	4.50	0.41	0.63	3.60	4.58	0.43	0.47	4.24	5.05
	$N=250$	0.44	0.58	1.35	1.56	0.33	0.46	1.58	2.00	0.22	0.21	1.81	2.26
	$N=500$	0.29	0.45	0.80	0.95	0.20	0.22	0.95	0.98	0.15	0.17	1.13	1.33
	$N=1000$	0.18	0.21	0.36	0.45	0.15	0.18	0.60	0.68	0.10	0.10	0.91	1.00
6	$N=100$	1.66	2.03	6.11	8.32	0.53	0.69	6.66	8.59	0.45	0.35	8.33	9.62
	$N=250$	0.62	0.86	2.29	2.81	0.32	0.39	2.65	3.07	0.27	0.24	3.01	4.23
	$N=500$	0.49	0.59	1.30	1.50	0.17	0.22	1.33	1.54	0.20	0.15	2.17	2.61
	$N=1000$	0.28	0.35	0.79	0.81	0.13	0.19	1.03	1.09	0.15	0.13	1.52	1.41
8	$N=100$	1.72	2.16	10.05	12.89	0.62	0.70	10.64	13.27	0.77	0.38	12.39	15.60
	$N=250$	0.80	0.97	3.27	4.04	0.36	0.43	3.55	4.32	0.37	0.22	6.16	6.26
	$N=500$	0.44	0.55	1.68	1.96	0.28	0.31	2.56	2.71	0.25	0.22	3.60	3.96
	$N=1000$	0.25	0.35	0.87	1.05	0.15	0.18	1.16	1.34	0.22	0.16	3.11	3.26
Loading													
2	$N=100$	0.46	0.80	1.34	1.58	0.24	0.38	1.34	1.59	0.17	0.19	1.33	1.60
	$N=250$	0.25	0.42	0.58	0.66	0.15	0.19	0.52	0.60	0.06	0.07	0.46	0.54
	$N=500$	0.12	0.18	0.30	0.32	0.06	0.10	0.26	0.29	0.05	0.05	0.24	0.28
	$N=1000$	0.07	0.10	0.15	0.16	0.06	0.09	0.20	0.21	0.05	0.05	0.16	0.17

Table 1 (continued)

N_{ind}		$r = 0.2$				$r = 0.4$				$r = 0.6$			
		GSCA$_R$	PLS$_A$	GSCA$_F$	PLS$_B$	GSCA$_R$	PLS$_A$	GSCA$_F$	PLS$_B$	GSCA$_R$	PLS$_A$	GSCA$_F$	PLS$_B$
4	$N = 100$	0.97	1.52	4.09	5.12	0.41	0.60	4.02	4.97	0.23	0.26	3.85	4.77
	$N = 250$	0.35	0.53	1.44	1.69	0.22	0.24	1.38	1.65	0.09	0.10	1.38	1.64
	$N = 500$	0.25	0.35	0.75	0.87	0.17	0.20	0.76	0.88	0.06	0.07	0.63	0.75
	$N = 1000$	0.16	0.15	0.29	0.34	0.09	0.10	0.33	0.37	0.04	0.04	0.32	0.38
6	$N = 100$	1.31	1.86	6.81	8.81	0.41	0.50	6.82	8.75	0.17	0.22	6.66	8.98
	$N = 250$	0.40	0.63	2.33	2.81	0.22	0.26	2.32	2.80	0.12	0.14	2.28	2.75
	$N = 500$	0.37	0.43	1.12	1.33	0.13	0.15	1.08	1.29	0.05	0.05	1.04	1.26
	$N = 1000$	0.22	0.26	0.58	0.66	0.08	0.09	0.52	0.60	0.07	0.07	0.56	0.66
8	$N = 100$	1.01	1.62	10.54	13.37	0.47	0.60	10.50	13.38	0.25	0.29	10.18	13.42
	$N = 250$	0.48	0.61	3.10	3.85	0.25	0.29	3.18	3.95	0.11	0.12	3.18	3.93
	$N = 500$	0.30	0.35	1.47	1.77	0.15	0.17	1.51	1.82	0.09	0.10	1.53	1.84
	$N = 1000$	0.16	0.19	0.72	0.86	0.11	0.11	0.69	0.83	0.07	0.07	0.76	0.90
Path													
2	$N = 100$	2.24	1.28	3.03	1.11	2.06	1.88	3.28	1.72	2.12	1.58	3.88	2.02
	$N = 250$	1.30	0.95	1.71	0.83	0.57	0.56	1.02	0.73	0.89	0.77	1.45	0.84
	$N = 500$	0.47	0.39	0.59	0.38	0.39	0.36	0.64	0.35	0.60	0.61	0.75	0.74
	$N = 1000$	0.23	0.25	0.29	0.26	0.45	0.47	0.51	0.50	0.43	0.45	0.42	0.46
4	$N = 100$	3.69	2.93	9.21	3.62	3.59	2.88	10.49	2.78	2.04	1.25	9.81	3.29
	$N = 250$	1.94	1.71	4.26	1.56	1.22	1.08	3.88	1.57	1.28	0.99	4.47	2.04
	$N = 500$	0.79	0.68	1.91	0.83	0.75	0.81	1.75	1.24	0.82	0.71	2.06	1.11
	$N = 1000$	0.57	0.65	0.95	0.76	0.48	0.47	1.04	0.52	0.38	0.36	1.03	0.63

Table 1 (continued)

N_{ind}		$r=0.2$				$r=0.4$				$r=0.6$			
		GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}	GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}	GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}
6	$N=100$	6.21	4.93	17.24	5.81	4.61	3.78	18.23	4.99	3.73	2.17	19.06	5.54
	$N=250$	1.81	1.91	6.28	2.83	1.65	1.40	6.91	2.93	0.99	0.52	6.71	2.56
	$N=500$	0.86	1.08	2.96	1.56	0.99	0.88	3.43	1.57	1.00	0.84	3.51	1.67
	$N=1000$	0.79	0.81	1.83	0.87	0.71	0.75	1.80	1.21	0.28	0.35	1.37	0.72
8	$N=100$	7.25	5.89	26.36	9.75	4.44	3.32	26.73	9.20	4.74	3.15	28.46	8.99
	$N=250$	2.70	3.07	9.46	4.33	2.21	1.93	10.06	3.29	1.18	0.66	9.45	3.23
	$N=500$	1.63	1.64	4.91	2.33	0.72	0.67	4.51	1.71	0.65	0.49	4.55	1.84
	$N=1000$	1.07	1.10	2.68	1.21	0.64	0.60	2.18	1.09	0.44	0.32	2.39	1.02

Table 2 Root mean square error (RMSE) of the estimators of each approach per experimental condition in the simulation study

N_{ind}		$r=0.2$				$r=0.4$				$r=0.6$			
		GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}	GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}	GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}
Weight													
2	$N=100$	0.07	0.10	0.13	0.14	0.05	0.07	0.14	0.16	0.04	0.05	0.17	0.19
	$N=250$	0.04	0.06	0.08	0.08	0.03	0.04	0.08	0.09	0.02	0.03	0.10	0.11
	$N=500$	0.03	0.04	0.05	0.06	0.02	0.03	0.06	0.07	0.02	0.02	0.07	0.08
	$N=1000$	0.02	0.03	0.04	0.04	0.01	0.02	0.04	0.05	0.01	0.01	0.05	0.06
4	$N=100$	0.07	0.09	0.16	0.18	0.04	0.05	0.18	0.21	0.03	0.03	0.24	0.27
	$N=250$	0.04	0.05	0.09	0.10	0.02	0.03	0.11	0.12	0.02	0.02	0.15	0.16
	$N=500$	0.03	0.04	0.06	0.07	0.02	0.02	0.08	0.08	0.01	0.01	0.10	0.11
	$N=1000$	0.02	0.03	0.05	0.05	0.01	0.02	0.05	0.06	0.01	0.01	0.07	0.08
6	$N=100$	0.06	0.08	0.17	0.19	0.03	0.04	0.20	0.23	0.03	0.02	0.34	0.38
	$N=250$	0.03	0.05	0.10	0.11	0.02	0.03	0.12	0.13	0.02	0.01	0.19	0.21
	$N=500$	0.02	0.03	0.07	0.07	0.01	0.02	0.08	0.09	0.01	0.01	0.13	0.15
	$N=1000$	0.02	0.02	0.05	0.05	0.01	0.01	0.06	0.06	0.01	0.01	0.09	0.10
8	$N=100$	0.05	0.06	0.18	0.21	0.03	0.03	0.24	0.27	0.03	0.02	0.42	0.48
	$N=250$	0.03	0.04	0.10	0.11	0.02	0.02	0.13	0.15	0.02	0.01	0.24	0.27
	$N=500$	0.02	0.03	0.07	0.08	0.01	0.01	0.09	0.10	0.01	0.01	0.16	0.18
	$N=1000$	0.01	0.02	0.05	0.05	0.01	0.01	0.06	0.07	0.01	0.01	0.11	0.13
Loading													
2	$N=100$	0.07	0.09	0.11	0.12	0.04	0.05	0.09	0.10	0.02	0.03	0.08	0.09
	$N=250$	0.03	0.05	0.06	0.07	0.02	0.03	0.05	0.06	0.01	0.02	0.04	0.05
	$N=500$	0.02	0.03	0.04	0.05	0.02	0.02	0.04	0.04	0.01	0.01	0.03	0.03
	$N=1000$	0.02	0.02	0.03	0.03	0.01	0.01	0.03	0.03	0.01	0.01	0.02	0.02
4	$N=100$	0.10	0.11	0.14	0.15	0.06	0.06	0.12	0.14	0.03	0.04	0.10	0.12
	$N=250$	0.06	0.06	0.08	0.09	0.04	0.04	0.07	0.08	0.02	0.02	0.06	0.06

Table 2 (continued)

N_{ind}		$r=0.2$				$r=0.4$				$r=0.6$			
		GSCA$_{\text{R}}$	PLS$_{\text{A}}$	GSCA$_{\text{F}}$	PLS$_{\text{B}}$	GSCA$_{\text{R}}$	PLS$_{\text{A}}$	GSCA$_{\text{F}}$	PLS$_{\text{B}}$	GSCA$_{\text{R}}$	PLS$_{\text{A}}$	GSCA$_{\text{F}}$	PLS$_{\text{B}}$
	$N=500$	0.04	0.04	0.06	0.06	0.02	0.03	0.05	0.05	0.01	0.02	0.04	0.04
	$N=1000$	0.03	0.03	0.04	0.04	0.02	0.02	0.03	0.04	0.01	0.01	0.03	0.03
6	$N=100$	0.10	0.11	0.15	0.17	0.06	0.06	0.14	0.16	0.04	0.04	0.12	0.15
	$N=250$	0.06	0.07	0.09	0.10	0.04	0.04	0.08	0.08	0.02	0.02	0.06	0.07
	$N=500$	0.04	0.05	0.06	0.07	0.03	0.03	0.05	0.06	0.02	0.02	0.04	0.05
	$N=1000$	0.03	0.03	0.04	0.05	0.02	0.02	0.04	0.04	0.01	0.01	0.03	0.03
8	$N=100$	0.10	0.10	0.17	0.19	0.06	0.07	0.16	0.19	0.04	0.04	0.15	0.19
	$N=250$	0.06	0.06	0.09	0.10	0.04	0.04	0.08	0.09	0.02	0.02	0.07	0.08
	$N=500$	0.04	0.04	0.06	0.07	0.03	0.03	0.05	0.06	0.02	0.02	0.04	0.05
	$N=1000$	0.03	0.03	0.04	0.05	0.02	0.02	0.04	0.04	0.01	0.01	0.03	0.03
Path													
2	$N=100$	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.09	0.09	0.10	0.10
	$N=250$	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
	$N=500$	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	$N=1000$	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
4	$N=100$	0.10	0.10	0.11	0.11	0.10	0.10	0.11	0.11	0.10	0.10	0.11	0.11
	$N=250$	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
	$N=500$	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	$N=1000$	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
6	$N=100$	0.10	0.10	0.13	0.13	0.10	0.10	0.14	0.13	0.10	0.10	0.14	0.13
	$N=250$	0.06	0.06	0.07	0.07	0.06	0.06	0.07	0.07	0.06	0.06	0.07	0.07
	$N=500$	0.04	0.04	0.04	0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.04	0.04
	$N=1000$	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

Table 2 (continued)

N_{ind}		$r = 0.2$				$r = 0.4$				$r = 0.6$			
		GSCA_R	PLS_A	GSCA_F	PLS_B	GSCA_R	PLS_A	GSCA_F	PLS_B	GSCA_R	PLS_A	GSCA_F	PLS_B
8	$N = 100$	0.10	0.11	0.17	0.16	0.10	0.10	0.16	0.16	0.10	0.10	0.16	0.16
	$N = 250$	0.06	0.06	0.07	0.07	0.06	0.06	0.07	0.07	0.06	0.06	0.07	0.07
	$N = 500$	0.04	0.04	0.05	0.05	0.04	0.04	0.05	0.05	0.04	0.04	0.05	0.04
	$N = 1000$	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03

computed as the mean of the absolute values of estimators in the same sub-models. As Table 2 exhibits, estimators of GSCA_R and PLS_A were more efficient than those of GSCA_F and PLS_B in general. Specifically, GSCA_F and PLS_B estimators yielded larger RMSE values than GSCA_R and PLS_A estimators for weights and loadings, and this gap did not disappear even with the large size of sample (i.e. $N = 1000$), across every condition. With respect to path coefficients, the RMSEs of GSCA_F and PLS_B estimators were still larger than those of GSCA_R and PLS_A estimators when $N = 100$ or 250, but the difference in RMSEs became smaller as N increased. Between the GSCA_R and PLS_A, the estimators of GSCA_R were at least equivalent to or more efficient than those of PLS_A in general. Aside from the four conditions (i.e. $r = 0.6$, $N_\text{ind} = 6$ or 8, and $N = 100$ or 250) of the 48 conditions, the RMSEs of GSCA_R estimators for weights were less than or equal to those of PLS_A estimators. For loading parameters, GSCA_R estimators led to equivalent or smaller RMSEs on average than PLS_A estimators across all the conditions. On the other hand, GSCA_R and PLS_A showed no substantial difference in the RMSEs of the estimators for path coefficients.

For the purpose of testing the utility of each approach as a tool for hypothesis testing, we calculated their type I error and statistical power. We constructed a 95% confidence interval for each estimate via 100 bootstrap sampling and calculated the relative frequency of the cases where the confidence interval failed to contain a zero value. That frequency can be interpreted as empirical type I error for the parameter of zero value and as statistical power for the parameter of nonzero value. Table 3 shows the average type I errors over the two null path coefficients for each approach in all the experimental conditions. Overall, every approach succeeded in controlling type I error around at 0.05 level (i.e. deviated from 0.05 by 0.02 or less), except GSCA_F. GSCA_F controlled type I error too strictly when $N = 100$ and $N_\text{ind} = 6$ or 8 so that its value was 0.01 or even 0.00. Table 4 depicts the average statistical power for the parameters in each sub-model, varying N, N_ind, and r. GSCA_R and PLS_A tended to have power equal to or higher than GSCA_F and PLS_B across all the simulation conditions. In particular, inequality between the two groups (i.e. GSCA_R and PLS_A versus GSCA_F and PLS_B) was observed to a greater degree when weight and path coefficients were estimated with a small size of sample relative to the large number of indicators (e.g. $N = 100$ and $N_\text{ind} \geq 6$). In comparison of GSCA_R with PLS_A, GSCA_R showed better performance in power than PLS_A given a small or medium size of sample (i.e. $N \leq 250$) in general, but the difference was negligible.

5 Summary and discussions

In this paper, we uncovered the limitation of previous DGPs for variance-based SEM and proposed a new DGP where components are constructed to well explain the variances of their indicators as well as those of endogenous components. Along with the development of the DGP for variance-based structural equation models, GSCA *with reflective indicators* and PLSPM *with mode A* were properly evaluated. Our simulation study showed that all modeling approaches of variance-based SEM were able to provide consistent and acceptably unbiased estimators for the parameters of

Table 3 Type I error obtained from variance-based SEM approaches across different experimental conditions in the simulation study

N_{ind}		$r=0.2$				$r=0.4$				$r=0.6$			
		GSCA$_{\text{R}}$	PLS$_{\text{A}}$	GSCA$_{\text{F}}$	PLS$_{\text{B}}$	GSCA$_{\text{R}}$	PLS$_{\text{A}}$	GSCA$_{\text{F}}$	PLS$_{\text{B}}$	GSCA$_{\text{R}}$	PLS$_{\text{A}}$	GSCA$_{\text{F}}$	PLS$_{\text{B}}$
2	$N=100$	0.05	0.05	0.04	0.04	0.06	0.06	0.05	0.05	0.07	0.07	0.05	0.05
	$N=250$	0.04	0.04	0.03	0.04	0.04	0.04	0.04	0.04	0.06	0.06	0.06	0.06
	$N=500$	0.06	0.06	0.05	0.06	0.06	0.06	0.06	0.06	0.07	0.07	0.07	0.07
	$N=1000$	0.04	0.04	0.04	0.04	0.06	0.06	0.06	0.06	0.06	0.06	0.05	0.06
4	$N=100$	0.05	0.06	0.03	0.05	0.05	0.05	0.02	0.04	0.06	0.06	0.03	0.05
	$N=250$	0.06	0.06	0.05	0.05	0.07	0.07	0.05	0.05	0.06	0.06	0.04	0.04
	$N=500$	0.05	0.04	0.04	0.04	0.05	0.05	0.04	0.04	0.07	0.06	0.06	0.06
	$N=1000$	0.05	0.05	0.05	0.05	0.06	0.06	0.06	0.05	0.06	0.06	0.06	0.06
6	$N=100$	0.04	0.05	0.01	0.04	0.07	0.06	0.01	0.04	0.07	0.06	0.01	0.04
	$N=250$	0.05	0.05	0.04	0.04	0.05	0.05	0.04	0.04	0.05	0.05	0.03	0.03
	$N=500$	0.06	0.06	0.05	0.06	0.06	0.07	0.05	0.05	0.05	0.05	0.04	0.04
	$N=1000$	0.05	0.05	0.05	0.04	0.05	0.05	0.04	0.04	0.07	0.07	0.05	0.05
8	$N=100$	0.04	0.04	0.00	0.04	0.06	0.06	0.00	0.05	0.06	0.06	0.01	0.06
	$N=250$	0.04	0.05	0.02	0.04	0.06	0.07	0.02	0.05	0.06	0.06	0.03	0.05
	$N=500$	0.05	0.06	0.05	0.06	0.05	0.05	0.03	0.04	0.05	0.05	0.04	0.04
	$N=1000$	0.04	0.04	0.04	0.04	0.06	0.06	0.05	0.04	0.05	0.06	0.05	0.05

Table 4 Statistical power of variance-based SEM approaches across different experimental conditions in the simulation study

N_{ind}		$r=0.2$				$r=0.4$				$r=0.6$			
		GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}	GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}	GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}
Weight													
2	$N=100$	0.98	0.99	0.96	0.93	1.00	1.00	0.91	0.87	1.00	1.00	0.81	0.75
	$N=250$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00	0.98	0.97
	$N=500$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$N=1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	$N=100$	0.98	0.93	0.60	0.51	1.00	0.98	0.38	0.30	1.00	1.00	0.20	0.15
	$N=250$	1.00	1.00	0.95	0.91	1.00	1.00	0.83	0.76	1.00	1.00	0.58	0.50
	$\underline{N}=500$	1.00	1.00	1.00	0.99	1.00	1.00	0.97	0.94	1.00	1.00	0.83	0.77
	$N=1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.94
6	$N=100$	0.96	0.87	0.25	0.20	1.00	0.97	0.12	0.10	1.00	1.00	0.05	0.04
	$N=250$	1.00	0.99	0.78	0.69	1.00	1.00	0.52	0.43	1.00	1.00	0.20	0.15
	$N=500$	1.00	1.00	0.95	0.92	1.00	1.00	0.81	0.74	1.00	1.00	0.39	0.33
	$N=1000$	1.00	1.00	0.99	0.99	1.00	1.00	0.95	0.93	1.00	1.00	0.65	0.58
8	$N=100$	0.97	0.86	0.09	0.09	1.00	0.96	0.04	0.04	0.99	0.99	0.02	0.02
	$N=250$	1.00	0.99	0.58	0.46	1.00	1.00	0.27	0.19	1.00	1.00	0.10	0.07
	$N=500$	1.00	1.00	0.86	0.79	1.00	1.00	0.54	0.46	1.00	1.00	0.19	0.16
	$N=1000$	1.00	1.00	0.98	0.96	1.00	1.00	0.80	0.74	1.00	1.00	0.35	0.31
Loading													
2	$N=100$	0.97	0.99	0.99	0.98	1.00	1.00	0.99	0.99	1.00	1.00	1.00	1.00
	$N=250$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$N=500$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$N=1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	$N=100$	0.97	0.97	0.91	0.86	1.00	1.00	0.96	0.93	1.00	1.00	0.99	0.98
	$N=250$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 4 (continued)

N_{ind}		$r=0.2$				$r=0.4$				$r=0.6$			
		GSCA$_R$	PLS$_A$	GSCA$_F$	PLS$_B$	GSCA$_R$	PLS$_A$	GSCA$_F$	PLS$_B$	GSCA$_R$	PLS$_A$	GSCA$_F$	PLS$_B$
	$N=500$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$N=1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	$N=100$	0.95	0.94	0.77	0.72	1.00	1.00	0.91	0.87	1.00	1.00	0.97	0.95
	$N=250$	1.00	1.00	0.99	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$N=500$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$N=1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
8	$N=100$	0.97	0.96	0.66	0.62	1.00	1.00	0.84	0.80	1.00	1.00	0.94	0.91
	$N=250$	1.00	1.00	0.99	0.97	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$N=500$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	$N=1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Path													
2	$N=100$	0.81	0.80	0.80	0.78	0.82	0.81	0.80	0.78	0.83	0.82	0.80	0.78
	$N=250$	0.93	0.92	0.92	0.92	0.93	0.92	0.92	0.91	0.93	0.92	0.92	0.91
	$N=500$	0.98	0.98	0.98	0.97	0.98	0.98	0.98	0.97	0.98	0.98	0.98	0.98
	$N=1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4	$N=100$	0.79	0.77	0.72	0.67	0.82	0.80	0.72	0.68	0.82	0.81	0.73	0.68
	$N=250$	0.92	0.91	0.91	0.88	0.92	0.92	0.90	0.89	0.93	0.92	0.91	0.89
	$N=500$	0.98	0.98	0.97	0.97	0.98	0.98	0.97	0.97	0.98	0.98	0.98	0.97
	$N=1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	$N=100$	0.80	0.77	0.65	0.59	0.82	0.81	0.65	0.59	0.82	0.82	0.66	0.60
	$N=250$	0.92	0.91	0.89	0.85	0.92	0.92	0.89	0.85	0.92	0.92	0.89	0.85
	$N=500$	0.98	0.97	0.97	0.95	0.98	0.98	0.97	0.95	0.98	0.98	0.97	0.96
	$N=1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 4 (continued)

N_{ind}		$r = 0.2$				$r = 0.4$				$r = 0.6$			
		GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}	GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}	GSCA_{R}	PLS_{A}	GSCA_{F}	PLS_{B}
8	$N = 100$	0.80	0.77	0.55	0.53	0.81	0.80	0.55	0.51	0.83	0.82	0.57	0.52
	$N = 250$	0.92	0.91	0.87	0.82	0.93	0.92	0.88	0.81	0.92	0.92	0.88	0.81
	$N = 500$	0.98	0.98	0.97	0.95	0.98	0.98	0.96	0.95	0.98	0.98	0.97	0.95
	$N = 1000$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

variance-based structural equation models. This result not only serves as empirical evidence to substantiate the appropriateness of the DGP proposed in this paper, but also cements GSCA's and PLSPM's positions as variance-based SEM approaches.

GSCA *with reflective indicators* and PLSPM *with mode A* turned out to recover parameters in a more efficient manner than GSCA *with formative indicators* and PLSPM *with mode B* under variance-based structural equation models. It would be attributed from the fact that the former approaches estimate weights while considering the variances of indicators as well as those of endogenous components. In addition, we found that compared to PLSPM *with mode A*, GSCA *with reflective indicators* provided more efficient estimators for weights and loadings. For the path coefficients, though, there were no substantial differences between PLSPM *with mode A* and GSCA *with reflective indicators*. These patterns were the same for GSCA *with formative indicators* and PLSPM *with mode B*, which is in accord with the simulation result of Hair et al. (2017).

In terms of hypothesis testing, GSCA *with reflective indicators* and PLSPM *with mode A* outperformed GSCA *with formative indicators* and PLSPM *with mode B* as well. While GSCA *with reflective indicators* and PLSPM *with mode A* controlled type I error at the pre-specified significance level (i.e. 0.05) equally well, GSCA *with reflective indicators* showed slightly higher power than PLSPM *with mode A*. Notably, GSCA *with formative indicators* outperformed PLSPM *with mode B* in statistical power. This tendency became more salient when the true path coefficient was prescribed to be low (i.e., b = 0.15; see Table 3 in Supplementary materials). This result is in contrast to the one reported by Hair et al. (2017) that PLSPM *with mode B* was better than GSCA *with formative indicators* in statistical power. The different results between the two studies may be due to the differences in prescribed signs of the path coefficients in a given structural model: given that standard errors are fixed, a statistical power is likely to be affected by the magnitude of biases, which are further dependent upon the signs of path coefficients in the structural model. We observed that a change in the sign of a path coefficient in DGP also influenced patterns of biases for all the other path coefficients in both GSCA and PLSPM. The effect was rather arbitrary so that some changes were advantageous to GSCA and the other to PLSPM. In this sense, higher power of PLSPM *with mode B* or GSCA *with formative indicators* in each study could stem from the prescribed values of path coefficients being advantageous to either of them.

Lastly, our simulation showed that the effects of the number of indicators per component and correlation between indicators for each component were different across modeling approaches of variance-based SEM. GSCA *with reflective indicators* and PLSPM *with mode A* benefited from the large number of indicators per component and the high level of average correlation between indicators for a component, as in factor analysis (Marsh et al. 1998), whereas those conditions rather had negative impacts on GSCA *with formative indicators* and PLSPM *with mode B*. This finding is consistent with Becker et al. (2013), in which an increase in correlation between indicators was found to be associated with lower RMSE for PLSPM *with mode A* but higher RMSE for PLSPM *with mode B*. According to their explication, high correlations between indicators can lead to multicollinearity, subsequently aggravating stability of weight estimation, especially for PLSPM *with mode B*. With indicators for a component being more correlated to each other, the multicollinearity problem is expected to be worse. Conversely, PLSPM *with mode A* is relatively free from this issue as it estimates weights by correlation (Becker et al. 2013; Rigdon 2012). However, it cannot be a sufficient reason why PLSPM *with mode A* or GSCA *with reflective indicators* performs better, despite of the risk of multicollinearity. It may be reasonable to conjecture that including additional indicators leads to an increase in the number of equations to be considered in estimating weight parameters for GSCA *with reflective indicators* and PLSPM *with mode A*, which in turn, would make their estimation process more stable.

Based on our findings, we provide a couple of recommendations for practitioners to utilize variance-based SEM approaches. First, if you want to construct nomological components in SEM framework or run SEM using the measures based on principal component analysis, you should select GSCA *with reflective indicators* or PLSPM *with mode A*. Without the prior preference on the two approaches, we suggest using GSCA *with reflective indicators* since it can help construct components more precisely. Using GSCA *with formative indicators* or PLSPM *with mode B* is recommended in particular case when the construction of components is specifically aimed at explaining the variances of endogenous components only.

Second, if you use GSCA *with reflective indicators* or PLSPM *with mode A*, it would be acceptable to increase the number of indicators with high correlation if possible. On the other hand, when drawing on GSCA *with formative indicators* or PLSPM *with mode B*, you should sift a few indicators with low correlation through a set of candidate indicators. However, you need to be cautious to add or remove indicators because such change may alter the conceptual meaning of components (Bollen 2017; Jarvis et al. 2003).

In spite of our comprehensive investigation on relative performances of the four SEM approaches, we overlooked two important criteria to evaluate their performance. The first one is another important property of an estimator—robustness to model-misspecification. As mentioned in Sect. 2, the limited estimation method adopted in PLSPM might allow PLSPM to be robust to model-misspecification, even though, in Hwang et al. (2010) and Hwang and Takane (2014)'s simulation study, the evidence to support this hypothesis was not found under factor-based structural equation models. On the other hand, GSCA *with reflective indicators* and PLSPM *with mode A* could be practically more robust to model-misspecification. SEM techniques are typically applied after all measurements are sufficiently validated (Bollen 1989; Chin 1998). In other words, SEM techniques are generally utilized in the situation where researchers are unsure of the true structural model but with validated measurement tools. In this case, GSCA *with formative indicators* and PLSPM *with mode B* would be subject to biased estimates for path coefficients, because their estimation of weights is contingent solely on allegedly specified paths among components. In contrast, GSCA *with reflective indicators* and PLSPM *with mode A* consider the variances of indicators as well in weights estimation, which may allow their estimators to be more robust against model-misspecification. Further research is required to test these hypotheses on the relative robustness under the various model constellations with components.

The second missing criteria is the one in the utterly different framework from parameter recovery—predictability. GSCA and PLSPM are essentially "prediction-oriented" approaches to SEM in that they can predict individual scores of every endogenous variable specified in the model, beyond simply estimating parameters (Cho et al. 2019; Sharma et al. 2018; Wold 1982). However, their relative predictive performance has never been properly evaluated even though its importance was acknowledged in the variance-based SEM scholarly community (Shmueli et al. 2016). Thus, the future research to compare the two variance-based SEM approaches needs to consider their performance on predictability. The DGP proposed in this paper may facilitate this future research.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Anderson EW, Fornell C (2002) Foundations of the American customer satisfaction index. Total Qual Manag 11(7):869–882. https://doi.org/10.1080/09544120050135425

Becker J-M, Rai A, Rigdon E (2013) Predictive validity and formative measurement in structural equation modeling: embracing practical relevance. In: Proceedings of the 34th international conference on information systems (ICIS), Milan, Italy

Bollen KA (1989) Structural equations with latent variables. Wiley, Hoboken. https://doi.org/10.1002/9781118619179

Bollen KA (1996) An alternative two stage least squares (2SLS) estimator for latent variable equations. Psychometrika 61(1):109–121. https://doi.org/10.1007/BF02296961

Bollen KA (2011) Evaluating effect, composite, and causal indicators in structural equation models. MIS Q 35(2):359. https://doi.org/10.2307/23044047

Bollen KA, Kirby JB, Curran PJ, Paxton PM, Chen F (2007) Latent variable models under misspecification two-stage least squares (2SLS) and maximum likelihood (ML) estimators. Soc Methods Res 36(1):48–86. https://doi.org/10.1177/0049124107301947

Chai T, Draxler RR (2014) Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. Geosci Model Dev 7(3):1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chin WW (1998) The partial least squares approach for structural equation modeling. In: Marcoulides GA (ed) Methodology for business and management. Modern methods for business research. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp 295–336

Cho G, Jung K, Hwang H (2019) Out-of-bag prediction error: a cross validation index for generalized structured component analysis. Multivar Behav Res. https://doi.org/10.1080/00273171.2018.1540340

Cronbach LJ, Meehl PE (1955) Construct validity in psychological tests. Psychol Bull 52(4):281–302

Dijkstra TK (2017) A perfect match between a model and a mode. In: Partial least squares path modeling: basic concepts, methodological issues and applications. Springer, Berlin, pp 55–80. https://doi.org/10.1007/978-3-319-64069-3_4

Efron B (1979) Bootstrap methods: another look at the jackknife. Ann Stat 7(1):1–26. https://doi.org/10.1214/aos/1176344552

Eklöf JA, Westlund AH (2002) The pan-European customer satisfaction index programme—current work and the way ahead. Total Qual Manag 13(8):1099–1106. https://doi.org/10.1080/09544120200000005

Fomby TB, Johnson SR, Hill RC (2011) Advanced econometric methods. Advanced econometric methods. Springer, New York. https://doi.org/10.1007/978-1-4419-8746-4

Fornell C, Johnson MD, Anderson EW, Cha J, Bryant BE (1996) The American customer satisfaction index: nature, purpose, and findings. J Mark 60(4):7. https://doi.org/10.2307/1251898

Gallier J, Quaintance J (2019) Algebra, topology, differential calculus, and optimization theory for computer science and engineering. Philadelphia, PA. Retrieved Feb 20, 2019, from https://www.cis.upenn.edu/~jean/math-basics.pdf

Gerbing DW, Hamilton JG (1994) The surprising viability of a simple alternate estimation procedure for construction of large-scale structural equation measurement models. Struct Equ Model A Multidiscip J 1(2):103–115. https://doi.org/10.1080/10705519409539967

Hair JF, Hult GTM, Ringle CM, Sarstedt M, Thiele KO (2017) Mirror, mirror on the wall: a comparative evaluation of composite-based structural equation modeling methods. J Acad Mark Sci 45(5):616–632. https://doi.org/10.1007/s11747-017-0517-x

Hwang H, Takane Y (2014) Generalized structured component analysis: a component-based approach to structural equation modeling. Chapman and Hall/CRC Press, New York

Hwang H, Malhotra NK, Kim Y, Tomiuk MA, Hong S (2010) A comparative study on parameter recovery of three approaches to structural equation modeling. J Mark Res 47(4):699–712. https://doi.org/10.2139/ssrn.1585305

Hwang H, Takane Y, Tenenhaus A (2015) An alternative estimation procedure for partial least squares path modeling. Behaviormetrika 42(1):63–78. https://doi.org/10.2333/bhmk.42.63

Hwang H, Sarstedt M, Cheah JH, Ringle CM (2019) A concept analysis of methodological research on composite-based structural equation modeling: bridging PLSPM and GSCA. Behaviormetrika. https://doi.org/10.1007/s41237-019-00085-5

Jarvis CB, MacKenzie SB, Podsakoff PM (2003) A critical review of construct indicators and measurement model misspecification in marketing and consumer research. J Consum Res 30(2):199–218. https://doi.org/10.1086/376806

Jöreskog KG (1970) Estimation and testing of simplex models. Br J Math Stat Psychol 23(2):121–145. https://doi.org/10.1111/j.2044-8317.1970.tb00439.x

Jöreskog KG (1978) Structural analysis of covariance and correlation matrices. Psychometrika 43(4):443–477. https://doi.org/10.1007/BF02293808

Lay DC, Lay SR, McDonald JJ (2015) Linear algebra and its applications, 576

Lohmöller J-B (1989) Latent variable path modeling with partial least squares. Springer, New York. https://doi.org/10.1007/978-3-642-52512-4

Marsh HW, Hau KT, Balla JR, Grayson D (1998) Is more ever too much? The number of indicators per factor in confirmatory factor analysis. Multivar Behav Res 33(2):181–220. https://doi.org/10.1207/s15327906mbr3302_1

Rego LL, Morgan NA, Fornell C (2013) Reexamining the market share-customer satisfaction relationship. J Mark 77(5):1–20. https://doi.org/10.1509/jm.09.0363

Reinartz W, Haenlein M, Henseler J (2009) An empirical comparison of the efficacy of covariance-based and variance-based SEM. Int J Res Mark 26(4):332–344. https://doi.org/10.1016/j.ijresmar.2009.08.001

Rigdon EE (2012) Rethinking partial least squares path modeling: in praise of simple methods. Long Range Plan 45(5–6):341–358. https://doi.org/10.1016/j.lrp.2012.09.010

Roldán JL, Sánchez-Franco MJ (2012) Variance-based structural equation modeling: guidelines for using partial least squares in information systems research. In: Mora M, Gelman O, Steenkamp AL, Raisinghani M (eds) Research methodologies, innovations and philosophies in software systems engineering and information systems. IGI Global, Hershey, pp 193–221. https://doi.org/10.4018/978-1-4666-0179-6.ch010

Sarstedt M, Hair JF, Ringle CM, Thiele KO, Gudergan SP (2016) Estimation issues with PLS and CBSEM: where the bias lies! J Bus Res 69(10):3998–4010. https://doi.org/10.1016/j.jbusres.2016.06.007

Sharma PN, Shmueli G, Sarstedt M, Danks N, Ray S (2018) Prediction-oriented model selection in partial least squares path modeling. Decis Sci 00:1–41. https://doi.org/10.1111/deci.12329

Shmueli G, Ray S, Velasquez Estrada JM, Chatla SB (2016) The elephant in the room: predictive performance of PLS models. J Bus Res 69(10):4552–4564. https://doi.org/10.1016/J.JBUSRES.2016.03.049

Tenenhaus M (2008) Component-based structural equation modelling. Total Quality Manag Bus Excell 19(7–8):871–886. https://doi.org/10.1080/14783360802159543

Tenenhaus M, Esposito Vinzi V, Chatelin Y-M, Lauro C (2005) PLS path modeling. Comput Stat Data Anal 48(1):159–205. https://doi.org/10.1016/J.CSDA.2004.03.005

Wold H (1982) Models for knowledge. In: Gani J (ed) The making of statisticians. Springer, New York, pp 189–212. https://doi.org/10.1007/978-1-4613-8171-6_1

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.