An approach to structural equation modeling with both factors and components: Integrated generalized structured component analysis

Article *in* Psychological Methods · July 2020

DOI: 10.1037/met0000336

CITATIONS
13

READS
1,255

7 authors, including:

Heungsun Hwang
McGill University
103 PUBLICATIONS **1,858** CITATIONS
SEE PROFILE

Gyeongcheol Cho
McGill University
18 PUBLICATIONS **146** CITATIONS
SEE PROFILE

Carl F. Falk
McGill University
46 PUBLICATIONS **1,098** CITATIONS
SEE PROFILE

Running Head: AN APPROACH TO SEM WITH FACTOR AND COMPONENT

An approach to structural equation modeling with both factors and components: Integrated generalized structured component analysis

Heungsun Hwang[1], Gyeongcheol Cho[1], Kwanghee Jung[2], Carl Falk[1], Jessica Flake[1], Min Jin Jin[3,4], Seung Hwan Lee[3,5]

1. McGill University, Montreal, Canada

2. Texas Tech University, Lubbock, USA

3. Clinical Emotion and Cognition Research Laboratory, Inje University, Goyang, Korea

4. Chung-Ang University, Seoul, Korea

5. Ilsan-Paik Hospital, Inje University, Goyang, Korea

May 26, 2020

Author Note

The research reported in this article was supported by the Ministry of Education and the National Research Foundation of Korea (NRF-2019S1A5A2A03052192) to the first author and by the Brain Research Program through the National Research Foundation of Korea from the Ministry of Science, ICT & Future Planning (NRF-2015M3C7A1028252) to the last author.

Correspondence concerning the article should be addressed to: Heungsun Hwang, Department of Psychology, McGill University, 2001 McGill College Avenue, Montreal, QC H3A 1G1, Canada. Email: heungsun.hwang@mcgill.ca.

An approach to structural equation modeling with both factors and components: Integrated generalized structured component analysis

Abstract

In this paper, we propose integrated generalized structured component analysis (IGSCA), which is a general statistical approach for analyzing data with both components and factors in the same model, simultaneously. This approach combines generalized structured component analysis (GSCA) and generalized structured component analysis with measurement errors incorporated ($GSCA_M$) in a unified manner and can estimate both factor- and component-model parameters, including component and factor loadings, component and factor path coefficients, and path coefficients connecting factors and components. We conduct two simulation studies to investigate the performance of IGSCA under models with both factors and components. The first simulation study assesses how existing approaches for structural equation modeling and IGSCA recover parameters. This study shows that only consistent partial least squares (PLSc) and IGSCA yield unbiased estimates of all parameters, whereas the other approaches always provided biased estimates of several parameters. As such, we conduct a second, extensive simulation study to evaluate the relative performance of the two competitors (PLSc and IGSCA), considering a variety of experimental factors (model specification, sample size, the number of indicators per factor/component, and exogenous factor/component correlation). IGSCA exhibits better performance than PLSc under most conditions. We also present a real data application of IGSCA to the study of genes and their influence on depression. Finally, we discuss the implications and limitations of this approach, and recommendations for future research.

Keywords: Structural equation modeling; common factor; component; covariance structure analysis; partial least squares path modeling; generalized structured component analysis;

consistent partial least squares; generalized structured component analysis with measurement errors incorporated, gene, depression

Translational Abstract

As psychology and many other sciences become interdisciplinary, there is an ever-increasing need for accommodating two statistical representations of constructs, i.e., common factors and components, at the same time and examining their relationships to aid in an understanding of human behavior and cognition from more diverse perspectives. For example, psychologists have increasingly been interested in the influences of genetic variation and/or altered brain activities on the variation of psychological constructs in cognition, personality, or mental disorders. Such psychological constructs have typically been represented by common factors, whereas genetic or imaging constructs, such as genes and brain regions, by components. We thus propose a general statistical approach, called integrated generalized structured component analysis (IGSCA), for estimating structural equation models with both factors and components. This approach combines two versions of generalized structured component analysis in a unified manner to estimate both factor- and component-model parameters, including component and factor loadings, component and factor path coefficients, and path coefficients connecting components and factors. We report on two simulation studies that establish IGSCA as a sensible method for estimating models with both factors and components, as compared to existing approaches. Finally, we demonstrate the potential of IGSCA in real data applications with an investigation of the effects of multiple genes on depression.

Structural equation modeling (SEM) is a general multivariate framework for specifying and examining a system of linear models that involve observed and latent variables. In the SEM literature, a latent variable is synonymous with a (common) factor (e.g., McDonald, 1999; Treiblmaier, Bentler, & Mair, 2011), which is not directly measurable and exists as an entity independent of observed variables, yet serving as the sole source for the covariation of the observed variables (e.g., Borsboom, Mellenbergh, & van Heerden, 2003). Observed variables linked to a factor are called effect or reflective indicators (Bollen, 1989, p. 65; Edwards & Bagozzi, 2000; MacCallum & Browne, 1993). If the model assumption holds, then the unique factor of an effect indicator, which is unexplained by the common factor, is assumed to be the sum of the specific factor and random measurement error of the effect indicator (e.g., Mulaik, 2010, pp. 132–133). Factors have been standard representations of various psychological constructs in cognition, personality, emotion, learning, and mental disorders (e.g., Borsboom et al., 2003). The measurement model that specifies the relationships between factors and their effect indicators is referred to as a reflective model.

In multivariate modeling, researchers can also contemplate a (weighted) composite or component of observed variables, which varies as a deterministic linear function of observed variables (e.g., Jarvis, MacKenzie, & Podsakoff, 2003). Observed variables forming a component are referred to as composite indicators (Bollen & Bauldry, 2011; Grace & Bollen, 2008). For example, according to the Grid-Enabled Measures (GEM) database created by the National Cancer Institute (www.gem-beta.org/public/home.aspx), activity space is defined as a “geographic area that people visit,” indicating that it may be treated as a (weighted) sum of all the physical locations that individuals visited. In the same database, body composition is considered an aggregation of body fat, muscle, and bone mineral in the entire body or specific

body sites, such as the hip, spine, and limbs. Furthermore, in genetic and neuroimaging studies, candidate genotypes, such as single nucleotide polymorphisms (SNPs), and brain voxel-level phenotypes are observed measures at particular genomic or brain locations, indicating that a different set of SNPs or voxels constitutes a different gene or brain region. Thus, a gene or brain region can be seen as a biological composite of SNPs or voxels, respectively (e.g., Jung, Takane, Hwang, & Woodward, 2012, 2016; Lee et al., 2016; Romdhani, Hwang, Paradis, Roy-Gagnon, & Labbe, 2015). The measurement model that specifies the relationships between components and their composite indicators is known as a composite model (e.g., Bollen & Diamantopoulos, 2017; Diamantopoulos & Winklhofer, 2001; Henseler et al., 2014; Schuberth, Henseler, & Dijkstra, 2018).

Depending on whether the measurement model involves either factors or components, SEM has evolved into two domains – factor-based and component-based (e.g., Jöreskog & Wold, 1982; Rigdon, 2012; Rigdon, Sarstedt, & Ringle, 2017; Tenenhaus, 2008). As the names denote, factor-based SEM involves the reflective model with factors only, whereas component-based SEM involves the composite model with components only. Methodologically, covariance structure analysis (CSA; Jöreskog, 1970, 1978) was developed for factor-based SEM, whereas partial least squares path modeling (PLSPM; Lohmöller, 1989; Wold, 1966, 1973, 1982) and generalized structured component analysis (GSCA; Hwang & Takane, 2004) were for component-based SEM. More recently, consistent partial least squares (PLSc; Dijkstra, 2010; Dijkstra & Henseler, 2015a, 2015b) and generalized structured component analysis with measurement errors incorporated ($GSCA_M$; Hwang, Takane, & Jung, 2017) were also developed for factor-based SEM.

The two SEM domains largely remain mutually exclusive, adhering to only either of factors and components. This methodological status quo is a barrier to a demand for representing both factors and components in the same model. Data with both factors and components can emerge from various social, behavioral and health sciences, and examining their relations facilitates an understanding of human behavior and mind from more diverse perspectives. For example, owing to ever-increasing advances in measurement tools, psychologists have more access to both genetic and neuroimaging data collected from the same individuals, and combine the two data sources with behavioral or cognitive outcomes to identify neuromechanisms linked to psychological disorders, clinical symptoms or cognitive tasks (e.g., Bookheimer et al., 2000; Hariri & Weinberger, 2003; Rasetti & Weinberger, 2011). SEM can a sensible choice for specifying and examining such gene-brain-behavior/cognition relationships based on theories or knowledge from previous studies, for example, in genome-wide whole brain association (e.g., Wang, Li, & Hakonarson, 2010) and brain connectivity analysis (e.g., Birnbaum & Weinberger, 2013; Friston, 1994). As stated earlier, it would be more plausible to represent genes and brain regions as components of SNPs and voxels, respectively, whereas psychological disorders, such as depression or post-traumatic stress disorder (PTSD), may be represented as factors. Relying on either of the SEM domains would not be sufficient for models that include both factors and components and their inter-relationships, such as a model where several genes influence PTSD severity.

In this paper, we propose a statistical approach to SEM with both factors and components. The proposed approach is a unified framework of GSCA and GSCA$_M$. As stated earlier, GSCA was developed for component-based SEM only. Conversely, GSCA$_M$ is its recent extension for factor-based SEM only, which contemplates both common and unique parts of

each indicator, as postulated in common factor analysis or factor-based SEM. As a consequence, $GSCA_M$ produces unbiased estimates of loadings and path coefficients in factor-based structural equation models.

By combining GSCA and $GSCA_M$ into a single framework, the proposed approach can provide estimates of parameters in structural equation models that include both factors and components. We term this approach *integrated generalized structured component analysis* (IGSCA). Strictly speaking, IGSCA still falls within the domain of component-based SEM, because it obtains components regardless of whether unique parts of indicators are taken into account, as will be discussed in the next section. Nevertheless, it simultaneously provides the estimates of factor loadings for reflective indicators and component loadings for composite indicators, as well as those of path coefficients involving factors and components.

It is worth mentioning that IGSCA can be chosen for estimating models with both factors and components. As summarized in Table 1, when the model contains factors only, IGSCA is expected to perform at most as similarly as to CSA or other methods for factor-based SEM. When the model includes components only, IGSCA is likely to perform comparably to GSCA at the most. Thus, in such situations modeling either factors or components only, there will be little practical advantages of using IGSCA. As with other SEM methods, furthermore, IGSCA does not necessarily protect against the situation where the researcher fits a model with an incorrectly specified dimension (e.g., a factor is misspecified as a component or vice versa), and we emphasize that theory is of ultimate importance in choosing an appropriate model.

In the next section, we provide the technical underpinnings of the proposed approach - IGSCA, including model specification and parameter estimation. We then discuss how existing approaches for factor- or component-based SEM and IGSCA are expected to estimate models

containing factors and components at the same time. In particular, we provide new theoretical expectations regarding parameters that appear only in models with both factors and components: path coefficients connecting factors and components. We subsequently conduct two simulation studies to empirically examine the performance of the existing SEM approaches and IGSCA. Next, we apply IGSCA to real data to further demonstrate its empirical usefulness. In this application, we investigate the effects of different genes on depression severity in a sample of Korean participants. Lastly, we summarize the previous sections and discuss the implications and limitations of the proposed approach.

Integrated Generalized Structured Component Analysis

Similarly to GSCA and GSCA_M, IGSCA involves specification of three sub-models: measurement, structural, and weighted relation models. The measurement model specifies the relationships between indicators and factors/components, whereas the structural model expresses the relationships between factors/components. CSA and PLSPM also involve these two sub-models. The weighted relation model, which is unique to GSCA and GSCA_M, is used to explicitly specify a component. Let $\mathbf{z}_1$ and $\mathbf{z}_2$ denote vectors of composite and effect indicators, respectively. Let $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ denote vectors of components and factors associated with $\mathbf{z}_1$ and $\mathbf{z}_2$, respectively. Assume that all indicators, components and factors are standardized to have zero means and unit variances. Let $\mathbf{C}_1$ and $\mathbf{C}_2$ denote matrices of loadings relating $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ to $\mathbf{z}_1$ and $\mathbf{z}_2$, respectively. Let $\mathbf{u}$ denote a vector of unique variables. Let $\mathbf{D}$ denote a diagonal matrix of unique loadings.

The measurement model for IGSCA is a combination of those for GSCA and GSCA_M. Specifically, the measurement model for GSCA is given as

$$\mathbf{z}_1 = \mathbf{C}_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1, \quad (1)$$

where $\boldsymbol{\varepsilon}_1$ is the residual term that represents the portion of $\mathbf{z}_1$ left unexplained by $\boldsymbol{\gamma}_1$, as in linear regression (Hwang & Takane, 2014, Chapter 2). The measurement model for GSCA_M is given as

$$\mathbf{z}_2 = \mathbf{C}_2\boldsymbol{\gamma}_2 + \mathbf{Du} + \boldsymbol{\varepsilon}_2, \quad (2)$$

where $\mathbf{C}_2\boldsymbol{\gamma}_2$ and $\mathbf{Du}$ indicate common and unique parts of $\mathbf{z}_2$, respectively, and $\boldsymbol{\varepsilon}_2$ denotes the portion of $\mathbf{z}_2$ left unexplained by their common and unique parts (Hwang et al., 2017). If the common factor analytic model also holds for a sample, $\boldsymbol{\varepsilon}_2$ will be zero (Velicer & Jackson, 1990). We assume that $\boldsymbol{\gamma}_2$ is uncorrelated with $\mathbf{u}$, i.e., $\boldsymbol{\gamma}_2\mathbf{u}' = \mathbf{u}\boldsymbol{\gamma}_2' = \mathbf{0}$, and $\mathbf{u}$ is columnwise orthonormalized, i.e., $\mathbf{uu}' = \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. Thus, the main difference between (1) and (2) is that the latter additionally considers the unique parts of indicators.

The measurement model for IGSCA is then given by

$$\begin{bmatrix}\mathbf{z}_1\\ \mathbf{z}_2\end{bmatrix} = \begin{bmatrix}\mathbf{C}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\gamma}_1\\ \boldsymbol{\gamma}_2\end{bmatrix} + \begin{bmatrix}\mathbf{0}\\ \mathbf{Du}\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}$$

$$\mathbf{z} = \mathbf{C}\boldsymbol{\gamma} + \mathbf{s} + \boldsymbol{\varepsilon}, \quad (3)$$

where $\mathbf{z} = [\mathbf{z}_1; \mathbf{z}_2]$, $\mathbf{C} = \begin{bmatrix}\mathbf{C}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{C}_2\end{bmatrix}$, $\boldsymbol{\gamma} = [\boldsymbol{\gamma}_1; \boldsymbol{\gamma}_2]$, $\mathbf{s} = [\mathbf{0}; \mathbf{Du}]$, and $\boldsymbol{\varepsilon} = [\boldsymbol{\varepsilon}_1; \boldsymbol{\varepsilon}_2]$.

Similarly, the weighted relation model for IGSCA is a simple mixture of those for GSCA and GSCA_M. The weighted relation model for GSCA is given as

$$\boldsymbol{\gamma}_1 = \mathbf{W}_1\mathbf{z}_1, \quad (4)$$

where $\mathbf{W}_1$ is a matrix of weights assigned to $\mathbf{z}_1$ (Hwang & Takane, 2014, Chapter 2). The weighted relation model for GSCA_M is given as

$$\boldsymbol{\gamma}_2 = \mathbf{W}_2(\mathbf{z}_2 - \mathbf{Du}), \quad (5)$$

where $\mathbf{W}_2$ is a matrix of weights assigned to $\mathbf{z}_2$, whose unique parts are removed (Hwang et al., 2017).

Then, the weighted relation model for IGSCA is expressed as

$$\begin{bmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_2 \end{bmatrix} \begin{bmatrix} \mathbf{z}_1 \\ \mathbf{z}_2 - \mathbf{Du} \end{bmatrix}$$

$$\boldsymbol{\gamma} = \mathbf{W}\mathbf{z}^*, \tag{6}$$

where $\mathbf{z}^* = [\mathbf{z}_1; \mathbf{z}_2 - \mathbf{Du}]$. This sub-model shows that both $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ represent components in nature rather than factors. Nonetheless, as shown in (5), $\boldsymbol{\gamma}_2$ is components of effect indicators, whose unique parts are removed. In this way, measurement errors in effective indicators are accounted for, rendering the parameters involving $\boldsymbol{\gamma}_2$ close to those from factor-based SEM (Hwang et al., 2017).

Let **B** denote a matrix of path coefficients relating $\boldsymbol{\gamma}$ among themselves. The structural model for IGSCA is expressed as

$$\boldsymbol{\gamma} = \mathbf{B}\boldsymbol{\gamma} + \boldsymbol{\zeta}, \tag{7}$$

where $\boldsymbol{\zeta}$ is the residual term for $\boldsymbol{\gamma}$. The **B** matrix contains path coefficients relating components to factors, as well as those among either factors or components only.

IGSCA integrates the sub-models into a unified formulation, as follows.

$$\begin{bmatrix} \mathbf{z} \\ \boldsymbol{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{C} \\ \mathbf{B} \end{bmatrix} \boldsymbol{\gamma} + \begin{bmatrix} \mathbf{s} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon} \\ \boldsymbol{\zeta} \end{bmatrix}$$

$$\boldsymbol{\psi} = \mathbf{A}\boldsymbol{\gamma} + \mathbf{v} + \mathbf{e}, \tag{8}$$

where $\boldsymbol{\psi} = [\mathbf{z}; \boldsymbol{\gamma}]$, $\mathbf{A} = [\mathbf{C}; \mathbf{B}]$, $\mathbf{v} = [\mathbf{s}; \mathbf{0}]$, and $\mathbf{e} = [\boldsymbol{\varepsilon}; \boldsymbol{\zeta}]$. This is called the IGSCA model. It is essentially of the same form as the GSCA_M model. However, (8) also includes components in the measurement model and path coefficients connecting factors and components in the structural model, which is distinctive from the GSCA_M model. IGSCA does not make distributional assumptions or impose structural constraints on **e**. Also, it does not impose any specific constraints on the covariance structure of composite indicators, leaving them to be correlated or

uncorrelated freely. On the other hand, in the paper, IGSCA imposes the ordinary, conditional independence among effect indicators (Mulaik, 2010, p. 167), keeping **D** to be a diagonal matrix, although this constraint can be relaxed by specifying **D** as a symmetric matrix rather than a diagonal one.

Let $\mathbf{e}_i$ denote the entire residual term in (8) for a single observation of a sample of N observations ($i = 1, \ldots, N$). To estimate all parameters in **W**, **A** and **v**, IGSCA seeks to minimize the following least squares criterion

$$\phi = \sum_{i=1}^{N} \mathbf{e}_i' \mathbf{e}_i \,, \tag{9}$$

subject to diag($\boldsymbol{\gamma}\boldsymbol{\gamma}'$) = $N\mathbf{I}$, $\boldsymbol{\gamma}_2\mathbf{u}' = \mathbf{u}\boldsymbol{\gamma}_2' = \mathbf{0}$, and $\mathbf{u}\mathbf{u}' = \mathbf{I}$. This criterion is essentially of the same form as that for GSCA$_M$. A main difference is that **v** (more precisely, **s** in (3)) contains additional zero elements that correspond to composite indicators $\mathbf{z}_1$. Thus, virtually the same iterative algorithm for GSCA$_M$ can be used to minimize the criterion. This algorithm alternates several steps, each of which estimates a set of parameters in a least squares sense with the other sets fixed, until no substantial differences in parameter estimates occurs between iterations. Specifically, in step 1, the two sets of component weights $\mathbf{W}_1$ and $\mathbf{W}_2$ in (6) are estimated via the corresponding step in the original GSCA algorithm for $\mathbf{W}_1$ or in the GSCA$_M$ algorithm for $\mathbf{W}_2$. In steps 2 and 3, the loadings and path coefficients in (3) and (7) are estimated. In the last step, the unique variables and loadings in (3) are updated for effect indicators (refer to Hwang et al., 2017). Although this estimation procedure of IGSCA amounts to a simple adoption of the existing algorithms, it is efficient to estimate both factor and component loadings as well as path coefficients involving factors and components, as will be demonstrated through the analyses of both simulated and real data in the next sections.

By minimizing the above least squares criterion, IGSCA estimates components (or their weights) in such a way that they maximize the variances of all endogenous indicators and components. If only the measurement model is specified, the components are obtained in the same way as those in (confirmatory) principal component analysis (e.g., Kiers, Takane, & ten Berge, 1996; ten Berge, 1993, p. 43), which maximize the variances of their own indicators (with or without their unique parts removed). If both measurement and structural models are specified, the components are obtained by applying principal component analysis and linear regression concurrently, maximizing the variances of their own indicators and endogenous components. Thus, the components in IGSCA can always be interpreted as low-dimensional approximations of their own indicators in the measurement model. This helps circumvent potential interpretational confounding of components that are typically considered in the composite model, where only component weights are specified (without component loadings) and estimated to predict their endogenous variables, so that their meaning can change depending on what they predict in the structural model (e.g., Howell, Breivik, & Wilcox, 2007; Kim, Shin, & Grover, 2010; Treiblmaier et al., 2011). In addition, it enables IGSCA to accommodate purely endogenous components that do not have effects on other components in the structural model.

Once all parameters are estimated, IGSCA can calculate indirect effects of variables as the products of relevant direct effects and also compute total effects as the sums of their direct and indirect effects. Examining indirect or total effects can be useful for providing additional information that testing direct effects only cannot give. For example, psychological disorders (i.e., factors) can be predicted by different genes (i.e., components), each of which is associated with multiple SNPs, as considered partly in the Empirical Application section. If researchers are further interested in the effects of an individual SNP for a given gene on psychological disorders,

these effects can be computed as the products of the SNP's component weight in $\mathbf{W}_1$ and the gene's direct effects (signified by its path coefficients in **B**) on psychological disorders. The indirect effects can be seen as the differential influences of an indicator on dependent variables mediated by its component.

Like GSCA and $GSCA_M$, IGSCA is also a non-parametric or distribution-free approach in the sense that it estimates parameters without recourse to distributional assumptions such as multivariate normality of indicators. As a trade-off of no reliance on distributional assumptions, it cannot estimate the standard errors of parameter estimates based on asymptotic (normal-theory) approximations. Instead, it utilizes the bootstrap method (Efron, 1979, 1982) to obtain the standard errors or confidence intervals of parameter estimates non-parametrically.

Current Statistical Approaches' Expected Behaviors of Estimating Models with Both Factors and Components

CSA is the de facto standard statistical approach for estimating factor-based structural equation models, although $GSCA_M$ and PLSc can also be used for estimating such models, as discussed earlier. These approaches generally provide unbiased estimates of factor loadings for effect indicators and factor path coefficients relating factors among themselves in factor-based models. On the other hand, PLSPM and GSCA are the main statistical approaches for estimating component-based structural equation models. PLSPM and GSCA generally yield unbiased estimates of component loadings for composite indicators and component path coefficients connecting components to each other in component-based models (e.g., Cho & Choi, 2019; Sarstedt, Hair, Ringle, Thiele, & Gudergan, 2016).

When a factor-based SEM approach (CSA, PLSc, or $GSCA_M$) is applied to models with both factors and components, it is expected to provide unbiased estimates of only factor loadings and factor path coefficients, but negatively biased estimates of component loadings and positively biased estimates of component path coefficients (Rhemtulla, van Bork, & Borsboom, 2020; Sarstedt et al., 2016). Conversely, when a component-based SEM approach (PLSPM or GSCA) is applied to models with factors and components, it is expected to provide unbiased estimates of only component loadings and component path coefficients, but positively biased estimates of factor loadings and negatively biased (attenuated) estimates of factor path coefficients (e.g., Dijkstra, 2010; Hwang, Malhotra, Kim, Tomiuk, & Hong, 2010; Velicer & Jackson, 1990).

A unique set of parameters that appears only in models with both factors and components are the path coefficients connecting factors and components. As we derive theoretically in Appendix 1, factor-based SEM approaches, such as CSA and $GSCA_M$, are expected to overestimate these path coefficients, whereas component-based SEM approaches, i.e., PLSPM and GSCA, are expected to underestimate them.

Prior to IGSCA, PLSc is thus far the only statistical approach that can be used for estimating models with both factor and components. We provide a brief description of this approach because it is a relatively novel development in the SEM literature. PLSc is a bias correction method that is built on Nunnally and Bernstein's (1994, pp. 241, 257) adjustment formula for the attenuated structural/path coefficients of components based on consistent reliabilities of the components. It begins with the assumption that the true measurement model is a particular type of reflective model, the so-called basic design (Wold, 1982), where each factor underlies at least two indicators, each of which loads on one and only one factor. This basic

design is analogous to what is known as an independent clusters structure in factor-based SEM (Flora, 2018, p. 267). PLSc then carries out two main steps sequentially. In the first step, PLSPM is applied to estimate component weights and components. In the second, a new reliability, denoted by ρ_A, for each component is calculated using the weight estimates, and subsequently used for estimating factor path coefficients. This reliability is also used to obtain consistent factor loading estimates per factor. PLSc was initially developed as a computationally efficient alternative for estimating loadings and path coefficients in factor-based models because the PLSPM algorithm usually converges rapidly (Dijkstra, 2010; Dijkstra & Henseler, 2015b). However, it was soon recognized that PLSc can also be used for estimating the path coefficient relating a component to a factor in the case that no ρ_A-based correction is applied to the component (Dijkstra & Henseler, 2015b). In the same way, it can estimate the loadings of both components and factors. Thus, PLSc is expected to provide unbiased estimates of all parameters in models with both factors and components, if the basic design holds for the reflective model (Dijkstra & Henseler, 2015b). It has indeed been regarded as “the method of choice” (Dijkstra & Henseler, 2015b, p. 311) for such models. However, there has been little empirical investigation into the performance of PLSc. In addition, the imposition of the basic design assumption can be restrictive in practice, excluding multidimensional measurement models that have been well researched (Asparouhov & Muthén, 2009). Conversely, IGSCA is a single-step approach that estimates all parameters simultaneously without recourse to the basic design assumption.

Table 1 provides a summary of the expected behaviors of the existing SEM approaches and IGSCA in estimating models with both factors and components. Nevertheless, no studies have been conducted to empirically assess the performance of all these approaches under such models. Thus, in the following section, we conduct a simulation study to evaluate how all the

approaches perform in recovering parameters in a model involving two components and a factor simultaneously.

=== Insert Table 1 about here ===

Simulation Study 1

In this study, we specified a data generating model that was composed of two exogenous components and one endogenous factor. Each factor or component was linked to three indicators. Figure 1 displays the specified model along with its prescribed loadings and path coefficients, where circles and hexagons are used to signify factors and components, respectively (e.g., Grace & Bollen, 2008). The loadings for the composite indicators z_1 to z_6 are component loadings, whereas those for the effect indicators z_7 to z_9 are factor loadings. The two path coefficients b_1 and b_2 denote the effects of the exogenous components on the endogenous factor. Besides their loadings, the composite indicators were additionally assigned weights to produce components, which are omitted to make the figure concise. The weights per component were fixed as .41, .37, and .39. The correlation between the two exogenous components (σ) was .3.

=== Insert Figure 1 about here ===

We considered four levels of sample size (N = 100, 200, 500, and 1000) and generated 1000 random samples from a multivariate normal distribution with zero means and the covariance matrix implied by the parameters per sample size. We provide a detailed description of our data generation procedure in Appendix 2. We applied the five existing SEM approaches (CSA, $GSCA_M$, PLSPM, GSCA, and PLSc) and IGSCA to each sample. We used the R package lavaan (version 0.5-16) (Rosseel, 2012) to apply CSA and wrote a MATLAB code for the other

approaches.[1] We fixed the first loading per factor/component to one to avoid scale indeterminacy in CSA, and fixed the sign of each factor/component to be positively correlated with the indicator with the largest loading to avoid sign indeterminacy in the other approaches (e.g., Henseler, Hubona, & Ray, 2016; Tenenhaus, Esposito Vinzi, Chatelin, & Lauro, 2005).

As parameter recovery measures, we calculated finite-sample properties, such as relative bias expressed as a percentage, standard deviation, and root mean square error, of the parameter estimates obtained from each approach. To conserve space, we focus here on reporting the relative biases of the parameters estimated from the six approaches for each experimental condition. All the properties of the parameter estimates per condition are fully provided in Supplementary Materials.

In the calculation of the parameter recovery measures, we removed any sample entailing non-convergence or convergence to improper solutions. Only CSA had such convergence problems when the sample size was small (N = 100), which is consistent with the literature (e.g., Anderson & Gerbing, 1984; Boomsma, 1982; 1985; Chen, Bollen, Paxton, Curran, & Kirby, 2001; Hwang et al., 2017). Specifically, when N = 100, nine samples was omitted for the convergence problems under CSA.

Table 2 presents the relative biases of the standardized loadings estimated from all the approaches over the different sample sizes. We regarded a relative bias greater than 10% in absolute value as indicative of an unacceptable degree of bias (e.g., Bollen, Kirby, Curran, Paxton, & Chen, 2007; Lei & Wu, 2012). CSA and $GSCA_M$ resulted in unbiased estimates of the factor loadings, whereas yielded negatively biased estimates of a few component loadings across all the sample sizes. As we have noted in the previous section, this is consistent with the

[1] The MATLAB code is available from the first author upon request.

literature (Rhemtulla et al., 2020; Sarstedt et al., 2016). On the other hand, GSCA and PLSPM provided positively biased estimates of the factor loadings regardless of the sample size. The inflated estimation of factor loadings by GSCA and PLSPM is also consistent with the literature (e.g., Dijkstra, 2010; Hwang et al., 2010; Sarstedt et al., 2016; Velicer & Jackson, 1990). As expected as well, both GSCA and PLSPM resulted in unbiased estimates of the component loadings across the same conditions. On the contrary, PLSc and IGSCA provided unbiased estimates of both factor and component loadings across all the sample sizes.

Most importantly, we now turn to examine the relative bias of the standardized path coefficients estimated from the six approaches (Table 3), which has not been examined in prior research. CSA and $GSCA_M$ yielded positively biased estimates of the path coefficients relating two components to a factor, whereas GSCA and PLSPM provided negatively biased estimates of them, regardless of the sample size. On the other hand, PLSc and IGSCA provided unbiased estimates of both path coefficients in all the sample sizes. This empirically supports our derivation regarding the behaviors of such path coefficients in Appendix 1.

=== Insert Tables 2 and 3 about here ===

To summarize, this is the first study to empirically evaluate the performance of the existing SEM approaches and IGSCA in estimating a model with both factors and components. Factor-based SEM approaches, such as CSA and $GSCA_M$, provided unbiased estimates of factor loadings, yet underestimated several component loadings and overestimated both path coefficients. Component-based SEM approaches, including PLSPM and GSCA, resulted in unbiased estimates of all component loadings, whereas underestimated the factor loadings and overestimated the path coefficients. Conversely, PLSc and IGSCA were the only approaches that provided unbiased estimates of all the loadings and path coefficients.

Although the present study may be perceived as limited in scope as it considered only two experimental factors (SEM approach and sample size) under a relatively simple data generating model, it provides empirical support for our theoretical expectations. Furthermore, only PLSc and IGSCA appear appropriate for estimating models with both factors and components as other approaches always resulted in biased estimates of several parameters, even under such a simple model. In the next section, therefore, we conduct another simulation study to examine the relative performance of PLSc and IGSCA while at the same time greatly expanding upon the generalizability of our simulation results.

Simulation Study 2

In the second study, we considered a more complex data generating model than the one used in the previous study. In particular, we used a model, where three factors and three components had effects on one component and one factor, as displayed in Figure 2. We manipulated four experimental factors for PLSc and IGSCA: (1) measurement model complexity, (2) the degree of the correlations of exogenous factors and components, (3) sample size, and (4) model specification. These experimental factors are frequently encountered in SEM simulations (Bandalos & Gagné, 2012; Paxton, Curran, Bollen, Kirby, & Chen, 2001). We did not additionally manipulate data distribution (e.g., normal vs. non-normal) as an experimental factor because both PLSc and IGSCA do not require distributional assumptions.

=== Insert Figure 2 about here ===

Three levels of measurement model complexity were determined by varying the number of indicators per factor/component (N_{ind} = 3, 5, and 7). The loading parameters per factor/component were fixed as follows: when N_{ind} = 3, c = .6, .7, and .8; when N_{ind} = 5, c

= .6, .6, .7, .8, and .8; when N_{ind} = 7, c = .6, .6, .7, .7, .7, .8, and .8. All composite indicators were also assigned weights to yield components, which are not displayed in Figure 2 to make the figure succinct.

We contemplated five different model specifications – correct, under-parameterized I, over-parameterized I, under-parameterized II, and over-parameterized II. The correct specification indicates that the fitted model was equivalent to the data generating model. The under-parameterized model I represents a misspecification of the measurement model, where the fitted model was more restricted than the data generating model, omitting two indicators for factors in the data generating model. The over-parameterized model I stands for another misspecification of the measurement model, where the fitted model incorrectly included additional component cross loadings, although the data generating model had no such cross loadings. The under-parameterized model II denotes a misspecification of the structural model, where the fitted model incorrectly excluded two path coefficients from the data generating model. The over-parameterized model II indicates a different misspecification of the structural model, in which the fitted model incorrectly contained two additional path coefficients despite that no such additional path coefficients existed in the data generating model. In Figure 2, the incorrectly omitted loadings and path coefficients in the under-parameterized models I and II are labeled A and B, respectively, whereas the additional component cross loadings and path coefficients in the over-parameterized models I and II are denoted by dashed arrows. In the figure, therefore, the data generating model is the one without the dashed arrows. The under-parameterized models I and II are the data generating model without the arrows labeled A and B, respectively. The over-parameterized models I and II are the data generating model with the dashed component cross loadings and dashed path coefficients added, respectively. PLSc can be

applied to the over-parameterized model I because the basic design should hold for the reflective measurement model with factors. In this specification, only components involved cross loadings. As stated earlier, PLSc does not need to correct component loadings or component scores (Dijkstra & Henseler, 2015b).

We also varied sample size (N = 100, 200, 500, and 1000) and the correlation of each pair of exogenous factors and components (σ = .1, .3, and .5). The correlations of some pairs of exogenous factors and components (i.e., γ_1 and γ_3, γ_1 and γ_5, γ_2 and γ_3, γ_2 and γ_6, γ_3 and γ_5, γ_4 and γ_5, and γ_4 and γ_6) were fixed to zero in order to keep the implied covariance matrix of indicators to be positive definite even when the correlation level σ was high. In sum, our simulation design consisted of a total of 180 data generating conditions (3 measurement model complexities × 5 model specifications × 4 sample sizes × 3 exogenous component/factor correlations). For each experimental condition, we generated 1000 random samples based on the data generation procedure proposed by Cho and Choi (2019).

Lastly, we generated all path coefficients in the model randomly from the set of {-.7, -.6, -.5, -.4, -.3, -.2, -.1, .1, .2, .3, .4, .5, .6, .7} for each random sample. This can alleviate the potential problem that our results are dependent on specific choices of path coefficient parameter values, thereby further increasing the generalizability of the results. We also had the signs of the exogenous factor/component correlations changed randomly per sample. Note that we retained the same loadings and exogenous factor/component correlations to maintain the validity of the measurement model.

We applied PLSc and IGSCA to each random sample per condition. In this study, we could not calculate the same recovery measures for their estimates as in the previous stimulation study (e.g., relative bias) because path coefficients were randomly chosen for each sample.

Instead, we computed the mean absolute error (MAE) of each set of loading and path coefficient estimates per sample as follows.

$$\text{MAE} = \frac{1}{I}\sum_{i=1}^{I}\left|\theta_i - \hat{\theta}_i\right|, \tag{10}$$

where I is the number of either loadings or path coefficients, and θ_i and $\hat{\theta}_i$ are the ith loading or path coefficient and its estimate obtained from each approach, respectively. The smaller the MAE, the closer the estimates are to parameters on average.

In the calculation of the MAE values per condition, we removed any sample entailing non-convergence or convergence to improper solutions. We found that only PLSc had cases with improper solutions, which was also reported in previous studies (e.g., Hwang et al., 2017; Rönkkö, McIntosh, & Aguirre-Urreta, 2016; Schamberger, Schuberth, Henseler & Dijkstra, 2020). This problem occurred in all conditions yet tended to occur more frequently when sample size was small (Table 4).

To conserve space, we focus here on reporting the average MAE values of the parameters estimated from PLSc and IGSCA. We provided the individual MAE values of the estimates from the approaches under all the conditions in Supplementary Materials. Figure 3 displays the average MAE values of the estimates from PLSc and IGSCA for each of the three experimental factors (the number of indicators, exogenous factor/component correlation, and sample size) under correct model specification. As shown in Figure 3, on average, the MAE values of the IGSCA estimates of loadings were consistently smaller than those of the PLSc estimates, regardless of the simulation conditions. The MAE values of both estimators approached zero when sample size increased, whereas the other conditions did not seem to influence their performance in recovering loading parameters. On the other hand, the MAE values of both

IGSCA and PLSc estimates of path coefficients were quite similar across the conditions. In general, for both approaches, the exogenous factor/component correlation had an adverse influence on the recovery of the path coefficients (i.e., the larger this correlation, the larger MAE), whereas the sample size had a positive impact (i.e., the larger the sample size, the smaller MAE). Conversely, the number of indicators had little influence on their recovery of path coefficients.

The average MAE values of the PLSc and IGSCA estimates under the under-parameterized model I were quite comparable to those obtained under the correct model in Figure 3. To further conserve space, therefore, these results were relegated to Supplementary Materials. As stated earlier, in the under-parameterized model I, an effect indicator was incorrectly removed from each of two factors. Removing the effect indicators had little impact on both approaches' recovery of remaining parameters, leading them to perform similarly to the case of correct model specification. This is consistent with that adding or omitting an effect indicator for a common factor does not change loading estimates of the other indicators for the factor (e.g., Widaman, 2018).

Figure 4 shows the average MAE values of the PLSc and IGSCA estimates under the over-parameterized model I. In all conditions, the PLSc and IGSCA estimates of loadings and path coefficients generally showed similar MAE patterns to those under correct model specification. Nonetheless, PLSc performed more poorly in recovering both loadings and path coefficients than under correct model specification. Consequently, the MAE values of the loadings and path coefficients estimated from PLSc were larger than their counterparts from IGSCA in most conditions. In the over-parameterized model I, a composite indicator for each of two components was incorrectly linked to another component, giving rise to additional cross

component loadings. PLSPM, on which PLSc relies solely for estimating component loadings, tends to yield more biased loading estimates than GSCA in models with cross loadings (Hwang et al., 2010; Hwang & Takane, 2014, Chapter 2), which can in turn affect the recovery of path coefficients negatively. This might lead PLSc to perform more poorly than IGSCA under the over-parameterized model I.

Figures 5 and 6 exhibit the average MAE values of the PLSc and IGSCA estimates under the under-parameterized model II and the over-parameterized model II, both of which involved misspecified structural models. Under both misspecifications, again, the PLSc and IGSCA estimates of loadings and path coefficients showed similar MAE patterns to those obtained under correct model specification. That is, in general, IGSCA recovered loadings better than PLSc, whereas they recovered path coefficients similarly. Furthermore, the average MAE values of the IGSCA estimates of both loadings and path coefficients largely remained similar to those under correct model speciation, suggesting that the two misspecifications of the structural model had no substantial influence on the performance of IGSCA. Conversely, PLSc recovered loadings better under the over-parameterized model II than under the under-parameterized model II. This performance of PLSc seems to be related to that PLSPM tends to better recover parameters in more complex structural models than those in simpler structural models (Henseler et al., 2014). As PLSc utilizes PLSPM for the first stage of parameter estimation, the positive effect of structural model complexity on PLSPM was likely to carry over to PLSc under the over-parametrized model II, which had a more complex structural model with additional path coefficients than the under-parameterized model II.

=== Insert Table 4 and Figures 3, 4, 5, and 6 about here ===

In sum, we investigated the relative performance of PLSc and IGSCA in recovering parameters in a complex model involving multiple factors and components, taking into account a variety of experimental conditions. To enhance the generalizability of our results, we did not fix path coefficients to specific values in advance and instead had them randomly chosen from a wide range of candidate values per sample. In general, IGSCA showed better performance in parameter recovery than PLSc. IGSCA always recovered loadings better than PLSc regardless of the experimental factors, while the two approaches recovered path coefficients similarly in most conditions except for the case of a misspecified model with cross component loadings, where IGSCA largely recovered path coefficients better. In addition, PLSc suffered from the occurrence of improper solutions in all conditions, whereas IGSCA did not encounter such a problem.

Empirical Application: The Gene and Depression Data

In this example, we investigate the influence of genes on the severity of depression. The example had a total of 231 Korean participants, which included 137 (59.3%) healthy volunteers recruited from community advertisements and 94 PTSD patients recruited from notices on the bulletin board at a university hospital in a northern suburban area of Seoul, Korea. The PTSD patients were diagnosed based on the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) by a psychiatrist, and healthy participants were also evaluated using the DSM-5 by a psychiatrist. There were 75 (32.5%) men and 156 women with a mean age of 46.10 years (SD = 13.49). Each participant signed a written form of informed consent approved by the Institutional Review Board at the hospital before participating in the study (IRB no. 2015-07-025), and all measurements and experiments were executed in accordance with the guidelines and regulations of the board.

To measure the severity of depression symptoms among the participants, the Korean translation of the Hospital Anxiety Depression Scale (HADS; Oh, Min, & Park, 1999) was administered. The HADS is a self-report rating scale that has two sub-scales, each of which consists of seven items for anxiety (HADS-A) and depression (HADS-D). We used the seven items for depression in the HADS-D. To control for the effect of alcohol-related problems as a covariate, the Alcohol Use Disorders Identification Test (AUDIT) was used to assess alcohol consumption, drinking behaviors, and alcohol-related problems. The AUDIT is a 10-item screening tool developed by the World Health Organization, and well-validated in Korea (B. O. Lee, Lee, Lee, Choi, & Namkoong, 2000). The AUDIT is assessed with a 5-point Likert scale ranging from 0 ("never") to 4 ("4 or more times a week").

All the participants had their blood sampled to extract DNA using NanoDrop® ND-1000 UV-Vis Spectrophotometer. Then, genomic DNA were diluted to 10 ng/$\mu\ell$ concentration at 96 well PCR plates. TaqMan SNP Genotyping Assays were obtained from Applied Biosystems. The probes were labeled with FAM or VIC dye at the 5' end and a minor-groove binder and non-fluorescent quencher at the 3' end. 2 µL of DNA was added to each 5 µL PCR reaction at 384 well reaction plates. SNP genotyping reactions were performed on ABI PRISM 7900HT Real-time PCR system. After the PCR amplification, allelic discrimination is performed at the same machines (ABI 7900HT). The allelic discrimination is an end point plate read. The SDS v2.4 software calculates the fluorescence measurements made during the plate read and plots Rn values based on the signals from each well. A total of 18 SNPs from 9 different DNAs were obtained for the study. For all SNPs, the wild, hetero, and mutant genotypes were coded as 1, 2, and 3, respectively. We selected nine genes that may be related to depression, including SLC6A4 (Holmes, Bogdan, & Pizzagalli, 2010), FKBP5 (Zobel et al., 2010), ADCYAP1R1 (Lowe et al.,

2015), BDNF (Sen et al., 2003), COMT (Åberg, Fandiño-Losada, Sjöholm, Forsell, & Lavebratt, 2011), HTR3A (Gatt et al., 2010), DRD2 (Vaske, Makarios, Boisvert, Beaver, & Wright, 2009), NR3C1 (Gałecka et al., 2013), and OXTR (McQuaid, McInnis, Stead, Matheson, & Anisman, 2013). Table 6 exhibits all the genes and their associated SNPs.

Depression symptoms are known to be affected by genetic polymorphism (Peper, Brouwer, Boomsma, Kahn, & Hulshoff Pol, 2007) and also influenced by other extraneous variables, such as gender (Piccinelli & Wilkinson, 2000; Sowell et al., 2007), age (Kessler et al., 2010; Salat et al., 2004), and alcohol-related problems (Boden & Fergusson, 2011; Durazzo et al., 2011). In addition, some studies have observed potential interaction effects of gender and age (Patten et al., 2016) and of gender and alcohol use (e.g., Danzo, Connell, & Stormshak, 2017; Edwards et al., 2014; Fleming, Mason, Mazza, Abbott, & Catalano, 2008; Miettunen et al., 2014; Nolen-Hoeksema, 2004) on depression symptoms. Accordingly, we hypothesized that the nine genes had direct effects on depression severity, while controlling for the direct and interaction effects of gender, age, and AUDIT as covariates on depression severity. Table 5 provides the correlation matrix of all the SNPs, depression items, and covariates, along with their means and standard deviations. There were no extreme observations (e.g., their z scores were greater than |3.3|) or skewed variables. As discussed earlier, we assumed that each of the nine genes was a component of SNPs, whose number per gene ranged from one to nine, whereas depression was a factor that underlay the seven items in the HADS-D. Figure 7 displays the hypothesized structural model.

== Insert Table 5 and Figure 7 about here ==

We applied IGSCA to fit the hypothesized model to the data. We used 1000 bootstrap samples to estimate the standard errors and 95% confidence intervals of the parameter estimates.

Tables 6 and 7 present the standardized loadings and path coefficients, respectively, estimated from IGSCA. All loading estimates were statistically significant and large in magnitude. We found that the pituitary adenylate cyclase activating polypeptide 1 receptor type I (ADCYAP1R1) gene had a statistically significant and positive effect on depression (b_3 = .18, SE = .07, 95% CI = [.03, .30]), suggesting that people with the mutant allele in ADCYAP1R1 may experience a higher level of depression severity. This finding is consistent with previous studies that the ADCYAP1R1 gene may be implicated in stress response processes (Hashimoto, Shintani, & Baba, 2006), trauma-related psychopathology (Roman et al., 2014), and depression (Aragam, Wang, & Pan, 2011), since the neuropeptide pituitary adenylate cyclase-activating polypeptide (PACAP) regulates activation of the hypothalamic-pituitary-adrenal axis after stressful experience (Lehmann, Mustafa, Eiden, Herkenham, & Eiden, 2013). The R^2 value of depression was .13.

== Insert Tables 6 and 7 about here ==

Discussion

We proposed a novel statistical approach to SEM with both factors and components. The proposed method, IGSCA, combines GCSA and $GSCA_M$ into a single framework. IGSCA estimates all parameters simultaneously by minimizing a single least squares criterion without recourse to distributional assumptions.

We conducted two simulation studies to evaluate the performance of IGSCA relative to existing approaches in the domains of factor- and component-based SEM. The first study evaluated the performance of all of the approaches. In line with our derivations, only PLSc and IGSCA are appropriate candidates for estimating models with both factors and components

because the other approaches always yielded biased estimates of some parameters – even under a simple model. The second study then focused on examining the relative performance of the two competing approaches (PLSc and IGSCA), considering a number of experimental conditions. The second study shows that IGSCA generally outperformed PLSc in that it recovered loadings better while recovering path coefficients equally or better. Also, IGSCA was less vulnerable to model misspecification and the occurrence of improper solutions.

We also applied IGSCA to examine the effects of several candidate genes on depression. We utilized biological pathway information to specify genes as components of SNPs, which are known to be related to depression symptoms. Moreover, we considered multiple genes simultaneously, taking into account potential correlations among the genes. This can also help avoid the multiple testing problem in genetic studies (Lee et al., 2016).[2] Several studies have also considered multiple genes at the same time, each of which was regarded as a component of SNPs (e.g., Romdhani et al., 2015). However, these previous studies did not consider factors at all because they relied on component-based statistical approaches. Conversely, by applying IGSCA, we modeled genes as components and depression as a factor, which, as outlined in our introduction, is more theoretically plausible. Based on this model, we found evidence that a particular gene (ADCYAP1R1) appears to have an effect on depression, which is consistent with previous research.

The development of IGSCA has both technical and empirical implications. IGSCA enables us to efficiently estimate structural equation models that are composed of both components and factors. It will make a technical contribution to bridging the two SEM domains,

[2] Although some researchers had suggested more stringent Type I error control for SEM (e.g., Cribbie, 2000), we did not employ such additional Type I error control as it still seems uncommon to do so. If desired, researchers may conduct adjusted tests that are based on the conventional Bonferroni correction or less conservative procedures such as the Šidák or Benjamini-Hochberg correction.

substantially expanding the scope and applicability of SEM. IGSCA can serve as a flexible tool for researchers to examine the relationships involving factors and components at the same time. The availability of such an analytic tool can allow researchers to consider a greater variety of variables emerged from different disciplines. For example, the integration of gene- and brain-level variables into studies in clinical child or pediatric psychology via IGSCA may provide insights for better comprehension of children's healthy development, as well as identifying which neurobiological pathways are involved in altered neurodevelopment and behavior. We believe that the demand for accommodating both factors and components continues to rise in psychology and other sciences that are becoming increasingly interdisciplinary.

Despite its technical and empirical implications, IGSCA has limitations. Notably, IGSCA does not offer a conventional test statistic for examining overall model-data consistency or comparing alternative models, such as the chi-square test of fit in CSA, partly because it does not require distributional assumptions. IGSCA can still allow examining the statistical significance of individual parameter estimates, helping to investigate consistency between data and a series of hypotheses that constitutes the model. This local fit evaluation may contribute to confirmation, rejection, or revision of the entire model. Nevertheless, it would be necessary to develop more formal procedures for evaluating overall model-data consistency and alternative models. Additionally, it would be worthwhile to develop a test statistic or index for evaluating model predictability, which has been highly recommended in SEM (Cho, Jung, & Hwang, 2019; Hair, Pieper, & Ringle, 2012; MacCallum & Austin, 2000; MacCallum, Roznowski, & Necowitz, 1992; Ringle, Sarstedt, & Straub, 2012).

Despite the promising results of the current paper, IGSCA is at an early stage of development relative to other existing SEM approaches. Additional work is needed to refine and

extend IGSCA to improve its data-analytic capability. For example, IGSCA in this paper has focused mainly on unidimensional measurement models where a single factor or component is associated with a set of indicators. Although IGSCA in theory can accommodate more complex models and we have considered a multidimensional model with cross loadings in our simulation study, more work is needed to develop and test IGSCA with various complex measurement models, including multitrait-multimethod (Campbell & Fiske, 1959), latent growth curve (Duncan, Duncan, & Strycker, 2006; Meredith & Tisak, 1990), bifactor (Holzinger & Swineford, 1937), and random intercept (Maydeu-Olivares & Coffman, 2006) models.

IGSCA is geared only for an aggregate sample analysis, which estimates parameters by pooling the data across observations under the assumption that all observations come from a single homogenous population. In some situations, however, it may be more reasonable to assume that observations are drawn from (unknown) heterogeneous subgroups in the population that have different parameters or relations among factors and/or components. For example, two different trends of antisocial behavior, such as life-course persistent and adolescent-limited, have been discussed in the literature (Moffitt, 1993). In the situations where such cluster-level heterogeneity is present, an aggregate sample analysis tends to yield biased estimates (e.g., DeSarbo & Cron, 1988; Jedidi, Jagpal, & DeSarbo, 1997; Muthén, 1989). Thus, future work may extend IGSCA to account for cluster-level heterogeneity through combining it with cluster analysis or latent class models.

IGSCA is currently based on the assumption that a set of indicators is always linearly related to a factor/component, as is also the case with most applications of other SEM approaches. However, in some cases, such an indicator-factor/component relationship may not be strictly linear. For instance, a factor/component may have a floor or ceiling effect on its

indicators (e.g., Bauer, 2005; Mooijaart & Bentler, 1986). Thus, future work may expand IGSCA to capture potentially nonlinear relationships between indicators and factors/components.

IGSCA has not yet been developed to deal with the issues of multicollinearity and variable selection. Multicollinearity makes it difficult to interpret the unique influences of predictor variables on outcomes, possibly leading to inferential error. Selection of an optimal set of variables is of use in facilitating the interpretation of a given model, eliminating uninformative variables (e.g., Baumann, Albert, & von Korff, 2002). To address these issues, we may consider combining IGSCA with the elastic net (Zou & Hastie, 2005) in a single framework, which includes the ridge (Hoerl & Kennard, 1970) and lasso (Tibshirani, 1996), following the prior development of ridge- and lasso-regularized GSCA (Hwang, 2009; Hwang & Takane, 2014, Chapter 9).

IGSCA cannot yet account for the dynamic nature of temporally (serially) correlated data, where the term dynamic refers to a process in which a state of a variable at a particular time may be influenced by a state of the same or other variables at previous times. In brain connectivity studies (Friston, 1994), such time series data are ubiquitous. For example, functional magnetic resonance imaging records changes in blood oxygenation over scans (time points), called blood-oxygen level dependent signals, while a subject is presented with stimuli or asked to perform a task. The development of dynamic GSCA (Jung et al., 2012) can be adapted to integrate multivariate autoregressive models into IGSCA to explicitly take into account the dynamic relationships in time series data.

IGSCA may not be robust to outliers because it estimates parameters via least squares. It would be desirable to develop robust estimation procedures for IGSCA, where observations are iteratively re-weighted based on their residuals. We note that such approaches have been

developed for GSCA (Hwang & Takane, 2014, Chapter 3) and may be adapted for use with IGSCA. To accommodate missing data, it may be desirable to develop methods that are more efficient than listwise deletion or mean substitution. For example, we may consider estimating missing observations iteratively based on the data and model specification as is already available for GSCA (e.g., Hwang & Takane, 2014, Chapter 3).

As with many SEM approaches, IGSCA requires researchers to formulate the nature of each dimension as a factor or component in advance. This formulation should be based on prior substantive theory or knowledge. Nevertheless, in practice, this decision may sometimes prove difficult if a strong theoretical foundation is lacking. In this case, we may consider applying a confirmatory tetrad analysis (Bollen & Ting, 1993, 2000), which tests the statistical significance of (non-redundant) model-implied vanishing tetrads per factor/component, in order to statistically evaluate whether having factors or components is more consistent with the data.

In closing, we believe that IGSCA has great potential to unify disparate research areas in psychology and other social and behavioral sciences where either factors or components dominate the research terrain. Although we have listed numerous limitations of IGSCA, many of these are common to traditional factor-based and component-based approaches, and these problems may be addressed by borrowing innovations that are already available for the very closely related GSCA approach (e.g., outliers, missing data, multicollinearity, variable selection, dynamic data, etc.). We hope to continue to make IGSCA more widely applicable and useful to both methodologists and applied researchers by addressing these issues in future research, and by also developing a software program for IGSCA in an accessible format, such as an R package. We expect that this will allow IGSCA to be useful for a wide range of real-world problems and allow more thorough investigations of its empirical performance.

Appendix 1: The impact of factor/component misspecification on the estimation of a structural correlation when the true model contains both factor and component

We consider a simple case, where the true model has a factor and a component that is a unit-weighted composite, as in Rhemtulla et al. (2020). We show how the estimation of the structural correlation between the component and factor in the model is affected (1) when the component is incorrectly specified as a factor (i.e., both are treated as factors), and (2) when the factor is erroneously specified as a component (i.e., both are components). Methodologically, case (1) corresponds to when a factor-based SEM approach (CSA, PLSc, or $GSCA_M$) is applied to the true model, whereas case (2) corresponds to when a component-based SEM approach (PLSPM or GSCA) is applied to the true model.

Let $\mathbf{x} = [x_1,\ldots, x_p]$ denote a p-dimensional vector of composite indicators. Let $\mathbf{y} = [y_1,\ldots, y_q]$ denote a q-dimensional vector of effect indicators. Let γ_1 denote the unit-weighted component of $\mathbf{x}$, i.e., $\gamma_1 = \mathbf{x}\mathbf{1}_p$, where $\mathbf{1}_p$ is a p-dimensional vector of ones. Let γ_2 denote the factor that underlies $\mathbf{y}$, based on the reflective model $y_j = \lambda_j\gamma_2 + u_j$, where λ_j and u_j are the loading and unique factor for y_j, respectively ($j = 1,\ldots,q$). We assume that $\text{var}(x_p) = \text{var}(y_q) = 1$, $\text{cov}(x_p, x_{p'}) = \text{cov}(y_q, y_{q'}) = m$, and $\text{cov}(x_p, y_q) = h$ ($p \neq p'$; $q \neq q'$). For simplicity, we further assume that all composite indicators are measured without error. The reflective model will perfectly reproduce the covariance matrix of $\mathbf{y}$, when $\lambda_j = 1$, $\text{var}(\gamma_2) = m$, and $\text{var}(u_j) = 1 - m$.

Then, the correlation between γ_1 and γ_2 is given by

$$\beta = \text{corr}(\gamma_1, \gamma_2) = \frac{\text{cov}(\mathbf{x}\mathbf{1}_p, \gamma_2)}{\sqrt{\text{var}(\mathbf{x}\mathbf{1}_p)}\sqrt{\text{var}(\gamma_2)}} = \frac{ph}{\sqrt{p + p(p-1)m}\sqrt{m}}. \quad \text{(A1.1)}$$

When γ_1 is incorrectly specified as another factor, say γ_3, based on the reflective model $x_s = \lambda_s\gamma_3 + u_s$ ($s = 1,\ldots, p$). This reflective model again perfectly reproduces the covariance matrix of $\mathbf{x}$ when $\lambda_s = 1$, $\text{var}(\gamma_3) = m$, and $\text{var}(u_s) = 1 - m$. Then, the correlation between γ_3 and γ_2 is given by

$$\beta_f = \text{corr}(\gamma_3, \gamma_2) = \frac{\text{cov}(\gamma_3, \gamma_2)}{\sqrt{\text{var}(\gamma_3)}\sqrt{\text{var}(\gamma_2)}} = \frac{h}{m}. \quad \text{(A1.2)}$$

(also see Bollen, 1989, p. 327).

When γ_2 is incorrectly specified as another component, denoted by $\gamma_4 = \mathbf{y1}_q$, where $\mathbf{1}_q$ is a q-dimensional vector of ones, then the correlation between γ_1 and γ_4 is given by

$$\beta_c = \text{corr}(\gamma_1, \gamma_4) = \frac{\text{cov}(\mathbf{x1}_p, \mathbf{y1}_q)}{\sqrt{\text{var}(\mathbf{x1}_p)}\sqrt{\text{var}(\mathbf{y1}_q)}} = \frac{pqh}{\sqrt{p + p(p-1)m}\sqrt{q + q(q-1)m}}. \quad \text{(A1.3)}$$

By dividing β_f by β, we have

$$\frac{\beta_f}{\beta} = \frac{\sqrt{p + p(p-1)m}\sqrt{m}}{pm}. \quad \text{(A1.4)}$$

This ratio will be greater than 1, unless $m = 1$. Thus, when the component in the true model is misspecified as a factor (i.e., case (1)), the structural correlation β tends to be overestimated. On the other hand, the ratio of β_c to β is given by

$$\frac{\beta_c}{\beta} = \frac{q\sqrt{m}}{\sqrt{q + q(q-1)m}}. \quad \text{(A1.5)}$$

This ratio will be smaller than 1 unless $m = 1$, indicating that when the factor in the true model is misspecified as a component (i.e., case (2)), the structural correlation β tends to be underestimated.

Appendix 2: The data generation procedure for the first simulation study

In our first simulation study, a structural equation model was considered that contained two exogenous components, one endogenous factor, and three indicators per component or factor. All the indicators, components, and factor were assumed to be standardized. Let $\mathbf{z}_X = \begin{bmatrix} \mathbf{z}_{X1} \\ \mathbf{z}_{X2} \end{bmatrix}$ denote a vector of composite indicators for the two components. Let $\mathbf{z}_Y$ denote a vector of effect indicators for the endogenous factor. Let $\boldsymbol{\gamma}_X = \begin{bmatrix} \gamma_{X1} \\ \gamma_{X2} \end{bmatrix}$ denote a vector of the components, and γ_Y denote the endogenous factor. Let $\boldsymbol{\Sigma}z_X = \begin{bmatrix} \boldsymbol{\Sigma}z_{X1} & \boldsymbol{\Sigma}z_{X1}z_{X2}' \\ \boldsymbol{\Sigma}z_{X1}z_{X2} & \boldsymbol{\Sigma}z_{X2} \end{bmatrix}$ denote the covariance matrix of the composite indicators, where $\boldsymbol{\Sigma}z_{X1}$ and $\boldsymbol{\Sigma}z_{X2}$ the covariance matrices of $\mathbf{z}_{X1}$ and $\mathbf{z}_{X2}$, respectively, and $\boldsymbol{\Sigma}z_{X1}z_{X2}$ is the covariance matrix between $\mathbf{z}_{X1}$ and $\mathbf{z}_{X2}$. Let $\boldsymbol{\Sigma}_X = \begin{bmatrix} 1 & \sigma \\ \sigma & 1 \end{bmatrix}$ denote the covariance matrix of the two exogenous components. Let $\mathbf{W}_X = \begin{bmatrix} \mathbf{W}_{X1} & 0 \\ 0 & \mathbf{W}_{X2} \end{bmatrix}$ denote a matrix of weights assigned to the composite indicators $\mathbf{z}_{X1}$ and $\mathbf{z}_{X2}$. Let $\mathbf{C}_X = \begin{bmatrix} \mathbf{c}_{X1} & 0 \\ 0 & \mathbf{c}_{X2} \end{bmatrix}$ denote a matrix of component loadings for $\mathbf{z}_{X1}$ and $\mathbf{z}_{X2}$, and $\mathbf{C}_Y$ denote a vector of factor loadings for $\mathbf{z}_Y$. Let $\boldsymbol{\varepsilon}_X = \begin{bmatrix} \boldsymbol{\varepsilon}_{X1} \\ \boldsymbol{\varepsilon}_{X2} \end{bmatrix}$ denote a vector of the residuals for $\mathbf{z}_{X1}$ and $\mathbf{z}_{X2,}$ and $\boldsymbol{\varepsilon}_Y$ denote a vector of the residuals for $\mathbf{z}_Y$. Let $\boldsymbol{\Sigma}\varepsilon_X = \begin{bmatrix} \boldsymbol{\Sigma}\varepsilon_{X1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}\varepsilon_{X2} \end{bmatrix}$ and $\boldsymbol{\Sigma}\boldsymbol{\varepsilon}_Y = \begin{bmatrix} d_1 & 0 & 0 \\ 0 & d_2 & 0 \\ 0 & 0 & d_3 \end{bmatrix}$ denote the covariance matrices of $\boldsymbol{\varepsilon}_X$ and $\boldsymbol{\varepsilon}_Y$, respectively. Let $\mathbf{B} = [\beta_1, \beta_2]$ denote a vector of path coefficients. Let ζ denote the residual for γ_Y in the structural model.

The specified model can be expressed by the following equations.

$$\mathbf{z}_X = \mathbf{C}_X\boldsymbol{\gamma}_X + \boldsymbol{\varepsilon}_X \quad \text{(A2.1)}$$

$$\mathbf{z}_Y = \mathbf{C}_Y\boldsymbol{\gamma}_Y + \boldsymbol{\varepsilon}_Y, \quad \text{(A2.2)}$$

$$\boldsymbol{\gamma}_X = \mathbf{W}_X\mathbf{z}_X \quad \text{(A2.3)}$$

$$\gamma_Y = \mathbf{B}\boldsymbol{\gamma}_x + \zeta, \quad \text{(A2.4)}$$

where $\text{cov}(\boldsymbol{\gamma}_X,\boldsymbol{\varepsilon}_X) = \mathbf{0}$, and $\text{cov}(\gamma_Y,\boldsymbol{\varepsilon}_Y) = \mathbf{0}$, and $\text{cov}(\gamma_Y,\zeta) = 0$.

For the simulation study, the prescribed loadings of effect indicators for a common factor were used to derive the covariance matrix of the effect indicators, whereas the prescribed covariance matrix of composite indicators for a component was used to derive their weights and loadings. Specifically, we pre-determined the values of $\boldsymbol{\Sigma}z_{X1}$, $\boldsymbol{\Sigma}z_{X2}$, $\mathbf{C}_Y$, $\mathbf{B}$, and σ, as follows:

$$\boldsymbol{\Sigma}z_{X1} = \boldsymbol{\Sigma}z_{X2} = \begin{bmatrix} 1 & .6 & .7 \\ & 1 & .5 \\ & & 1 \end{bmatrix}, \mathbf{C}_Y = \begin{bmatrix} c_{y1} \\ c_{y2} \\ c_{y3} \end{bmatrix} = \begin{bmatrix} .6 \\ .7 \\ .8 \end{bmatrix}, \mathbf{B} = [-.7, .5], \text{ and } \sigma = .3, .5, \text{ or } .7.$$

Given the prescribed parameter values, $\mathbf{W}_X = \begin{bmatrix} \mathbf{w}_{X1} & 0 \\ 0 & \mathbf{w}_{X2} \end{bmatrix}$ was obtained by

$\mathbf{w}_{X1} = (\boldsymbol{\Sigma}z_{X1})^{-1/2}\mathbf{q}_{X1}$ and $\mathbf{w}_{X2} = (\boldsymbol{\Sigma}z_{X2})^{-1/2}\mathbf{q}_{X2}$, where $\mathbf{q}_{X1}$ and $\mathbf{q}_{X2}$ are the eigenvectors corresponding to the largest eigenvalues of $\boldsymbol{\Sigma}z_{X1}$ and $\boldsymbol{\Sigma}z_{X2}$, respectively (see Cho & Choi, 2020). Then,

$\mathbf{C}_X = \begin{bmatrix} \mathbf{C}_{X1} & 0 \\ 0 & \mathbf{C}_{X2} \end{bmatrix}$ was obtained by $\mathbf{c}_{X1} = \mathbf{w}_{X1}\boldsymbol{\Sigma}_{z_{X1}}$ and $\mathbf{c}_{X2} = \mathbf{w}_{X2}\boldsymbol{\Sigma}_{z_{X2}}$, and $\boldsymbol{\Sigma}\varepsilon_X = \begin{bmatrix} \boldsymbol{\Sigma}\varepsilon_{X1} & 0 \\ 0 & \boldsymbol{\Sigma}\varepsilon_{X2} \end{bmatrix}$

was by $\boldsymbol{\Sigma}\varepsilon_{X1} = \boldsymbol{\Sigma}z_X - \mathbf{c}_{X1}'\mathbf{c}_{X1}$ and $\boldsymbol{\Sigma}\varepsilon_{X2} = \boldsymbol{\Sigma}z_{X2} - \mathbf{c}_{X2}'\mathbf{c}_{X2}$. Subsequently,

$\boldsymbol{\Sigma}z_X = \begin{bmatrix} \boldsymbol{\Sigma}z_{X1} & \boldsymbol{\Sigma}z_{X1}z_{X2}' \\ \boldsymbol{\Sigma}z_{X1}z_{X2} & \boldsymbol{\Sigma}z_{X2} \end{bmatrix}$ was given by $\boldsymbol{\Sigma}z_X = \mathbf{C}_X'\boldsymbol{\Sigma}_X\mathbf{C}_X' + \boldsymbol{\Sigma}\varepsilon_X$. We generated $\mathbf{z}_X$ from a multivariate normal distribution with zero means and $\boldsymbol{\Sigma}_{Zx}$. We then derived $\boldsymbol{\gamma}_X$ from (A2.3) and

generated γ_Y from a normal distribution with zero mean and var($\boldsymbol{\varepsilon}_Y$), where var($\boldsymbol{\varepsilon}_Y$) =1 – $\mathbf{B}'\mathbf{B}$ – $2\sigma\beta_1\beta_2$. We also generated $\boldsymbol{\varepsilon}_Y$ from a multivariate normal distribution with zero mean and $\boldsymbol{\Sigma}\varepsilon_Y$, where $\boldsymbol{\Sigma}\varepsilon_Y = \begin{bmatrix} 1-c_{y1}^{\,2} & 0 & 0 \\ 0 & 1-c_{y2}^{\,2} & 0 \\ 0 & 0 & 1-c_{y3}^{\,2} \end{bmatrix}$. Finally, we generated $\mathbf{z}_Y$ from (A2.2). In this way, we could obtain $\mathbf{z} = \begin{bmatrix} \mathbf{z}_X \\ \mathbf{z}_Y \end{bmatrix}$.

References

Åberg, E., Fandiño-Losada, A., Sjöholm, L. K., Forsell, Y., & Lavebratt, C. (2011). The functional Val158Met polymorphism in catechol-O- methyltransferase (COMT) is associated with depression and motivation in men from a Swedish population-based study. *Journal of Affective Disorders*, *129*(1–3), 158–166. https://doi.org/10.1016/j.jad.2010.08.009

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*(2), 155–173. https://doi.org/10.1007/BF02294170

Aragam, N., Wang, K. S., & Pan, Y. (2011). Genome-wide association analysis of gender differences in major depressive disorder in the Netherlands NESDA and NTR population-based samples. *Journal of Affective Disorders*, *133*(3), 516–521. https://doi.org/10.1016/j.jad.2011.04.054

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Bandalos, D. L., & Gagné, P. (2012). Simulation methods in structural equation modeling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling.* (pp. 92–108). New York, NY: Guilford Press.

Bauer, D. J. (2005). The role of nonlinear factor-to-indicator relationships in tests of measurement equivalence. *Psychological Methods*, *10*(3), 305. https://doi.org/10.1037/1082-989X.10.3.305

Baumann, K., Albert, H., & von Korff, M. (2002). A systematic evaluation of the benefits and

hazards of variable selection in latent variable regression. Part I. Search algorithm, theory and simulations. *Journal of Chemometrics*, *16*(7), 339–350. https://doi.org/10.1002/cem.730

Birnbaum, R., & Weinberger, D. R. (2013). Functional neuroimaging and schizophrenia: A view towards effective connectivity modeling and polygenic risk. *Dialogues in Clinical Neuroscience*, *15*(3), 279–289. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3811100/

Boden, J. M., & Fergusson, D. M. (2011). Alcohol and depression. *Addiction*, *106*(5), 906–914. https://doi.org/10.1111/j.1360-0443.2010.03351.x

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley. https://doi.org/10.1002/9781118619179

Bollen, K. A., & Bauldry, S. (2011). Three Cs in measurement models: Causal indicators, composite indicators, and covariates. *Psychological Methods*, *16*(3), 265–284. https://doi.org/10.1037/a0024448

Bollen, K. A., & Diamantopoulos, A. (2017). In defense of causal-formative indicators: A minority report. *Psychological Methods*, *22*(3), 581–596. https://doi.org/10.1037/met0000056

Bollen, K. A., Kirby, J. B., Curran, P. J., Paxton, P. M., & Chen, F. (2007). Latent variable models under misspecification two-stage least squares (2SLS) and maximum likelihood (ML) estimators. *Sociological Methods and Research*, *36*(1), 48–86. https://doi.org/10.1177/0049124107301947

Bollen, K. A., & Ting, K. (1993). Confirmatory tetrad analysis. In P. Marsden (Ed.), *Sociological methodology 1993* (pp. 147–175). Washington, DC: American Sociological Association.

Bollen, K. A., & Ting, K. (2000). A tetrad test for causal indicators. *Psychological Methods*, *5*(1), 3–22. https://doi.org/10.1037/1082-989X.5.1.3

Bookheimer, S. Y., Strojwas, M. H., Cohen, M. S., Saunders, A. M., Pericak-Vance, M. A., Mazziotta, J. C., & Small, G. W. (2000). Patterns of brain activation in people at risk for Alzheimer's disease. *New England Journal of Medicine, 343*(7), 450–456. https://doi.org/10.1056/NEJM200008173430701

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: causality, structure, prediction* (Part 1, pp. 149–173). Amsterdam, Netherlands: North Holland.

Boomsma, Anne. (1985). Nonconvergence, improper solutions, and starting values in lisrel maximum likelihood estimation. *Psychometrika, 50*(2), 229–242. https://doi.org/10.1007/BF02294248

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*(2), 81–105. https://doi.org/10.1037/h0046016

Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research, 29*(4), 468–508. https://doi.org/10.1177/0049124101029004003

Cho, G., & Choi, J. Y. (2020). An empirical comparison of generalized structured component analysis and partial least squares path modeling under variance-based structural equation models. *Behaviormetrika, 47*(1), 243–272. https://doi.org/10.1007/s41237-019-00098-0

Cho, G., Jung, K., & Hwang, H. (2019). Out-of-bag prediction error : A cross validation index for generalized structured component analysis. *Multivariate Behavioral Research*, 1–9. https://doi.org/10.1080/00273171.2018.1540340

Cribbie, R. A. (2000). Evaluating the importance of individual parameters in structural equation modeling: The need for type I error control. *Personality and Individual Differences*, *29*(3), 567–577. https://doi.org/10.1016/S0191-8869(99)00219-6

Danzo, S., Connell, A. M., & Stormshak, E. A. (2017). Associations between alcohol-use and depression symptoms in adolescence: Examining gender differences and pathways over time. *Journal of Adolescence*, *56*, 64–74. https://doi.org/https://doi.org/10.1016/j.adolescence.2017.01.007

DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, *5*(2), 249–282. https://doi.org/10.1007/BF01897167

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, *38*(2), 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845

Dijkstra, T. K. (2010). Latent variables and indices: Herman Wold's basic design and partial least squares. In V. Esposito Vinzi, W. W. Chin, J. Henseler, & H. Wang (Eds.), *Handbook of Partial Least Squares: Concepts, Methods and Applications* (pp. 23–46). Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-32827-8_2

Dijkstra, T. K., & Henseler, J. (2015a). Consistent and asymptotically normal PLS estimators for linear structural equations. *Computational Statistics and Data Analysis*, *81*, 10–23. https://doi.org/10.1016/j.csda.2014.07.008

Dijkstra, T. K., & Henseler, J. (2015b). Consistent partial least squares path modeling. *MIS Quarterly, 39*(2), 297–316. https://doi.org/10.25300/misq/2015/39.2.02

Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, issues, and applications* (2nd ed.). Mahwah, NJ: Erlbaum.

Durazzo, T. C., Tosun, D., Buckley, S., Gazdzinski, S., Mon, A., Fryer, S. L., & Meyerhoff, D. J. (2011). Cortical thickness, surface area, and volume of the brain reward system in alcohol dependence: Relationships to relapse and extended abstinence. *Alcoholism: Clinical and Experimental Research, 35*(6), 1187–1200. https://doi.org/10.1111/j.1530-0277.2011.01452.x

Edwards, A. C., Joinson, C., Dick, D. M., Kendler, K. S., Macleod, J., Munafò, M., … Heron, J. (2014). The association between depressive symptoms from early to late adolescence and later use and harmful use of alcohol. *European Child & Adolescent Psychiatry, 23*(12), 1219–1230. https://doi.org/10.1007/s00787-014-0600-5

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods*, *5*(2), 155–174. https://doi.org/10.1037/1082-989X.5.2.155

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, *7*(1), 1–26. https://doi.org/10.1214/aos/1176344552

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970319

Fleming, C. B., Mason, W. A., Mazza, J. J., Abbott, R. D., & Catalano, R. F. (2008). Latent growth modeling of the relationship between depressive symptoms and substance use

during adolescence. *Psychology of Addictive Behaviors*, *22*(2), 186–197. https://doi.org/10.1037/0893-164X.22.2.186

Flora, D. B. (2018). *Statistical methods for the social and behavioral sciences. A model-based approach*. Thousand Oaks, CA: Sage.

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, *2*(1–2), 56–78. https://doi.org/10.1002/hbm.460020107

Gałecka, E., Szemraj, J., Bieńkiewicz, M., Majsterek, I., Przybyłowska-Sygut, K., Gałecki, P., & Lewiński, A. (2013). Single nucleotide polymorphisms of NR3C1 gene and recurrent depressive disorder in population of Poland. *Molecular Biology Reports*, *40*(2), 1693–1699. https://doi.org/10.1007/s11033-012-2220-9

Gatt, J. M., Williams, L. M., Schofield, P. R., Dobson-Stone, C., Paul, R. H., Grieve, S. M., … Nemeroff, C. B. (2010). Impact of the HTR3A gene with early life trauma on emotional brain networks and depressed mood. *Depression and Anxiety*, *27*(8), 752–759. https://doi.org/10.1002/da.20726

Grace, J. B., & Bollen, K. A. (2008). Representing general theoretical concepts in structural equation models: The role of composite variables. *Environmental and Ecological Statistics*, *15*(2), 191–213. https://doi.org/10.1007/s10651-007-0047-7

Hair, J. F., Pieper, T. M., & Ringle, C. M. (2012). The use of partial least squares structural equation modeling in strategic management research: A review of past practices and recommendations for future applications. *Long Range Planning*, *45*(5–6), 320–340. https://doi.org/10.1016/J.LRP.2012.09.008

Hariri, A. R., & Weinberger, D. R. (2003). Functional neuroimaging of genetic variation in serotonergic neurotransmission. *Genes, Brain and Behavior*, *2*(6), 341–349.

https://doi.org/10.1046/j.1601-1848.2003.00048.x

Hashimoto, H., Shintani, N., & Baba, A. (2006). New insights into the central PACAPergic system from the phenotypes in PACAP- and PACAP receptor-knockout mice. *Annals of the New York Academy of Sciences*, *1070*(1), 75–89. https://doi.org/10.1196/annals.1317.038

Henseler, J., Dijkstra, T. K., Sarstedt, M., Ringle, C. M., Diamantopoulos, A., Straub, D. W., … Calantone, R. J. (2014). Common beliefs and reality about PLS: Comments on Rönkkö and Evermann (2013). *Organizational Research Methods*, *17*(2), 182–209. https://doi.org/10.1177/1094428114526928

Henseler, J., Hubona, G., & Ray, P. A. (2016). Using PLS path modeling in new technology research: Updated guidelines. *Industrial Management and Data Systems*, *116*(1), 2–20. https://doi.org/10.1108/IMDS-09-2015-0382

Holmes, A. J., Bogdan, R., & Pizzagalli, D. A. (2010). Serotonin transporter genotype and action monitoring dysfunction: A possible substrate underlying increased vulnerability to depression. *Neuropsychopharmacology*, *35*(5), 1186–1197. https://doi.org/10.1038/npp.2009.223

Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika*, *2*(1), 41–54. https://doi.org/10.1007/BF02287965

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods*, *12*(2), 205–218. https://doi.org/10.1037/1082-989X.12.2.205

Hwang, H., Malhotra, N. K., Kim, Y., Tomiuk, M. A., & Hong, S. (2010). A comparative study on parameter recovery of three approaches to structural equation modeling. *Journal of Marketing Research*, *47*(4), 699–712. https://doi.org/10.2139/ssrn.1585305

Hwang, H., & Takane, Y. (2004). Generalized structured component analysis. *Psychometrika,*

69(1), 81–99. https://doi.org/10.1007/BF02295841

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

Hwang, H., Takane, Y., & Jung, K. (2017). Generalized structured component analysis with uniqueness terms for accommodating measurement error. *Frontiers in Psychology, 8*, 2137. https://doi.org/10.3389/fpsyg.2017.02137

Jarvis, C. B., MacKenzie, S. B., & Podsakoff, P. M. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research*, *30*(2), 199–218. https://doi.org/10.1086/376806

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, *16*(1), 39–59. Retrieved from http://www.jstor.org/stable/184129

Jöreskog, K. G. (1970). Estimation and testing of simplex models. *British Journal of Mathematical and Statistical Psychology*, *23*(2), 121–145. https://doi.org/10.1111/j.2044-8317.1970.tb00439.x

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

Jöreskog, K. G., & Wold, H. (1982). The ML and PLS techniques for modeling with latent variables: Historical and comparative aspects. In H. Wold & K. G. Jöreskog (Eds.), *Systems under indirect observation: Causality, structure, prediction, part I* (pp. 263–270). Amsterdam, Netherlands: North Holland.

Jung, K., Takane, Y., Hwang, H., & Woodward, T. S. (2012). Dynamic GSCA (Generalized

Structured Component Analysis) with applications to the analysis of effective connectivity in functional neuroimaging data. *Psychometrika*, *77*(4), 827–848. https://doi.org/10.1007/s11336-012-9284-2

Jung, K., Takane, Y., Hwang, H., & Woodward, T. S. (2016). Multilevel dynamic generalized structured component analysis for brain connectivity analysis in functional neuroimaging data. *Psychometrika*, *81*(2), 565–581. https://doi.org/10.1007/s11336-015-9440-6

Kessler, R. C., Birnbaum, H., Bromet, E., Hwang, I., Sampson, N., & Shahly, V. (2010). Age differences in major depression: Results from the national comorbidity survey replication (NCS-R). *Psychological Medicine*, *40*(2), 225–237. https://doi.org/10.1017/S0033291709990213

Kiers, H. A. L., Takane, Y., & ten Berge, J. M. F. (1996). The analysis of multitrait-multimethod matrices via constrained components analysis. *Psychometrika*, *61*(4), 601–628. https://doi.org/10.1007/BF02294039

Kim, G., Shin, B., & Grover, V. (2010). Research note: Investigating two contradictory views of formative measurement in information systems research. *MIS Quarterly*, *34*(2), 345–365. https://doi.org/10.2307/20721431

Lee, B. O., Lee, C. H., Lee, P. G., Choi, M. J., & Namkoong, K. (2000). Development of Korean version of alcohol use disorder identification test (AUDIT-K): Its reliability and validity. *Journal of Korean Academy for Addiction Psychiatry*, *4*(2), 85–94. Retrieved from https://ir.ymlib.yonsei.ac.kr/handle/22282913/172500

Lee, S., Choi, S., Kim, Y. J., Kim, B. J., Hwang, H., & Park, T. (2016). Pathway-based approach using hierarchical components of collapsed rare variants. *Bioinformatics*, *32*(17), i586–i594. https://doi.org/10.1093/bioinformatics/btw425

Lehmann, M. L., Mustafa, T., Eiden, A. M., Herkenham, M., & Eiden, L. E. (2013). PACAP-deficient mice show attenuated corticosterone secretion and fail to develop depressive behavior during chronic social defeat stress. *Psychoneuroendocrinology*, *38*(5), 702–715. https://doi.org/10.1016/j.psyneuen.2012.09.006

Lei, P.-W., & Wu, Q. (2012). Estimation in structural equation modeling. In R. H. Holye (Ed.), *Handbook of structural equation modeling.* (pp. 164–180). New York, NY: Guilford Press.

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

Lowe, S. R., Pothen, J., Quinn, J. W., Rundle, A., Bradley, B., Galea, S., … Koenen, K. C. (2015). Gene-by-social-environment interaction (GxSE) between ADCYAP1R1 genotype and neighborhood crime predicts major depression symptoms in trauma-exposed women. *Journal of Affective Disorders*, *187*, 147–150. https://doi.org/10.1016/j.jad.2015.08.002

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*(1), 201–226. https://doi.org/10.1146/annurev.psych.51.1.201

MacCallum, R. C., & Browne, M. W. (1993). The use of causal indicators in covariance structure models: Some practical issues. *Psychological Bulletin*, *114*(3), 533–541. https://doi.org/10.1037/0033-2909.114.3.533

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344–362. https://doi.org/10.1037/1082-989X.11.4.344

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Erlbaum.

McQuaid, R. J., McInnis, O. A., Stead, J. D., Matheson, K., & Anisman, H. (2013). A paradoxical association of an oxytocin receptor gene polymorphism: Early-life adversity and vulnerability to depression. *Frontiers in Neuroscience, 7*, 128. https://doi.org/10.3389/fnins.2013.00128

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122. https://doi.org/10.1007/BF02294746

Miettunen, J., Murray, G. K., Jones, P. B., Mäki, P., Ebeling, H., Taanila, A., … Moilanen, I. (2014). Longitudinal associations between childhood and adulthood externalizing and internalizing psychopathology and adolescent substance use. *Psychological Medicine*, *44*(8), 1727–1738. https://doi.org/DOI: 10.1017/S0033291713002328

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial behavior: A developmental taxonomy. *Psychological Review*, *100*(4), 674–701. https://doi.org/10.1037/0033-295X.100.4.674

Mooijaart, A., & Bentler, P. M. (1986). Random polynomial factor analysis. In E. Diday, Y. Escoufier, L. Lebart, J. Pages, Y. Schektman, & R. Romassone (Eds.), *Data analysis and informatics, IV: Proceedings of the fourth international symposium on data analysis and informatics* (pp. 241–250). Amsterdam, Netherlands: Elsevier.

Mulaik, S. (2010). *Foundations of factor analysis*. *Foundations of factor analysis* (2nd ed.). New York: Chapman and Hall/CRC Press. https://doi.org/https://doi.org/10.1201/b15851

Muthén, B. O. (1989). Factor structure in groups selected on observed scores. *British Journal of Mathematical and Statistical Psychology*, *42*(1), 81–90. https://doi.org/10.1111/j.2044-8317.1989.tb01116.x

Nolen-Hoeksema, S. (2004). Gender differences in risk factors and consequences for alcohol use and problems. *Clinical Psychology Review*, *24*(8), 981–1010. https://doi.org/https://doi.org/10.1016/j.cpr.2004.08.003

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Oh, S. M., Min, K. J., & Park, D. B. (1999). A study on the standardization of the hospital anxiety and depression scale for Koreans: A comparison of normal, depressed and anxious groups. *Journal of Korean Neuropsychiatric Association*, *38*(2), 289–296. Retrieved from https://koreamed.org/article/0055JKNA/1999.38.2.289

Patten, S. B., Williams, J. V. A., Lavorato, D. H., Wang, J. L., Bulloch, A. G. M., & Sajobi, T. (2016). The association between major depression prevalence and sex becomes weaker with age. *Social Psychiatry and Psychiatric Epidemiology*, *51*(2), 203–210. https://doi.org/10.1007/s00127-015-1166-3

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*(2), 287–312. https://doi.org/10.1207/S15328007SEM0802_7

Peper, J. S., Brouwer, R. M., Boomsma, D. I., Kahn, R. S., & Hulshoff Pol, H. E. (2007). Genetic influences on human brain structure: A review of brain imaging studies in twins. *Human Brain Mapping*, *28*(6), 464–473. https://doi.org/10.1002/hbm.20398

Piccinelli, M., & Wilkinson, G. (2000). Gender differences in depression: Critical review. *British Journal of Psychiatry*, *177*(06), 486–492. https://doi.org/10.1192/bjp.177.6.486

Rasetti, R., & Weinberger, D. R. (2011). Intermediate phenotypes in psychiatric disorders. *Current Opinion in Genetics & Development*, *21*(3), 340–348.

https://doi.org/10.1016/J.GDE.2011.02.003

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. https://doi.org/10.1037/met0000220

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

Rigdon, E. E., Sarstedt, M., & Ringle, C. M. (2017). On comparing results from CB-SEM and PLS-SEM: Five perspectives and five recommendations. *Marketing ZFP*, *39*(3), 4–16. https://doi.org/10.15358/0344-1369-2017-3-4

Ringle, C. M., Sarstedt, M., & Straub, D. W. (2012). Editor's comments: A critical look at the use of PLS-SEM in "MIS Quarterly." *MIS Quarterly*, *36*, iii–xiv. https://doi.org/10.2307/41410402

Roman, C. W., Lezak, K. R., Hartsock, M. J., Falls, W. A., Braas, K. M., Howard, A. B., … May, V. (2014). PAC1 receptor antagonism in the bed nucleus of the stria terminalis (BNST) attenuates the endocrine and behavioral consequences of chronic stress. *Psychoneuroendocrinology*, *47*, 151–165. https://doi.org/10.1016/j.psyneuen.2014.05.014

Romdhani, H., Hwang, H., Paradis, G., Roy-Gagnon, M. H., & Labbe, A. (2015). Pathway-based association study of multiple candidate genes and multiple traits using structural equation models. *Genetic Epidemiology*, *39*(2), 101–113. https://doi.org/10.1002/gepi.21872

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Salat, D. H., Buckner, R. L., Snyder, A. Z., Greve, D. N., Desikan, R. S. R., Busa, E., … Fischl,

B. (2004). Thinning of the cerebral cortex in aging. *Cerebral Cortex*, *14*(7), 721–730. https://doi.org/10.1093/cercor/bhh032

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research*, *69*(10), 3998–4010. https://doi.org/10.1016/j.jbusres.2016.06.007

Schuberth, F., Henseler, J., & Dijkstra, T. K. (2018). Confirmatory composite analysis. *Frontiers in Psychology*, *9*, 2541. https://doi.org/10.3389/fpsyg.2018.02541

Sen, S., Nesse, R. M., Stoltenberg, S. F., Li, S., Gleiberman, L., Chakravarti, A., … Burmeister, M. (2003). A BDNF coding variant is associated with the NEO personality inventory domain neuroticism, a risk factor for depression. *Neuropsychopharmacology*, *28*(2), 397–401. https://doi.org/10.1038/sj.npp.1300053

Sowell, E. R., Peterson, B. S., Kan, E., Woods, R. P., Yoshii, J., Bansal, R., … Toga, A. W. (2007). Sex differences in cortical thickness mapped in 176 healthy individuals between 7 and 87 years of age. *Cerebral Cortex*, *17*(7), 1550–1560. https://doi.org/10.1093/cercor/bhl066

ten Berge, J. M. F. (1993). *Least squares optimization in multivariate analysis*. Leiden, Netherlands: DSWO Press.

Tenenhaus, M. (2008). Component-based structural equation modelling. *Total Quality Management and Business Excellence*, *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. *Computational Statistics & Data Analysis*, *48*(1), 159–205. https://doi.org/10.1016/J.CSDA.2004.03.005

Treiblmaier, H., Bentler, P. M., & Mair, P. (2011). Formative constructs implemented via common factors. *Structural Equation Modeling*, *18*(1), 1–17. https://doi.org/10.1080/10705511.2011.532693

Vaske, J., Makarios, M., Boisvert, D., Beaver, K. M., & Wright, J. P. (2009). The interaction of DRD2 and violent victimization on depression: An analysis by gender and race. *Journal of Affective Disorders*, *112*(1–3), 120–125. https://doi.org/10.1016/j.jad.2008.03.027

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: some further observations. *Multivariate Behavioral Research*, *25*(1), 97–114. https://doi.org/10.1207/s15327906mbr2501_12

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164–e164. https://doi.org/10.1093/nar/gkq603

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnajah (Ed.), *Multivariate analysis* (pp. 391–420). NewYork, NY: Academic Press.

Wold, H. (1973). Nonlinear iterative partial least squares (NIPALS) Modelling: Some current developments. In P. R. Krishnaiah (Ed.), *Multivariate analysis–III* (pp. 383–407). New York, NY: Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-12-426653-7.50032-6

Wold, H. (1982). Soft modeling: The basic design and some extensions. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction, part II* (pp. 1–54). Amsterdam, Netherlands: North Holland.

Zobel, A., Schuhmacher, A., Jessen, F., Hfels, S., Von Widdern, O., Metten, M., … Schwab, S.

G. (2010). DNA sequence variants of the FKBP5 gene are associated with unipolar depression. *International Journal of Neuropsychopharmacology*, *13*(5), 649–660. https://doi.org/10.1017/S1461145709991155

Table 1. A summary of expected behaviors of different SEM approaches in estimating parameters of models with both factors and components. 0 = no bias, + = positive bias, and - = negative bias. CSA = covariance structure analysis, $GSCA_M$ = generalized structured component analysis with measurement errors incorporated, PLSPM = partial least squares path modeling, GSCA = generalized structured component analysis, PLSc = consistent partial least squares, and IGSCA = integrated generalized structured component analysis.

	Factor loadings	Component loadings	Paths connecting factors only	Paths connecting components only	Paths connecting factors and components
CSA	0	-	0	+	+
$GSCA_M$	0	-	0	+	+
PLSPM	+	0	-	0	-
GSCA	+	0	-	0	-
PLSc	0	0	0	0	0
IGSCA	0	0	0	0	0

Table 2. Relative biases expressed as percentages of standardized loadings obtained from different SEM approaches in Simulation Study 1. c = component loading and f = factor loading. CSA = covariance structure analysis, $GSCA_M$ = generalized structured component analysis with measurement errors incorporated, PLSPM = partial least squares path modeling, GSCA = generalized structured component analysis, PLSc = consistent partial least squares, and IGSCA = integrated generalized structured component analysis.

N		z_1 (c)	z_2 (c)	z_3 (c)	z_4 (c)	z_5 (c)	z_6 (c)	z_7 (f)	z_8 (f)	z_9 (f)
100	CSA	-3.52	-15.53	-8.50	-2.08	-16.89	-8.76	-0.58	-0.34	-0.28
	$GSCA_M$	2.41	-17.67	-11.06	1.72	-17.98	-10.45	0.61	0.91	0.82
	PLSPM	-0.24	-0.49	-0.24	-1.42	-1.73	-1.64	23.51	16.49	7.81
	GSCA	-0.15	-0.25	-0.25	-0.07	-0.31	-0.12	24.65	16.24	7.51
	PLSc	-0.24	-0.49	-0.24	-1.42	-1.73	-1.64	-0.82	-0.50	-0.89
	IGSCA	-0.12	-0.27	-0.24	-0.06	-0.34	-0.09	0.04	0.37	0.52
200	CSA	-3.32	-15.53	-8.14	-2.11	-16.71	-8.83	0.20	-0.16	-0.10
	$GSCA_M$	3.25	-18.09	-11.49	2.37	-18.42	-11.36	0.73	0.35	0.54
	PLSPM	-0.06	-0.15	-0.13	-0.39	-0.70	-0.44	24.42	16.64	7.96
	GSCA	-0.05	-0.07	-0.08	-0.07	-0.18	-0.12	25.42	16.34	7.68
	PLSc	-0.06	-0.15	-0.13	-0.39	-0.70	-0.44	0.04	-0.08	-0.38
	IGSCA	-0.02	-0.12	-0.06	-0.04	-0.22	-0.10	0.51	0.03	0.41
500	CSA	-3.20	-15.34	-8.16	-1.93	-16.60	-8.75	-0.16	0.03	0.05
	$GSCA_M$	3.58	-18.23	-11.87	2.55	-18.55	-11.47	-0.11	-0.01	0.63
	PLSPM	0.04	-0.03	-0.04	-0.03	-0.22	-0.17	24.23	16.79	8.08
	GSCA	0.03	0.06	-0.05	0.01	-0.04	-0.06	25.28	16.44	7.77
	PLSc	0.04	-0.03	-0.04	-0.03	-0.22	-0.17	-0.14	0.07	-0.17
	IGSCA	0.04	0.02	-0.03	0.03	-0.09	-0.03	-0.21	-0.07	0.49
1000	CSA	-3.32	-15.44	-8.09	-2.02	-16.54	-8.77	0.10	-0.16	0.07
	$GSCA_M$	3.67	-18.44	-12.00	2.48	-18.57	-11.58	0.08	-0.05	0.25
	PLSPM	0.00	-0.02	-0.04	-0.10	0.05	-0.18	24.48	16.70	8.08
	GSCA	-0.02	0.03	-0.04	-0.04	0.03	-0.07	25.52	16.36	7.75
	PLSc	0.00	-0.02	-0.04	-0.10	0.05	-0.18	0.23	-0.27	0.00
	IGSCA	-0.01	-0.01	-0.01	-0.02	-0.02	-0.04	-0.03	-0.04	0.19

Table 3. Relative biases expressed as percentages of standardized path coefficients obtained from different SEM approaches in Simulation Study 1. CSA = covariance structure analysis, $GSCA_M$ = generalized structured component analysis with measurement errors incorporated, PLSPM = partial least squares path modeling, GSCA = generalized structured component analysis, PLSc = consistent partial least squares, and IGSCA = integrated generalized structured component analysis.

N		CSA	$GSCA_M$	PLSPM	GSCA	PLSc	IGSCA
100	b_1	12.81	10.82	-13.01	-12.45	-0.15	-1.40
	b_2	16.51	15.15	-11.99	-12.03	1.03	-0.96
200	b_1	12.53	12.44	-12.85	-12.73	0.01	-0.62
	b_2	15.91	16.59	-12.49	-12.73	0.43	-0.60
500	b_1	12.80	13.79	-12.71	-12.72	0.24	-0.02
	b_2	15.75	17.87	-12.77	-12.92	0.18	-0.24
1000	b_1	12.66	14.00	-12.88	-12.93	0.07	-0.03
	b_2	15.54	18.14	-13.01	-13.12	-0.07	-0.25

Table 4. The number of samples with improper solutions per condition over 1000 samples in PLSc. M1 = correct model, M2 = under-parameterized I, M3 = over-parameterized I, M4 = under-parameterized II, and M5 = over-parameterized II.

		$N_{ind} = 3$					$N_{ind} = 5$					$N_{ind} = 7$				
σ	N	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5	M1	M2	M3	M4	M5
0.1	100	423	492	418	507	400	386	404	390	470	358	451	467	451	525	435
	200	288	357	288	386	273	265	277	264	350	245	311	311	311	404	309
	400	146	188	146	211	131	145	153	145	198	132	167	163	166	229	158
	1000	77	93	78	127	76	82	84	82	127	77	86	97	86	131	81
0.3	100	396	479	401	479	366	378	412	378	441	361	441	444	444	523	421
	200	267	334	266	372	241	260	289	261	353	232	328	334	326	402	293
	400	154	175	153	238	141	152	163	152	214	132	175	173	176	247	154
	1000	96	129	97	148	78	116	124	116	154	97	106	108	107	166	88
0.5	100	400	485	396	487	366	391	417	397	467	350	409	422	412	505	383
	200	303	363	304	383	264	272	293	275	347	234	298	293	300	379	275
	400	187	217	186	244	161	167	175	166	230	135	190	196	192	256	154
	1000	121	155	123	198	110	127	147	129	188	108	131	142	131	193	107

Table 5. The correlations, means, and standard deviations (SD) of the gene and depression data. z_1 = HAD2, z_2 = HAD4, z_3 = HAD6, z_4 = HAD8, z_5 = HAD10, z_6 = HAD12, and z_7 = HAD14, z_8 = rs25531, z_9 = rs9296158, z_{10} = rs3800373, z_{11} = rs1360780, z_{12} = rs9470080, z_{13} = rs4713916, z_{14} = rs4713919, z_{15} = rs6902321, z_{16} = rs56311918, z_{17} = rs3798345, z_{18} = rs2267735, z_{19} = rs6265, z_{20} = rs4680, z_{21} = rs4633, z_{22} = rs1062613, z_{23} = rs2075652, z_{24} = rs258747, z_{25} = rs53576, z_{26} = gender, z_{27} = age, and z_{28} = AUDIT.

D d	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	z_{10}	z_{11}	z_{12}	z_{13}	z_{14}	z_{15}	z_{16}	z_{17}	z_{18}	z_{19}	z_{20}	z_{21}	z_{22}	z_{23}	z_{24}	z_{25}	z_{26}	z_{27}	z_{28}
z_1	1																											
z_2	.50	1																										
z_3	.55	.63	1																									
z_4	.27	.37	.35	1																								
z_5	.47	.44	.53	.40	1																							
z_6	.48	.60	.63	.46	.45	1																						
z_7	.45	.36	.44	.24	.45	.38	1																					
z_8	.02	-.03	-.07	.08	-.03	.03	-.08	1																				
z_9	.00	-.05	-.06	-.00	-.05	-.08	-.01	-.06	1																			
z_{10}	-.08	-.05	-.14	-.01	-.07	-.08	-.01	-.01	.79	1																		
z_{11}	-.09	-.06	-.11	.01	-.05	-.08	-.02	-.03	.81	.93	1																	
z_{12}	.04	.01	-.04	.01	-.01	-.07	-.02	-.04	.92	.75	.78	1																
z_{13}	-.02	.04	-.05	.05	-.00	.01	-.05	-.01	.69	.79	.84	.79	1															
z_{14}	-.02	.03	-.09	.03	-.01	-.03	-.04	.01	.67	.69	.74	.77	.82	1														
z_{15}	.07	.05	.02	.02	.02	-.03	-.04	-.06	.85	.65	.69	.93	.82	.75	1													
z_{16}	-.02	.04	-.00	.05	-.00	.04	-.08	-.03	.62	.68	.75	.64	.82	.69	.68	1												
z_{17}	-.01	-.03	-.07	.03	-.02	-.00	-.02	-.04	.73	.81	.85	.68	.87	.72	.72	.78	1											
z_{18}	.07	.08	.12	.11	.12	.06	.06	-.04	.06	-.01	.02	.07	.04	.07	.06	-.00	.01	1										
z_{19}	-.04	-.07	-.01	-.02	-.05	-.06	.02	-.03	.09	.06	.08	.10	.04	.13	.05	.02	.04	-.01	1									
z_{20}	-.12	-.10	-.00	.06	.06	-.00	-.00	-.03	.11	.05	.06	.07	-.00	.02	.04	.01	.05	-.08	.07	1								
z_{21}	-.14	-.14	-.03	.03	.04	-.02	-.04	-.04	.12	.06	.04	.07	-.02	-.03	.04	-.01	.04	-.08	.04	.93	1							
z_{22}	.08	.09	.08	-.04	.02	.01	.14	.00	.03	.03	.01	.03	-.01	-.03	.01	-.08	-.00	-.03	.05	.03	.03	1						
z_{23}	-.10	-.01	-.02	-.10	-.10	.01	-.07	.03	-.04	-.02	.01	-.03	.04	.01	.01	.06	.03	-.01	-.04	.01	.02	-.06	1					
z_{24}	.01	-.02	.07	-.00	.01	.02	.02	-.07	.06	.03	.04	.05	.05	.07	.05	.03	.07	.02	.12	.14	.12	.05	.13	1				
z_{25}	-.08	-.05	-.03	.10	-.11	-.02	-.13	.04	.07	-.01	.02	.04	.03	-.01	.05	.10	.08	.03	-.03	.02	.06	-.01	-.00	-.13	1			
z_{26}	-.14	-.18	-.17	-.10	-.12	-.19	-.07	.03	-.04	-.08	-.07	-.01	-.08	-.04	-.00	-.17	-.07	.17	-.05	.05	.04	.09	-.05	-.02	-.02	1		
z_{27}	.03	-.08	-.04	.03	.08	-.10	-.01	-.01	-.00	-.03	-.02	.04	-.01	.04	.03	-.02	-.02	.07	.11	-.01	.01	.08	-.16	.12	-.03	.02	1	
z_{28}	.13	.09	.14	.06	.07	.19	.02	-.06	.06	.08	.07	.06	.14	.08	.08	.11	.11	-.05	.04	.00	.01	.01	.10	-.01	-.02	-.44	-.21	1
Mean	1.51	0.91	1.68	1.60	1.03	1.25	0.82	1.25	1.58	1.41	1.43	1.61	1.42	1.52	1.56	1.29	1.35	1.97	1.88	1.55	1.54	1.17	1.83	1.53	1.77	1.68	46.10	3.04
SD	1.13	0.81	0.79	0.75	0.95	1.01	0.82	0.43	0.63	0.59	0.60	0.64	0.58	0.64	0.62	0.49	0.55	0.70	0.72	0.60	0.59	0.41	0.74	0.63	0.72	0.47	13.46	3.61

Table 6. The standardized loadings estimated from integrated generalized structured component analysis for the gene and depression data.

Factor/component	Indicator	Estimate	SE	95% CI
Depression	z_1 = HADS2	0.68	0.04	[0.60, 0.76]
	z_2 = HADS4	0.74	0.03	[0.68, 0.80]
	z_3 = HADS6	0.81	0.04	[0.74, 0.88]
	z_4 = HADS8	0.51	0.05	[0.40, 0.60]
	z_5 = HADS10	0.66	0.05	[0.56, 0.75]
	z_6 = HADS12	0.76	0.04	[0.68, 0.83]
	z_7 = HADS14	0.56	0.06	[0.43, 0.68]
SLC6A4	z_8 = rs25531	1.00	0	[1.00, 1.00]
FKBP5	z_9 = rs9296158	0.89	0.02	[0.85, 0.92]
	z_{10} = rs3800373	0.89	0.03	[0.83, 0.93]
	z_{11} = rs1360780	0.93	0.01	[0.90, 0.95]
	z_{12} = rs9470080	0.91	0.01	[0.88, 0.93]
	z_{13} = rs4713916	0.93	0.01	[0.90, 0.95]
	z_{14} = rs4713919	0.86	0.02	[0.80, 0.90]
	z_{15} = rs6902321	0.89	0.02	[0.85, 0.92]
	z_{16} = rs56311918	0.83	0.02	[0.78, 0.88]
	z_{17} = rs3798345	0.90	0.02	[0.85, 0.93]
ADCYAP1R1	z_{18} = rs2267735	1.00	0	[1.00, 1.00]
BDNF	z_{19} = rs6265	1.00	0	[1.00, 1.00]
COMT	z_{20} = rs4680	0.98	0.01	[0.97, 0.99]
	z_{21} = rs4633	0.98	0.01	[0.97, 0.99]
HTR3A	z_{22} = rs1062613	1.00	0	[1.00, 1.00]
DRD2	z_{23} = rs2075652	1.00	0	[1.00, 1.00]
NR3C1	z_{24} = rs258747	1.00	0	[1.00, 1.00]
OXTR	z_{25} = rs53576	1.00	0	[1.00, 1.00]

Table 7. The standardized path coefficients estimated from integrated generalized structured component analysis for the gene and depression data.

			Estimate	SE	95% CI
Genes	SLC6A4	→ Depression (b_1)	0.01	0.07	[-0.12, 0.14]
	FKBP5	→ Depression (b_2)	-0.06	0.06	[-0.18, 0.07]
	ADCYAP1R1	→ Depression (b_3)	0.18	0.07	[0.03, 0.30]
	BDNF	→ Depression (b_4)	-0.05	0.07	[-0.18, 0.08]
	COMT	→ Depression (b_5)	-0.03	0.07	[-0.18, 0.11]
	HTR3A	→ Depression (b_6)	0.10	0.08	[-0.05, 0.26]
	DRD2	→ Depression (b_7)	-0.10	0.07	[-0.23, 0.03]
	NR3C1	→ Depression (b_8)	0.04	0.07	[-0.09, 0.17]
	OXTR	→ Depression (b_9)	-0.07	0.07	[-0.21, 0.06]
Covariates	Gender	→ Depression (b_{10})	-0.24	0.08	[-0.39, -0.07]
	Age	→ Depression (b_{11})	-0.06	0.07	[-0.21, 0.08]
	AUDIT	→ Depression (b_{12})	0.09	0.09	[-0.08, 0.25]
	Gender × Age	→ Depression (b_{13})	-0.10	0.06	[-.023, 0.02]
	Gender × AUDIT	→ Depression (b_{14})	0.07	0.06	[-0.06, 0.18]

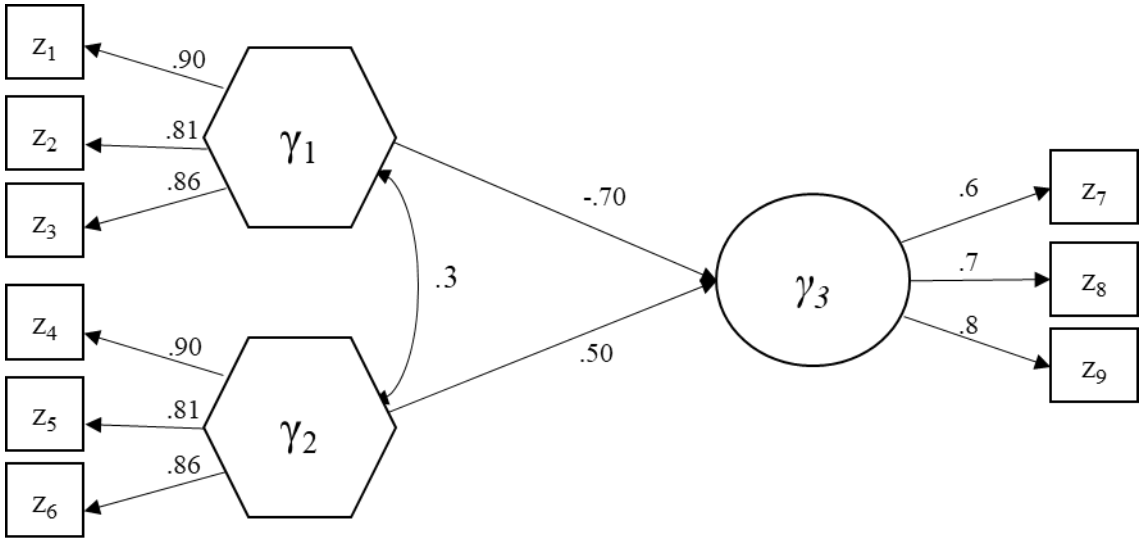

Figure 1. The data generating structural equation model specified for the first simulation study. Squares denote indicators, and circles and hexagons represent factors and components, respectively. Arrows signify loadings or path coefficients. All weights for composite indicators and residual terms are omitted.

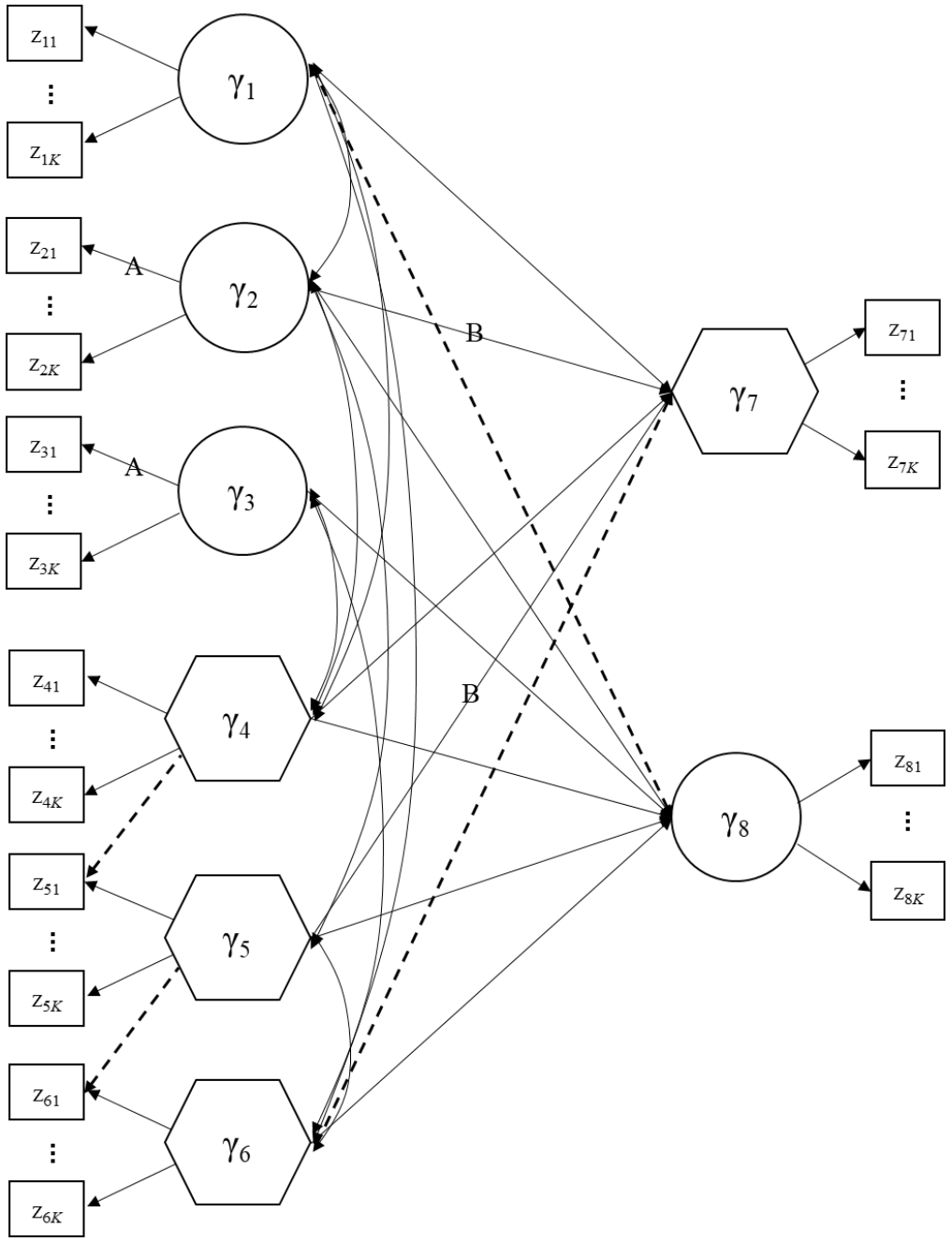

Figure 2. The data generating structural equation model specified for the second simulation study. Squares denote indicators, where z_{pk} is the *k*th indicator for the *p*th factor/component ($k = 1,\ldots,K$). Circles and hexagons represent factors and components, respectively. Arrows signify loadings or path coefficients in the data generating model. Dashed arrows indicate additional loadings or path coefficients in the over-parameterized models. The two loadings labeled A and the two path coefficients labeled B are excluded in the under-parameterized models I and II, respectively. All weights for composite indicators and residual terms are omitted.

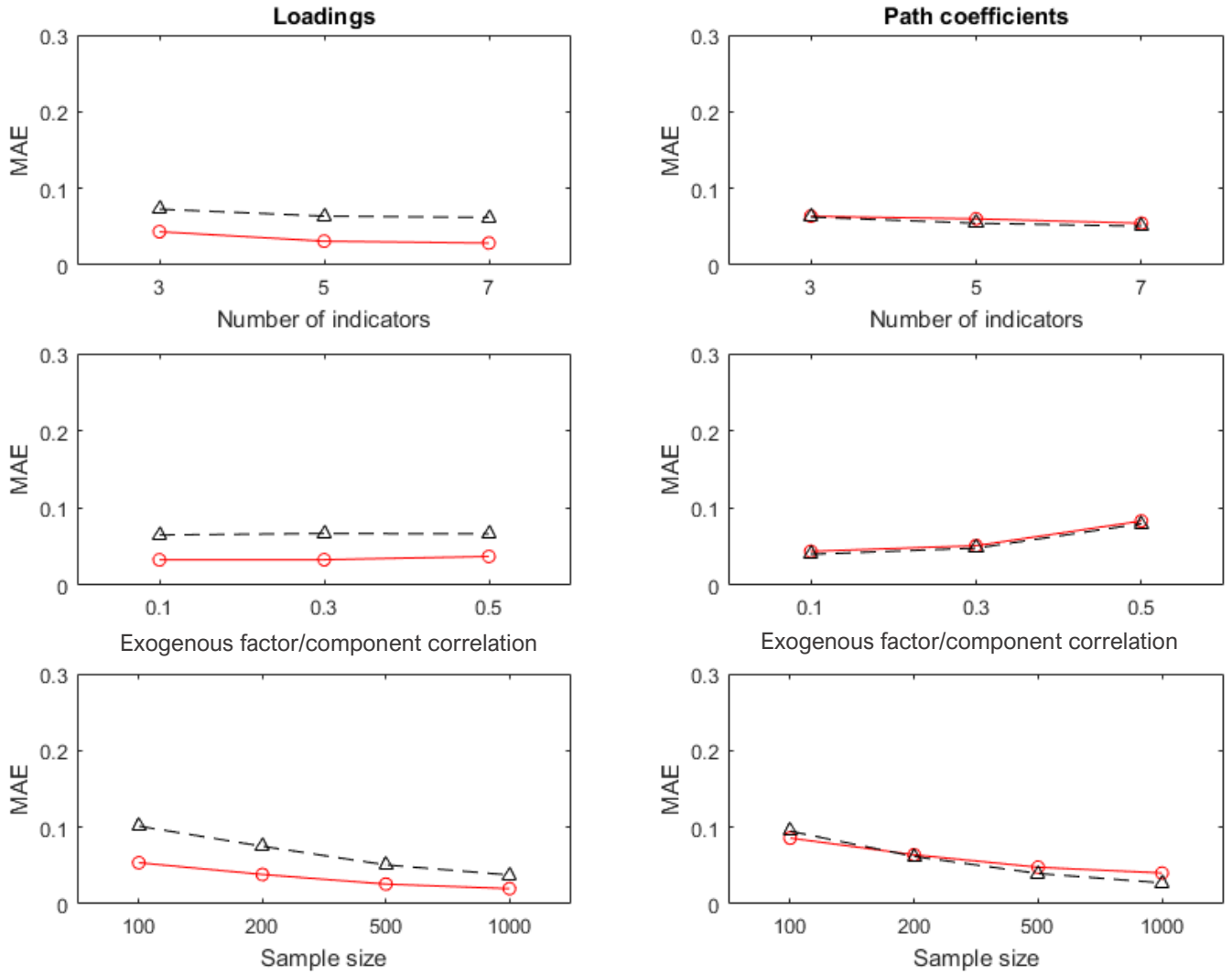

Figure 3. Average mean absolute error (MAE) values of the estimates of integrated generalized structured component analysis (IGSCA) and consistent partial least squares (PLSc) per condition under correct specification (o: IGSCA and Δ: PLSc).

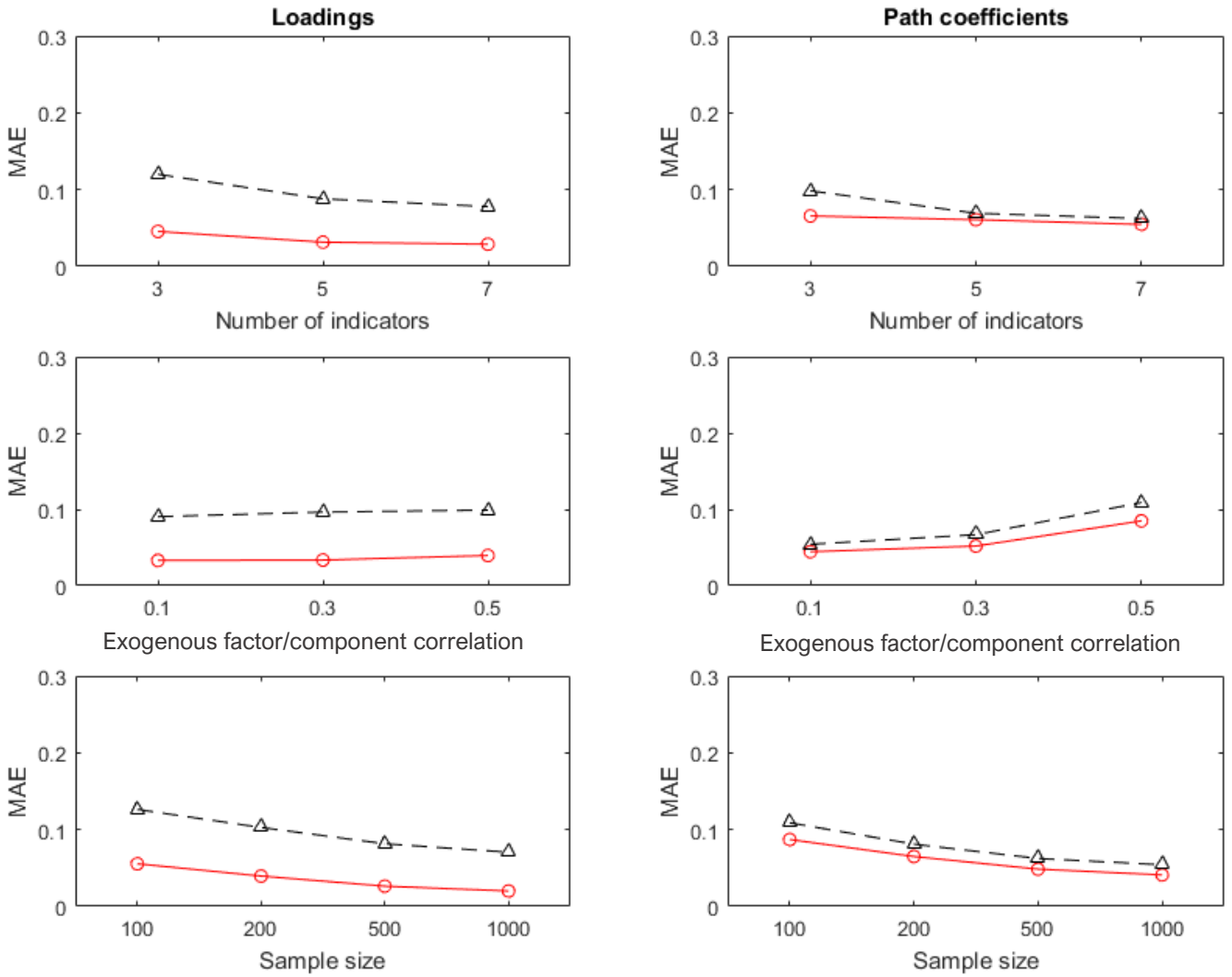

Figure 4. Average mean absolute error (MAE) values of the estimates of integrated generalized structured component analysis (IGSCA) and consistent partial least squares (PLSc) per condition under the over-parameterized model I (o: IGSCA and Δ: PLSc).

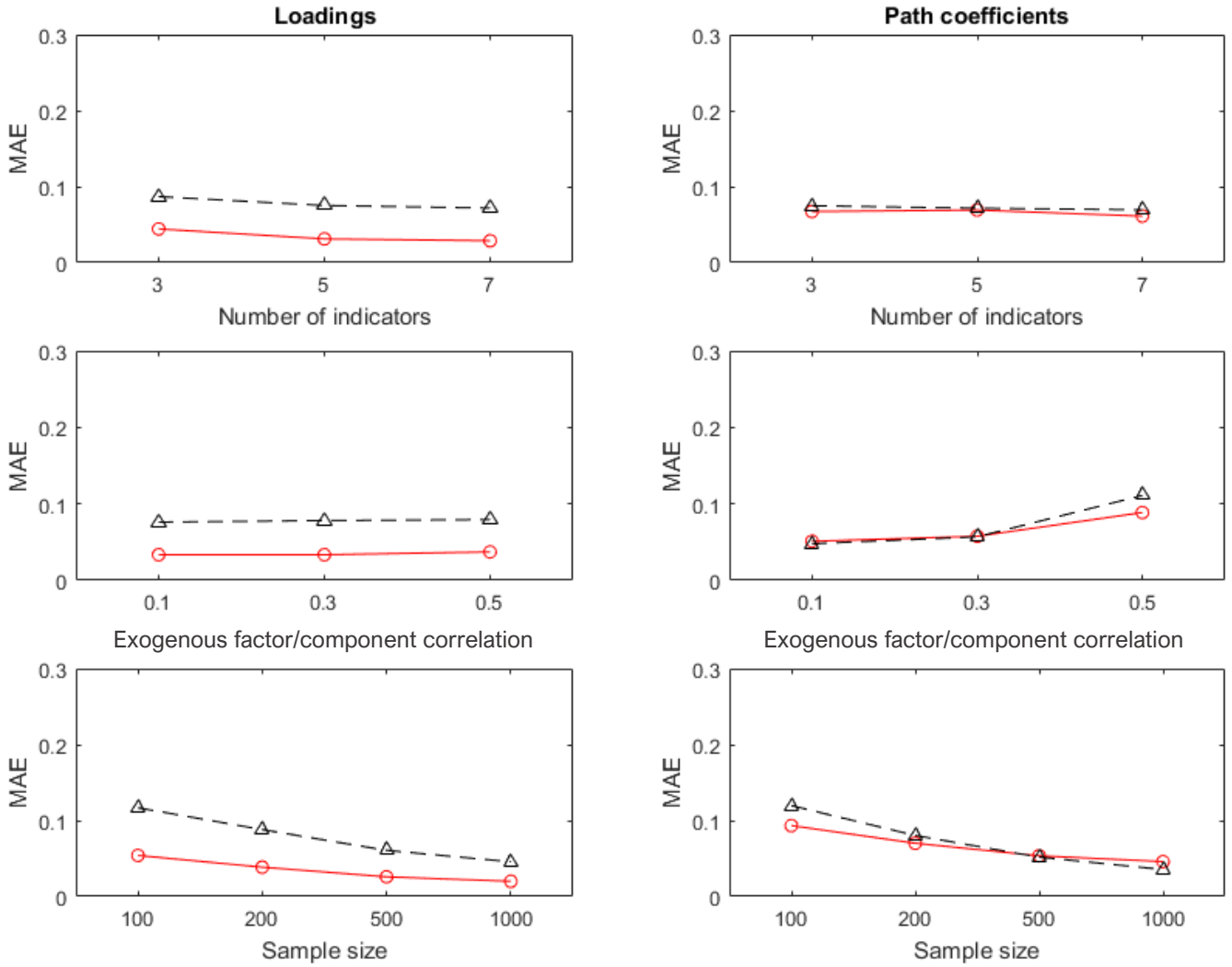

Figure 5. Average mean absolute error (MAE) values of the estimates of integrated generalized structured component analysis (IGSCA) and consistent partial least squares (PLSc) per condition under the under-parameterized model II (o: IGSCA and Δ: PLSc).

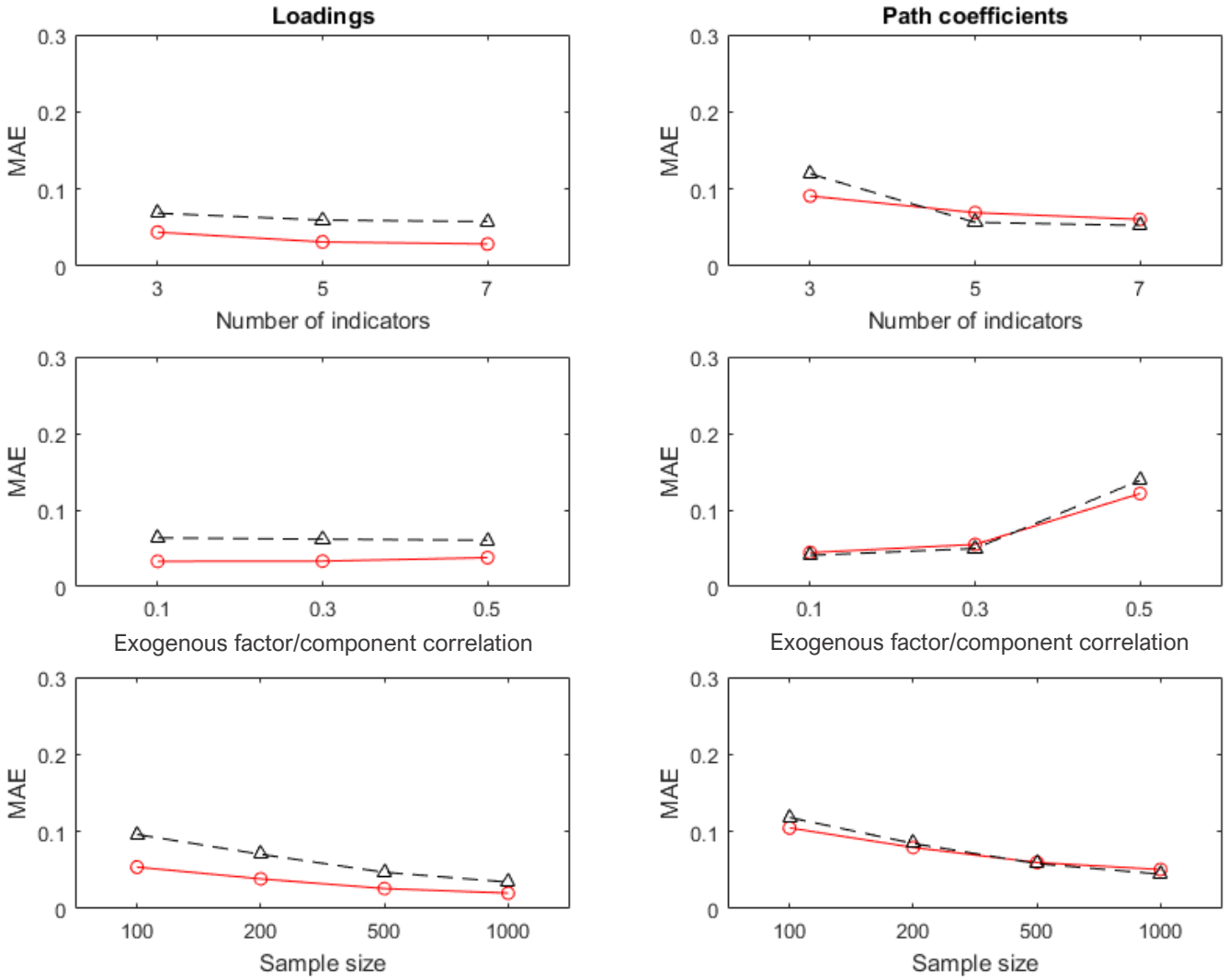

Figure 6. Average mean absolute error (MAE) values of the estimates of integrated generalized structured component analysis (IGSCA) and consistent partial least squares (PLSc) per condition under the over-parameterized model II (o: IGSCA and Δ: PLSc).

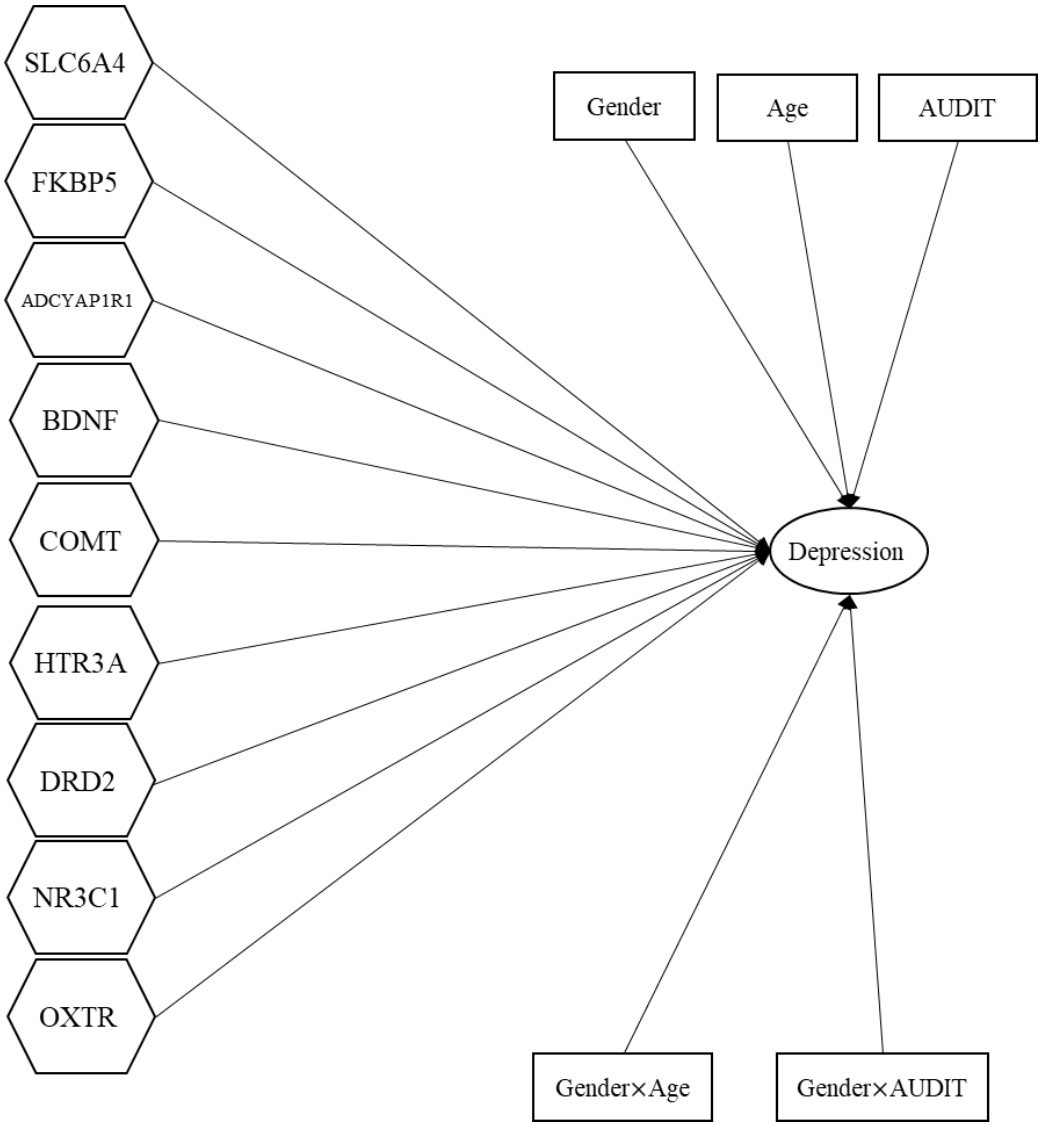

Figure 7. The structural model hypothesized for the gene and depression data. No residual terms are displayed.