UNIVERSITY OF OKLAHOMA
GRADUATE COLLEGE

ACCESSING THE IMPACT OF UNMODELED SYSTEMATIC ERROR VARIANCE ON MEASUREMENT INVARIANCE TESTS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirement for the

Degree of

MASTER OF SCIENCE

By

HE HUANG

Norman, Oklahoma

2022

ACCESSING THE IMPACT OF UNMODELED SYSTEMATIC ERROR VARIANCE ON MEASUREMENT INVARIANCE TESTS

A THESIS APPROVED FOR THE

DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Hairong Song

Dr. Dingjing Shi

Dr. Jordan Loeffelman

Acknowledgements

I would say thank you to Dr. Hairong Song for admitting me as a graduate student. I think it is my honor to be a graduate student at the University of Oklahoma. I also would like to say thank you to my laboratory mates, Yaqi and Adon, for emotional and academic supports. Last, I would love to say thank you to Jichun Zhang (University of Chicago) and Ting Wang (Fudan University). They are two of my friends met in Shanghai, I could not go through the hardest time of graduate school without them.

Table of Contents

List of Tables

List of Figures

Abstract

Systematic error variance (SEV) is one of sources that make a measurement noninvariant (DeShon, 2004). In the confirmatory factor analysis (CFA), researchers use the bifactor or correlated uniqueness (CU) model to control the SEV. This study aims to examine the impacts of SEVs on the multiple groups mean comparison, and evaluate the methods used to control the SEVs in the framework of multiple group CFA. In Monte Carlo simulation, multiple groups data contaminated by different SEV distributions are generated, then, the bifactor and the CU model were used to fit the data. The original model, which assumed no SEVs, was also used as the baseline model. Results show that uncontrolled SEV could affect the estimation of mean difference. Among three models, the bifactor overperformed the other two models in most conditions if it yields converged results. This study also provided an empirical example to demonstrate how to select appropriate methods in multiple group CFA. Implications of these results for applied researchers are discussed.

Keywords: Measurement invariance, confirmatory factor analysis, systematic variance

Introduction

When researchers are interested in group mean difference of a psychological construct, they may use a self-report inventory or scale to assess it. If means of these groups are different, they can conclude that the difference is attributed to group variables (e.g., culture, time). This comparison assumes that the measurement instrument is actually measuring the same construct across groups. In other words, researchers must make sure that the measurement is invariant across groups before making group mean comparison. Therefore, the measurement invariance test becomes an integral part of multiple-group analysis (DeShon, 2004). Measurement invariance test can be also conducted in the framework of the confirmatory factor analysis (CFA). CFA permits the comparison in the term of latent variables so that the measurement error is controlled (Thompson & Green, 2006). In the CFA, the measurement invariance test is applying restraints on a sequence of sets of parameters. The test begins with examining whether the factor structure is the same across groups by setting model configuration identical across groups. Then to restrain all factor loadings as equivalent across groups to test weak or metric invariance; then all intercepts to test strong or scalar invariance; then residuals to test strict invariance. Strong invariance is a necessary condition for construct means comparison.

Systematic error variance (SEV) is one of potential sources which make a measurement non-invariant. SEV is shared error variance among variables. Because it is usually ascribed to using of measurement instruments, the term SEV is believed to be exchangeable for method variance (Richardson et al., 2009; Podsakoff et al., 2003). It frequently appears in research data which is collected by self-report scale (Richardson et al., 2009). Most researchers deem the SEV as a detrimental impact on model parameter estimations; unmodeled systematic error variance usually cause poor model fitting and incorrect model parameters estimations (Podsakoff, et al., 2003; Spector, 2006).

In CFA, researchers may rely on modification indices or their prior knowledge to locate items affected by SEV then control it by modeling the SEV. Currently, two measures are commonly used: the bifactor model and the correlated uniqueness (CU) model (Lance, et al., 2002). By applying the appropriate measure, SEV can be taken out from variance of constructs. Although both methods are widely used in CFA studies, it is still controversial surrounding which method is more appropriate to model the systematic error variance. Two methods not only represent different assumptions of SEV, but also have their advantages and weaknesses (Lance, et al., 2002; Podsakoff et al., 2002; Conway, et al., 2004; Lance, et al., 2010).

Many studies have assessed effectiveness of the bifactor model and CU model in controlling SEV in the analysis of covariance structure, yet results indicated both models are not perfect (e.g., Richardson, et al., 2009; Williams, et al., 2015). In the other aspect, Geiser, et al. (2014) found that systematic error variance caused by measurement instrument can also alter the mean structure. However, few studies touched the topic that examining the effectiveness of both methods in recovering true mean structure (Cheung & Chan, 2002). The main purpose of this paper is to examine the effectiveness of two methods in controlling systematic error variance and whether they can lead to proper conclusion about latent factor mean difference among groups. This study presents a detailed examination of both methods by using Monte Carlo method to simulate different conditions of systematic error variance. Both methods will also be compared with a baseline model which assumes the SEV absent (no controlling). This study concludes with a real data demonstration to give social science researchers suggestions about how to choose appropriate SEV controlling strategy.

The paper will be presented in following order: first, the paper will introduce the measurement invariance and multiple group CFA, then the paper summarized the previous research about systematic error variance and methods used to control the systematic error variance. In the second section, the paper will present the simulation study and results. In the third section, the paper will present a real data demonstration.

Multiple-Group Confirmatory Factor Analysis and Measurement Invariance

In CFA model, a scale consists of n continuous, observable variables Y, and it measures k continuous latent variables η. The CFA model can be expressed as:

$$y_{ij} = \alpha_j + \Sigma_{k=1}^{K} \lambda_{jk} \eta_{jk} + \epsilon_{ij}, (1)$$

Where $i = 1, \dots, I$ for research subjects and $j = 1, \dots, J$ for observable variables or items of a scale; α_j represents the mean of sample, λ_{jk} represents the factor loadings of item j on latent variable k, ϵ_{ij} represents the measurement errors. According to this equation, the effects of latent variables on observable variables are additive.

In multiple group CFA, suppose there are total G groups, the superscript g denote that this parameter is group-specific, equation 1 can also be written as matrices multiplication:

$$y^g = \tau^g + \Lambda^g \eta^g + \epsilon^g, \quad (2)$$

where y^g is a $n \times 1$ vector of scores on n measured variables (or indicator) for each individual. τ^g is $n \times 1$ vector of intercepts on n indicators. Λ is a $n \times k$ matrix of the factor loadings. η is an $n \times 1$ vector of latent variable scores. ϵ is a $n \times 1$ vector of residuals. If take square at both sides, then we obtained the covariance matrix. The covariance matrix of this model can be expressed by equation:

$$\Sigma^g = \tau^g \tau^{g'} + \Lambda^g \Psi^g \Lambda^{g'} + \Theta_\epsilon^g, \quad (3)$$

Where Σ is a $n \times n$ matrix of covariance among n indicators. Ψ is $k \times k$ matrix of covariance among k latent factors, Θ_ϵ^g is a diagonal matrix of the variance components of errors.

In measurement invariance test, to test the configural invariance, researchers should make sure the locations of zeros and non-zero cells are identical across covariance matrices of each group. To test weak invariance, researchers should set equity on factor loadings across groups ($\Lambda^g = \Lambda$). To test strong invariance, researchers then should set equity on both factor loadings and intercepts across groups ($\Lambda^g = \Lambda, \tau^g = \tau$). To test strict invariance, researcher should set equity on factor loadings, intercepts, and residuals covariance matrix ($\Lambda^g = \Lambda, \tau^g = \tau, \Theta^g = \Theta$). When these conditions are met, it is safe to say that measurement invariance is hold across groups, and group mean comparison becomes feasible (Meredith, 1993). To estimate group mean difference, simple set latent factor mean η of an arbitrary group as 0 and leave that of other groups being freely estimated.

Previous Research of Systematic Error Variance

Measurement error variance is variance that cannot be explained by the construct of interests, it can be further partitioned into unique variance and systematic error variance (Spector, 1994; Spector & Brannick, 1995). The latter is defined as shared residual variance among a clutch of variables. Although the error variance is believed as threat to measurement validity and reliability, the systematic error variance is more concerning. For example, Spector (2006) believes the correlation between two variables will be inflated by systematic error variance.

The concept of systematic error variance was initially introduced by Fiske and Campbell (1959): By examining the multitrait-multimethod (MTMM) matrix, they found if two traits are measured by the same method, the measured correlation is the combination of traits and method. Later, other researchers found other sources of systematic error variance, for example: Common sources or raters (e.g., Eden & Leviatin, 1975; Guzzo, et al., 1986); item characteristics (e.g., Thomas & Kilmann, 1975); item context (e.g., Salancik & Pfefffer, 1977; Harrison & McLaughlin, 1993). Unfortunately, due to lacking appropriate analytic tools, researchers are still debating on how to locate the source of SEV and to which extent the systematic error variance has impact on analysis (Spector, 1987; Williams, et al., 1989; Bagozzi, Yi, 1990).

So far, the discussion of systematic error variance is mostly about the method variance. Researchers agree that method variance is the shared variance between observable variables that is ascribed to the way information was collected, while the definition of method varies (Maul, 2013). Most researchers deem all systematic error variance as method variance, since method variance is 'something like systematic variance not attributable to trait under consideration' (Golding, 1977, p.93). This idea is supported by many other researchers (e.g., Fiske, 1982; Bagozzi, 1984; Baumgartner & Steenkamp, 2001; Johnson, 2011; Messick, 1991; Siemsen et al. 2010, Weijters et al., 2010; Edwards, 2008). As Fiske (1982, p.82) noted, the definition of method 'encompassed potential influence at several levels of abstractions', therefore, the correlated residuals may be the combination of multiple method effect. Meanwhile, other researchers hold a narrower definition of method effect (Lance et al, 2009; Sechrest, 2000). According to Lance (2009, 2010), method variance should be able to trace back to certain measurement facets, like item similarity in content, structure, or format which elicit similar response. Meanwhile, the raters' tendency, measurement occasion or situation and item order may be excluded from measurement facets (Podsakoff, et al., 2012). In this perspective, the method variance is a subtype of systematic error variance. Besides, researchers like Spector believe the method effect is an 'urban legend', since the systematic error variance is caused by a limited number of people that cannot report accurately (Brannick et al., 2010; Spector, 2006). As Spector concluded that systematic error variance is the result of 'biases that affect particular sets of variables' (Brannick et al., 2010, p.417).

So far, the discussion about the systematic error variance is limited to examination of covariance structure of multiple constructs in single-group studies, in which correlations of multiple traits are suspected to be contaminated by systematic error variance. However, the single-trait measurement is also not free from systematic error. When researchers use CFA model to estimate a trait without controlling the systematic error variance, they may obtain the incorrect factor loadings. To illustrate, in CFA model, the covariance between two indicators is expressed as:

$$COV_{ij} = \lambda_{ci}\psi\lambda_{cj},$$

where λ_i, λ_j are factor loadings of indicator of i & j, ψ is the variance of latent the factor. In some model identification process, the latent factor mean is set as 0 and variance as 1, then the equation can be rewrite as:

$$COV_{ij} = \lambda_{ci}\lambda_{cj},$$

However, when these two indicators are contaminated by systematic error variance from some unknown sources, the covariance of two indicators can be decomposed into the covariance explained by the construct and the unknown source. Therefore, it can be expressed by equation:

$$COV_{ij} = \lambda_{ci}\lambda_{cj} + \lambda_{si}\psi_s\lambda_{sj},$$

If we assuming the systematic error variance is caused by a measurement facet based on narrow definition of the method effect. Then $\lambda_{si}, \lambda_{si}$ are factor loading of the latent method factor. We can also assume that the systematic error variance is just correlated residuals, so the equation can be also expressed as:

$$COV_{ij} = \lambda_{ci}\lambda_{cj} + COV_{\epsilon},$$

Where COV_{ϵ} represents the correlated residuals. In both cases, if researchers try to identify their model without modeling the systematic error variance, their factor loadings are potentially biased.

Statistical Methods to Control Systematic Error Variance

Some techniques are invented to estimate systematic error variance, for example, the MTMM matrix (Fiske & Campbell, 1959) and marker technique (Lindell & Whitney, 2001). Due to the improvement of statistical computation, the CFA is becoming a popular tool. With help of the CFA model, researchers could not only estimate both covariance structure and mean structure, but also separate systematic error variance from measurement error (Geiser, et al., 2014). Applying bifactor or CU models are the two most common strategies in controlling of SEV, the selection of strategies are based on researchers' knowledge about SEV or other practical considerations.

Bifactor model

Bifactor model, also named as correlated-trait-correlated-method (CTCM), specified a general latent factor that accounts for covariance among all indicators and group factors that account for additional covariance among subsets of indicators (Reise, 2012). The bifactor model assumes that systematic error variance is caused by other unmodeled constructs (e.g., measurement facets), so that SEV are specified as group factors. In bifactor model, single measurement can be decomposed into construct component C_j, other group factor S_k, and unexplained residuals E_{jk}:

$$y_{jk} = \lambda_{Cjk} C_j + \lambda_{Sjk} S_k + E_{jk},$$

where λ_{Cjk} are factor loadings on construct and λ_{Sjk} are factor loadings on group factors, the construct factors and groups factors are assumed to be uncorrelated (Eid et al., 2003; Geiser, et al., 2014).

So far, much research has assessed the performance of bifactor model in a specific application: using bifactor model to control the correlation that is caused by method effect between an independent variable and dependent variable. Many Monte Carlo simulation studies show that bifactor model could accurately recover the true correlation when the model is correctly specified (Conway et al., 2004; Hoogland & Boomsma, 1998; Lance et al., 2007; Le et al., 2009).

Richardson et al. (2009) compared three correctional measures in controlling of method effect and concluded that the bifactor model is not recommended. Specifically, three correctional methods are: correctional marker technique, CFA marker technique, and ULMC technique (bifactor). The first two techniques introduce a conceptually independent ‘marker’ construct whose measurement instrument is identical to independent/dependent variables (e.g. two constructs are measured by self-report scale). While, for the former one, all correlations between constructs would be assessed by a composed score. Since the marker construct is conceptually independent with constructs of interests (e.g., two constructs of interests are cognitive ability and math ability, the marker variable can be sexual orientation), any correlation between marker variable and constructs must be caused by method. The true correlation between two constructs of interests is the measured correlation minus method correlation. In CFA marker technique, all constructs are assessed as latent factors; an indicator has factor loadings on both construct of interest and marker latent variable. In ULMC technique, a method factor is specified. Each indicator has factor loadings on both construct factor and a method factor (see Figure 1). However, across various simulation conditions, the bifactor model is either equal or worse to the CFA marker technique or original model. Later, Williams et al. (2015) revised Richardson et al. (2009), they also reached the same conclusion about the bifactor model.

Beside these simulation studies, other researchers suggest that the bifactor model may cause the inadmissible model estimation (e.g., standardized factor loadings larger than 1.0) or non-convergence (Lance, 2002; Grayson & Marsh, 1994). Based on evidence above, it can be assumed that if the bifactor model is applied to control systematic error variance in multiple group CFA, biased results may be obtained.

Correlated Uniqueness (CU) model

CU model is invented to overcome the nonconvergence and inadmissible solution which cause by using bifactor model (Kenny,1976; Kenny, 1979; Kenny & Berman,1980; Marsh, 1989). The CU model can be written as:

$$y_{ij} = \lambda_{Sij} S_i + \delta_{ij},$$

where δ_{ij} is the combination of unexplained residuals and systematic error variance (Lance et al., 2002). From this expression, the CU model assumes the systematic error variance does not have mean structure.

Despite the flexibility and popularity of the CU model, there are some weaknesses also noticeable. As Lance et al. (2002) summarized, the CU model lacks theoretical soundness, because it does not separate the variance of method effect from other systematic or non-systematic error variance. On the other hand, the CU model could not estimate the correlations between method effects. Moreover, since CU model does not model the confounding effects on measurement of construct mean, it is also criticized for 'creating unmeasured variable problems' (James, 1980). Most serious issue is model estimation. Corway et al. (2004) found the estimations of construct factor loadings and correlation may be biased under the CU model, therefore, lead to inaccurate inference of construct

Systematic Error Variance in Multiple-Group CFA

Comparing with single group study, controlling SEV is trickier in multiple group studies, because SEV may interact with grouping factors. In the general linear model, the term interaction is referred as that the relation between two variables varies over different levels of the third variable. In multiple group CFA, the interaction is that systematic error variance is not identically distributed across groups. In other words, the systematic error variance altered the factor loadings in different extent across groups, which causes the measurement noninvariance. This issue has not received enough attention from researchers.

Some researchers indicate that interaction of SEV is prevalent. For example, the Rosenberg Self-esteem scale (SES) is widely used since its establishment and translated into more than 28 languages (Rosenberg, 1979). Schmitt and Allik (2005) analyzed the SES data from 53 countries and found the factorial structure of SES is not invariant across all countries. From the analysis of factorial structure, they concluded that neutral response bias, which is prevalent in collective cultures, may cause the configural noninvariance. This research implies that some systematic error variance is culturally specific. In other SES and cross-culture studies, using different samples yield contradicting results: some research sample from western populations found negative method effect exists (Motl & DiStefano, 2002; Horan, et al., 2003; Quilty, et al., 2006); other research sample from both eastern and western populations found both negative and positive method effect exist (Wang et al., 2001; Wu, 2008). Such evidence suggest researcher should not assume that the influence of systematic error variance is homogeneously exerted on each group.

Although the systematic error variance has been studied for more than 60 years, previous studies are focus on the covariance structure, few studies explored the influence of systematic error variance on modeling mean structure (Geiser, et al, 2014). While many studies indicate that systematic error variance has mean structure and its properties are similar to psychological constructs (Spector, et al., 2019; Lance, et al. 2011; Maul, 2013; Chen, et al., 2012; Pohl, & Steyer, 2010; Lance, et al., 2010). Based on this assumption, a single measurement is the linear combination of constructs:

$$y_j = \lambda_{cj} C_j + \lambda_{sj} S_j + \epsilon_j,$$

in which S_j is the mean of unexplained latent factor, C_j is the mean of latent factor of construct. In model fitting, it is unclear where such influence would exert on. According to research of method effect, the method effect would inflate the correlation between two constructs. Therefore, it can be assumed that when the systematic error variance is unmodeled, the factor loadings of affected indicators are also inflated, and the mean of unmodeled variances will sneaked into the mean of construct, instead of staying on intercepts of indicators.

Another issue is that minor misspecifications is tolerated in current measurement invariance tests. To find a baseline model, many methodological papers or books suggest that conducting CFA without any restrictions at each group separately, then move to multiple group CFA test if there is no group has large deviation; or testing configural invariance directly (e.g., Bowen & Masa, 2015; Schoot, et al., 2012; Milfont & Fisher, 2010). Therefore, misspecifications in certain groups are averaged by the whole sample, which may lead to inaccurate conclusions of estimation of factor mean in certain groups.

In summary, the impact of systematic error variance on mean structure analysis in multiple group CFA has not received enough attentions from researchers. And this issue is complicated by nonequivalent distribution of SEV. Although bifactor and CU model are extensively used to control the SEV, studies also suggested potential problems in applications of these model in single-group study. Therefore, it is worth to examine both models under the influence of SEV in the framework of multiple group CFA.

Therefore, this study has three objectives. For the first one, I want to know whether minor model misspecifications will affect the estimation of group mean differences. It represents a scenario in which items affected by SEV are inconsistent with researchers' priors. The second objective is examining whether the amount of SEV would affect the estimation of group mean difference. This issue will be examined in two dimensions: incorrect estimation is caused by 1. the overall amount of SEV, or 2. the difference amount of SEV across groups. The third objective is examining whether to apply SEV controlling methods will lead to appropriate conclusions about group mean differences; or, in other words, finding out the strategy of applying appropriate SEV controlling method. Both bifactor and CU model will be used to fit a same dataset and their performance will be compared.

This study would not provide opinions on definitions of the SEV or the method effect. Because this study is interested in mean structure analysis, I assume the SEV also have mean structure. Therefore, the properties of SEV in this study is closer to Lance et al. (2002) and the SEV will be specified as a latent factor.

Method

Simulation Models

The simulated the data is assumed to be collected by a scale which is administrated among multiple populations. This 6-item scale is developed in one population and has a clearly defined single-dimension construct. Previous research suggested 3 out of 6 items are contaminated by systematic error variance, so the researchers consider modifying the original model to control the variance.

Model configurations

The simulation models are inspired by Richardson et al. (2009); however, each unit of analysis is of mean structure. Figure 2 presents 5 configurations of models which were used to generate data. The first condition represents the exact fitting, in which exactly 3 items were contaminated by systematic error variance and researcher has successfully identified them. The second condition represents under fitting, in which 5 items were contaminated by systematic error variance but only 3 are controlled by researchers for both groups. The third condition represents the over fit; in which only 2 items contaminated by systematic error variance, but researchers thought 3 items. The fourth condition represents one of interaction of systematic error variance; in the first group, 3 items are under the influence of systematic error variance, however, the second group is freed of systematic error variance; this condition is named as interaction 1. The fifth condition represents another type of interaction: only 3 items in the first group are under the influence of systematic error variance, but 5 items in the second group; this condition is named as interaction 2.

Fixed parameters

The factor loadings loaded on construct, intercepts, latent factor variance and residual variance are set to be unchanged across all simulation conditions. For simplicity, all factor loadings are set to .8, which represent to medium to high reliability; the variance of construct factor is set to 1.0; the intercepts of indicators are set to 0; and the residual variance of indicators are set to .36 ($1 -.8^2$).

Systematic error variance

The systematic error variance functions as a latent factor, and it is named as error factor. This study wants to know whether the overall amount or the difference in amount of SEV have influences on estimation of mean difference. For the former one, named as equal SEV condition, is achieved by setting the average factor loadings as equivalent across groups; the average factor loadings are chosen from .2, .3, .4, and .5, which represent low, low-medium, medium, and high amount of SEV. For the later one, named as unequal SEV condition, the average factor loading always set to .2 in the second group; in the first, the average factor loading is chosen from .3, .4 and .5, which represent low, medium, and high amounts of difference in SEV.

In each level, factor loadings are not identical within a group, they vary around their average values. For example, when average factor loading is set as .2, if the error factor has three indicators, then three factor loadings are .1, .2, and .3; if the error factor has only two indicators, then two factor loadings are .15 and .25; if the error factor has five indicators, then first three factor loadings follow the rule of three factor loadings, and two extra factor loadings are fixed to .2 and .3 across all conditions. Since in the interaction 1 condition, the second group is freed of SEV, therefore, influence of SEV on two groups cannot be homogeneous. The interaction 1 condition would be independent from rest of simulation models and has 4 levels of factor loadings; for other simulation models, they are used to simulate both equal and unequal SEV conditions.

Sample size

Four sample sizes are used, they are 100, 300, 500, and 1000. They represent the sample size commonly appear in CFA studies.

Latent factors mean difference

One purpose of this study is to answer two questions: whether the systematic error variance will increase the chance of type I error when two groups are equivalent; and whether systematic error variance will increase the type II error when two groups are different. Therefore, two latent factor mean differences are used: 0 for detecting type I error, and .25 for type II error. The reason .25 is used for type II error detection is because it represents the medium effect size that can be detected by the smallest sample size (200 in total) in this experimental design. If the systematic error variance would change the estimation of mean structure, the chance of type I error will be higher as the sample size increase, while the chance of type II error will be higher as sample size decrease.

Data Generation

The process of data generation can be expressed as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} .8 & \lambda_{s1} \\ .8 & \lambda_{s2} \\ .8 & \lambda_{s3} \\ .8 & 0/.2 \\ .8 & 0/.3 \\ .8 & 0 \end{bmatrix} \times \begin{bmatrix} \eta_c \\ \eta_s \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}.$$

$\lambda_{s1}, \lambda_{s2}, \lambda_{s3}$ are factor loadings loaded on error factor. η_c is the mean of construct, the first group is either 0 or .25, and the second group is always 0. η_s is the mean of error factor, the first group is set to .5 and 0 for the second group.

In summary, there are totally $5 \times 4 \times 4 + 4 \times 3 \times 4 = 128$ simulation conditions, and in each condition, two group mean differences are set for testing type I and II error rates.

Analysis procedure

Simulation and analysis are performed through the R platform. An R program is designed for data generation and collection. Two R libraries were primarily used: ‘simsem’ and ‘lavaan’. ‘Simsem’ is an opensource R library developed by the University of Kansas and it is used for data simulation; ‘lavaan’ is also an open-source library used for model fitting created under the main developer Yves Roseel.

In each simulation condition, the program will simulate 1000 pairs of samples for each construct mean difference conditions, totally 2000 pairs of samples. After generation of one dataset, the ‘lavaan’ will use bifactor, CU, and single-factor models to fit this data. Figure 3 shows the configurations of models used to fit the data. At this stage, the program will directly access the mean different with strong invariance setting. Because the CU model is unable to set residual covariances as equivalent across two groups. Therefore, when the bifactor model is used to fit the data, partial invariance is applied, in which the factor loadings loaded on the error factors were freely estimated. After fitting, following information will be extracted: estimated mean difference, p-value associated the mean difference. Nonconvergent model fittings would be discarded. This process repeated for 1000 times.

To make all conditions comparable, the absolute error (AE) was introduced. AE is calculated by estimated mean difference minus true difference. In model identification, the first factor loading is set as 1, which is the default setting of most SEM programs (e.g., lavaan, mplus, amos). By doing so, the estimate latent mean difference is re-scaled by the first factor loading (in this study, the true estimated differences are 0 and .20, respect to 0 and .25).

Two indices are used to evaluate whether these models could obtain correct conclusions about the mean difference: specificity and sensitivity. The specificity is operationalized as the rate of not committing to type I error (true negative rate); the sensitivity is operationalized as the rate of not committing to the type II error (true positive rate).

Another two criteria are used to evaluate the performance of models: accuracy and precision. The accuracy is operationalized as the mean of absolute error; the precision is the width of 95% interval of AE after 1000 simulations.

Results

Rate of Converged Results

Table 1 and 2 present the number of converged results after 2000 times of model fitting. Results indicated that non-converged results could only occur with the bifactor model, while both the CU and single factor model had not any non-converged results. Also, for both equal and unequal SEV conditions, the rate of convergent results increased as the sample size and overall amount of SEV increasing. One exception was the condition of over fitting, the rate of nonconvergent results was higher than other conditions in both equal and unequal SEV conditions, and the increase of sample size did not improve the rate of convergent results.

Assessing Accuracy and Precision

Table 1 and 2 also present the average AE of each model in every simulation conditions. The AE represents the deviation of estimated group mean difference from the true difference. The accuracy is operationalized as the average AE after 2000 times of simulation. The average AE closer to 0, a model has higher probability to obtain true group mean differences. Results showed that the bifactor model overperformed the CU and the single factor models in most simulation conditions. In conditions of exact and over fitting, the average AE of the bifactor model was almost equal to 0 across all simulation conditions. In other simulation models, the average AE of the bifactor model was lower than 0, which indicates the bifactor model underestimated the group mean difference in these simulation conditions. However, such results were still better than the CU model or the single factor model.

In the perspectives of sample size or amounts of SEV, the accuracy of all three models were not affected by sample size. However, increase of amount of SEV have different impact on three models. For bifactor model, in each simulation model configuration, the range of accuracy was smaller than .018 (maximum in the interaction 2, equal SEV). However, for the CU model, the range of accuracy could be as high as .089 (maximum in the over, equal SEV), .187 for the single-factor model (maximum in the exact, equal SEV). Also, as the amount of SEV increasing, the accuracy of the CU and single factor model become worse.

On exception was interaction 1 model. In which case, the bifactor model had the worst performance. In the condition of interaction 1, the bifactor model highly estimated the group mean difference (average AE = .3), and such estimates were not affected by sample sizes or amounts of SEV. Estimates from the CU model and the single-factor model are slightly lower than the bifactor model (.209 to .255 for CU model; .118 to .251 for single-factor model). Also, as the amounts of SEV increasing, the accuracy also improved.

When the bifactor model did not model all SEVs, it was likely obtaining inaccurate estimates. In the overfitting and interaction 2 conditions, in which 5 items were influenced by SEV in either one or two groups, the average AE of the bifactor model ranged from -.102 to -.82 in both equal and unequal SEV conditions. Also, the average AEs seemed to be not affected by the sample size or the amounts of SEV. Although the inaccuracy in estimation, the bifactor model still overperformed the CU and the single factor model; in the under fitting and the interaction 2 conditions, the average AE of the bifactor model was closer to the 0 than the CU or the single factor models across all simulation conditions.

The precision is measured by width of 95% interval of AE after 2000 times of simulation, the narrower width indicates the better precision. Table 1 and table 2 show that the width of 95% interval which was calculated by the upper end minus the lower end. The precision did not have visible changes across the fitting models, the simulation models, and type of SEVs; the precision only improved as the sample size increasing. The only exception was the condition of interaction 1, in which the widths of intervals were slightly higher than other simulation models.

Assessing Specificity and Sensitivity

The specificity is referred as the true negative rate, or the rate of non-significant results when the true difference is zero. As table 1 and 2 show, the bifactor model still overperformed other models. In the conditions of exact fitting and overfitting, the specificity of the bifactor model could be remaining around .95 across all simulation conditions. However, for CU model, its specificity could reach above .9 only when sample size is small, however, when sample size was becoming larger, the specificity decreased. In other conditions of simulation models, the specificity of the bifactor model failed to remain above .9, but it was still much higher than the CU and the single factor model in both equal and unequal SEV conditions. One exception is the condition of interaction 1, in which single factor model had the best performance in specificity. However, all three models almost got 0 in specificity when sample size is larger than 300.

Also, compared with conditions of equal and unequal SEV, the specificity of the bifactor model does not have considerable differences. However, specificity of the CU model was slightly higher in equal SEV condition than unequal SEV condition.

The sensitivity is referred to true positive rate, or the rate of significant results when there is a difference between two groups. The sensitivity is also related to statistical power. Many researchers agree that the power above .80 is deemed as acceptable (Bezeau & Graves, 2001). Table 1 and 2 shows the bifactor model still overperformed the CU model and the single factor model, while the CU model was slightly better than the single factor model, across all simulation conditions. For the bifactor model, in most simulation models, when sample size was larger than 300, its power reached acceptable level. For the CU model, it required sample size larger than 500 to reach an acceptable power. One exception was interaction 2, in which the bifactor model failed to reach above .8 in most simulation conditions in both equal and unequal SEV. And equal or unequal SEV seem having no effects on sensitivity.

A Self-Esteem Example

This study provided an example of how to select appropriate SEV controlling method in multiple group CFA, based on the information obtained from the simulation study. The data is from a project in which the Rosenberg Self-Esteem Scale (RSE) was administrated among Chinese, Japanese, and American college students. There are totally 844 participants, 280 Americans, 378 Chinese and 186 Japanese.

The RSE was developed by Rosenberg (1965), it is the most widely used instrument to measure the globe self-esteem (Marsh, 1996). So far, the RSE has been translated into many languages and administrated among at least 53 countries or areas (Schmitt & Allik, 2005). The RSE is a Likert scale which consists of 10 items, participants make responses form one (strongly disagree) to five (strongly agree). 5 out of 10 items are positively worded (e.g., 'On the whole, I am satisfied with myself') and the rest is negatively worded (e.g., 'At times, I am no good at all'). Although RES is designed for measuring a unidimensional self-esteem, some studies also reported a two-factorial structure (e.g., Carmines & Zeller, 1979; Marsh, 1996). Majority of researchers agreed on that the multiple-factorial structure is accounted for response style to differently worded items, such response styles are also called positive or negative method effect (e.g., Marsh, 1996; Risko, Oakman & Evan, 2006). Therefore, if the RSE data is collected from one group, researchers can suspect either its half of items is contaminated by one type of SEV, or all items contaminated by two types of SEV.

To determine the factorial structure of each group, 7 models were selected from previous studies to fit the data of each group (e.g., Wu, 2008). Figure 4 presents the figures of these models. Seven models are: 1. original model which assumes no SEV; 2. Bifactor model which assumes the negative method effect exists; 3. Bifactor model which assumes the positive method effect exists; 4. CU model which assumes the negative method effect exists; 5. CU model which assumes the positive method effect exists; 6. Bifactor model assumes both method effects exist; 7. CU model assumes both method effects exist.

Table 3 presents results of all fitting attempts. The original model was used as the baseline model. The chi-square difference test indicated whether the model fitting improvement is significant or not. Results show, for the US and China, after applying any models which specifies only one method effect, the model fitting had significant improvements. Therefore, both positive and negative method effects existed in US and Chinese population (p < .000). For the Japanese data, the model fittings also had significant improvements, but the improvement is relatively small after applying models with positive method effect specified ($\chi^2(6) = 14.16$, p =.038 for CU model, $\chi^2(6) = 19.23$, p = .028 for bifactor model). Also, compared model 4 and model 7, the improvement was not significant anymore ($\chi^2(10) = 14.31$, p = .159). Therefore, positive method effect might slightly contaminate Japanese data.

Based on CFA results, it is safe to conclude that researchers should apply a model with both positive and negative method effects specified. Also, simulation study suggests that the bifactor model is superior to CU model. Therefore, a bifactor model with two method effect factors should be applied. However, this model unable to yield a converged result. By reviewing the CFA results, model was unable to yield converged results in US sample. Therefore, the US sample was discarded, and only Chinese and Japanese samples would be compared. Setting Chinese sample as the baseline and the estimated mean difference is -.806 (*z=-12.222*, *p<.000*). Therefore, the null hypothesis was rejected and there was a significant difference in self-esteem between the Chinese and Japanese people.

Discussion

In CFA, both bifactor and CU models are commonly used to control the systematic error variance, researchers make decision on model choice based on the fit indices: good fit indices indicate that the model has correctly specified the SEV. However, the misspecification and unequal distribution of SEV are issues unique to multiple group CFA. Plus, the mean structure analysis is also a unique objective of multiple group CFA. Therefore, commonly used fit indices (e.g., CFI, SRMR) may not be exclusive criteria for model selection. This study addressed two issues, the first is that how do the bifactor and CU model control the different types of SEVs in multiple group study; the second is that how SEV affects the estimation of group mean difference. Results indicate that the bifactor model has most accurate estimation of the mean difference, if it is able to yield the converged results. When bifactor models encounters the nonconvergent results, an overfitted CU model could be an alternative, though the results would still be biased. Also, whether the amount of SEVs distributes equally would not affect estimation a lot; however, configural noninvariance caused by SEVs and total amount of SEVs are more serious issues

All stimulation models can also be divided into two categories: for exact fitting, over fitting and interaction 1, all SEV are modeled; for under fitting and interaction 2, not all SEVs are modeled. In the first category, the bifactor model perfectly estimated the group mean difference, whereas the CU model is likely to underestimate the group mean difference. Such results indicate that by specifying the SEV as an error factor, the bifactor model could successfully estimate the mean structure of the error factor and partial it out from construct. In the contrary, the CU model also correctly specified the SEVs, the failed to obtain the estimations as accurate as the bifactor models. One explanation is because the CU model is unable to specify the mean structure of SEVs, it is less effective in taking the mean of SEVs out from the construct.

Results from the condition of the interaction 1 seem contradicting to the conclusions that draw above. Comparing with the over fitting, the interaction 1 is also a type of over fitting but yielded worst estimation, while the former one yield best estimations. One explanation is that when SEVs are totally absent, but model has specified SEV, extra parameters which were designed to control SEVs would extract covariance from the construct of interests at the second groups. It can explain that under the interaction 1 simulation model, all models are likely to overestimate the group mean difference. If SEVs are present, regardless its amount, they would serve as a reference so that model would not extract variance from the construct.

In the second category, in which not all SEVs are modeled, models with SEV controlling (bifactor, CU model) overperformed the models without SEV controlling (single factor model). And the bifactor is better than the CU model across all simulation conditions. From conclusions above, it is safe to concluded that the bifactor model have better estimations in group mean difference than the CU model.

These findings are also complementary to previous studies about the methods used to control SEVs. In the framework of covariance structure analysis, Williams and O'Boyle (2015) concluded that correctly specified bifactor model could be able to make expected error near to zero (Conway et al., 2004; Lance et al., 2007; Le et al., 2009; Marsh & Bailey, 1991). This study also indicates the correctly specified or overly specified bifactor model could also obtain accurate parameter estimation in the framework of mean structure analysis.

The major weakness of the bifactor model is non-converged results. Simulation study indicates that probability of nonconvergent results reduces as the amount of SEV or sample size increasing. In applied studies, researchers may not be able to increase the sample size. However, when nonconvergent results occur, it may indicate the SEV they try to model is low in amount. In other words, the ratio of the amount of SEV and the number of degrees of freedom costed in modeling this SEV is relatively low. Eid (2000) named this issue as over factorization and proposed one solution which reduced the number of parameters in fitting model. Eid (2000) and Eid et al. (2003) proposed the CTCM-1 (pronounced as 'CTCM minus one'): if a model have k error factors, only k-1 factors will be specified. However, by doing so, situation of exact or over fitting may be converted to under fitting. According to simulation results of this study, bias will be introduced. Therefore, researchers should also consider exact or over fitting CU model and determine which one could yield most accurate results.

Limitation

This study is not freed of limitations. This study is only simulated simplest SEV conditions, in which there is only one SEV resource and specified as a latent factor. In the stage of model fitting, the best results yielded by the bifactor model may be due to that bifactor model is closer to the true model than the CU model. Also, the SEV in this study actually reflects the narrow definition of method effects, which is proposed by Lance et al (2003). Due to lack of theoretical understanding of SEVs, it is unclear in how to simulate data with CU model, meanwhile, mean structure of SEVs included. According to the board definition of the method effect, the SEV may present in the data in a more complicated way. Following conditions are not simulated in this study: 1. there may be more than one type of SEV exist; 2. Some indicators have factor loadings loaded on more than one error factors; 3. covariance exists among two error factors. When these conditions are introduced, different conclusions about the bifactor and CU models may be obtained.

Direction for Future Research

Future studies can examine whether existed fit indices are able to be used as indicators of accuracy in mean difference estimations. Currently, five fit indices (CFI, SRMR, RMSEA, chi-square, AIC) are most widely used to evaluate the goodness of fit in CFA. Researchers are likely to choose the model with best fit indices because it indicates the model fits the data best. However, Lance et al. (2007) found that true model may not have the best fit indices. Therefore, it worth to know the means of fit indices to the estimation of group mean difference.

Conclusion

The promising findings from this study will help applied researchers to understand the impact of SEV on the measurement invariance test and the properties the SEV controlling methods. This study was found that unmodeled SEVs will alter estimation of latent group mean difference in multiple CFA. It was found that the bifactor model could accurate estimate the group mean difference when it is correctly or overly specified. It was also found that the CU model is less effective in controlling the SEV, therefore, the bifactor model is the preferable SEV controlling method. However, in certain conditions, bifactor model may yield worst estimations among available methods. These findings inform applied researchers in choosing appropriate SEV controlling strategies.

Reference

Bagozzi, R. P. (1984). Expectancy-value attitude models an analysis of critical measurement issues. *International Journal of Research in Marketing*, 1 (4), 295–310. https://doi.org/https://doi.org/10.1016/0167-8116(84)90017-X

Bagozzi, R. P., & Yi, Y. (1990). Assessing method variance in multitrait-multimethod matrices: The case of self-reported affect and perceptions at work. *Journal of Applied Psychology*, 75(5), 547–560. https://doi.org/10.1037/0021-9010.75.5.547

Baumgartner, H., & Steenkamp, J.-B. E. M. (2001). Response Styles in Marketing Research: A Cross-National Investigation. Journal of Marketing Research, 38(2), 143–156. https://doi.org/10.1509/jmkr.38.2.143.18840

Bezeau S, Graves R. Statistical power and effect sizes of clinical neuropsychology research. *J Clin Exp Neuropsychol*. 2001 Jun;23(3):399-406.

Bowen, N. K. & Masa, R. D. (2015). Conducting measurement invariance tests with ordinal data: A guide for social work researchers. *Journal of the Society for Social Work and Research*, 6 (2), 229–249. https://doi.org/10.1086/681607

Brannick, M. T., Chan, D., Conway, J. M., Lance, C. E. & Spector, P. E. (2010). What is method variance and how can we cope with it? a panel discussion. *Organizational Research Methods*, 13 (3), 407–420. https://doi.org/10.1177/1094428109360993

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

Carmines, E. G., & Zeller, R. A. (1979). Reliability and validity assessment. Beverly Hills, CA: Sage.

Chen, F. F., Hayes, A., Carver, C. S., Laurenceau, J. P. & Zhang, Z. (2012). Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*, 80 (1), 219–251. https://doi.org/10.1111/j.1467-6494.2011.00739.x

Cheung, M. W.-l. & Chan, W. (2002). Reducing uniform response bias with ipsative measurement in multiple-group confirmatory factor analysis. *Structural Equation Modeling*, 9 (1), 55–77. https://doi.org/10.1207/S15328007SEM0901_4

Conway, J. M., Lievens, F., Scullen, S. E. & Lance, C. E. (2004). Bias in the correlated uniqueness model for MTMM data. *Structural Equation Modeling*, 11 (4), 535–559. https://doi.org/10.1207/s15328007sem1104_3

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science*, 46, 137–149.

Eden, D., & Leviatan, U. (1975). Implicit Leadership Theory as a Determinant of the Factor Structure Underlying Supervisory Behavior Scales. *Journal of Applied Psychology*, 60, 736-741. http://dx.doi.org/10.1037/0021-9010.60.6.736

Edwards JR. 2008. To prosper, organizational psychology should overcome methodological barriers to progress. *J. Organ. Behav*. 29:469–91

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65, 241–261.

Eid, M., Lischetzke, T., Nussbeck, F. W. & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator ct-c(m-1) model. *Psychological Methods*, 8 (1), 38–60. https://doi.org/10.1037/1082-989X.8.1.38

Fiske, D., Brinberg, D., & Kidder, L. (Eds.) (1982). New directions for methodology of social and behavioral science: Forms of validity in research. San Francisco: Jossey-Bass.

Geiser, C., Burns, G. L. & Servera, M. (2014). Testing for measurement invariance and latent mean differences across methods: Interesting incremental information from multitrait-multimethod studies. *Frontiers in Psychology*, 5 (OCT), 1–19. https://doi.org/10.3389/fpsyg.2014.01216

Golding, S. L. (1977). Method variance, inadequate constructs, or things that go bump in the night*? Multivariate Behavioral Research*, 12 (1), 89–98. https://doi.org/10.1207/s15327906mbr1201_7

Guzzo, R. A., Wagner, D. B., Maguire, E., Herr, B., & Hawley, C. 1986. Implicit theories and the evaluation of group process and performance. *Organizational Behavior and Human Decision Processes*, 37: 279 –295

Harrison, D. A., & McLaughlin, M. E. (1993). Cognitive processes in self-report responses: Tests of item context effects in work attitude measures. Journal of Applied Psychology, 78(1), 129–140. https://doi.org/10.1037/0021-9010.78.1.129

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

Horan, P. M., DiStefano, C. & Motl, R. W. (2003). Wording effects in self-esteem scales: Methodological artifact or response style? *Structural Equation Modeling: A Multidisciplinary Journal*, 10 (3), 435–455. https://doi.org/10.1207/S15328007SEM1003_6

James, L. R. (1980). The unmeasured variables problem in path analysis. *Journal of Applied Psychology*, 65, 415– 421.

Lance, C. E., Dawson, B., Birkelbach, D. & Hoffman, B. J. (2010). Method effects, measurement error, and substantive conclusions. *Organizational Research Methods*, 13 (3), 435–455. https://doi.org/10.1177/1094428109352528

Lance, C. E., Noble, C. L. & Scullen, S. E. (2002). A critique of the correlated trait-correlated method and correlated uniqueness models for multitrait-multimethod data. *Psychological Methods*, 7 (2), 228–244. https://doi.org/10.1037/1082-989X.7.2.228

Lance, C. E., Woehr, D. J. & Meade, A. W. (2007). Case study: A Monte Carlo investigation of assessment center construct validity models. *Organizational Research Methods*, 10 (3), 430–448. https://doi.org/10.1177/1094428106289395

Le, H., Schmidt, F. L. & Putka, D. J. (2009). The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organizational Research Methods*, 12 (1), 165–200. https://doi.org/10.1177/1094428107302900

Lindell, M. K., & Whitney, D. J. (2001). Accounting for common method variance in cross-sectional research designs. *Journal of Applied Psychology*, *86*(1), 114-121. https://doi.org/10.1037/0021-9010.86.1.114

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology*, 70 (4), 810–819. https://doi.org/10.1037/0022-3514.70.4.810

Marsh, H. W. & Bailey, M. (1991). Confirmatory factor analyses of multitrait-multimethod data: a comparison of alternative models. *Applied Psychological Measurement*, 15 (1), 47–70. https://doi.org/10.1177/014662169101500106

Marsh, H. W. & Grayson, D. (1994). Longitudinal stability of latent means and individual differences: A unified approach. *Structural Equation Modeling: A Multidisciplinary Journal*, 1 (4), 317–359. https://doi.org/10.1080/10705519409539984

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology*, 4 (April), 1–13. https://doi.org/10.3389/fpsyg.2013.00169

Messick S. 1991. Psychology and methodology of response styles. *In Improving the Inquiry in Social Science: A Volume in Honor of Lee J. Cronbach*, ed. RE Snow, DE Wiley, pp. 161–200. Hillsdale, NJ: Erlbaum

Milfont, T. & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, ISSN 2011-7922, Vol. 3, Nº. 1, 2010, pags. 111-130, 3. https://doi.org/10.21500/20112084.857

Motl, R. W. & DiStefano, C. (2002). Longitudinal invariance of self-esteem and method effects associated with negatively worded items. *Structural Equation Modeling: A Multidisciplinary Journal*, 9 (4), 562–578. https://doi.org/10.1207/S15328007SEM0904_6

Podsakoff, P. M., MacKenzie, S. B., Lee, J. Y. & Podsakoff, N. P. (2003). Common Method Biases in Behavioral Research: A Critical Review of the Literature and Recommended Remedies. *Journal of Applied Psychology*, 88 (5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Pohl, S. & Steyer, R. (2010). Modeling common traits and method effects in multitrait-multimethod analysis. *Multivariate Behavioral Research*, 45 (1), 45–72. https://doi.org/10.1080/00273170903504729

Quilty, L. C., Oakman, J. M., & Risko, E. (2006). Correlates of the Rosen- berg Self-Esteem Scale method effects. *Structural Equation Modeling*, 13, 99–117.

Quilty, L. C., Oakman, J. M. & Risko, E. (2009). Structural equation modeling: A correlates of the Rosenberg self-esteem scale method effects. *Psychology*, (May 2012), 37–41.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47 (5), 667–696. https://doi.org/10.1080/00273171.2012.715555

Richardson, H. A., Simmering, M. J. & Sturman, M. C. (2009). a tale of three perspectives. *Organizational Research Methods*, 12 (4), 762–800. https://doi.org/10.1177/1094428109332834

Salancik, G. R., & Pfeffer, J. (1977). Who gets power--and how they hold on to it: A strategic-contingency model of power. *Organizational Dynamics*, 5(3), 2–21. https://doi.org/10.1016/0090-2616(77)90028-6

Schmitt, D. P. & Allik, J. (2005). Simultaneous administration of the Rosenberg self-esteem scale in 53 nations: Exploring the universal and culture-specific features of global self-esteem. *Journal of Personality and Social Psychology*, 89 (4), 623–642. https://doi.org/10.1037/0022-3514.89.4.623

Siemsen, E., Roth, A. & Oliveira, P. (2010). Common method bias in regression models with linear, quadratic, and interaction effects. *Organizational Research Methods*, 13 (3), 456–476. https://doi.org/10.1177/1094428109351241

Spector, P. E. (1987). Method variance as an artifact in self-reported affect and perceptions at work: Myth or significant problem? *Journal of Applied Psychology*, 72 (3), 438–443. https://doi.org/10.1037/0021-9010.72.3.438

Spector, P. E. (1994). Using self-report questionnaires in OB research: A comment on the use of a controversial method, *Journal of Organizational Behavior*, 15, 385-392.

Spector, P. E. and Brannick, M. T., (1995). The Nature and Effects of Method Variance in Organizational Research. *Psychology Faculty Publications*. 568. https://digitalcommons.usf.edu/psy_facpub/568

Spector, P. E. (2006). Method Variance in Organizational Research: Truth or Urban Legend? *Organizational Research Methods*, 9(2), 221–232. https://doi.org/10.1177/1094428105284955

van de Schoot, R., Lugtig, P. & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, 9 (4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Thomas, K. W., and Kilmann, R. H. (1975) The Social Desirability Variable in Organizational Research: An Alternative Explanation for Reported Findings. *Academy of Management Journal*, 18(4), 471-482.

Thompson, M. S., Green, S. B. (2006). Evaluating between-group differences in latent variable means. In Hancock, G. R., Mueller, R. O. (Eds.), *Structural Equation Modeling: A Second Course* (pp. 119-169). Greenwich, CT: Information Age.

Weijters, B., Cabooter, E. & Schillewaert, N. (2010). The effect of rating scale format on response styles: The number of response categories and response category labels. *International Journal of Research in Marketing*, 27 (3), 236–247. https://doi.org/https://doi.org/10.1016/j.ijresmar.2010.02.004

Williams, L. J., Cote, J. A., & Buckley, M. R. (1989). Lack of method variance in self-reported affect and perceptions at work: Reality or artifact? Journal of Applied Psychology, 74(3), 462–468. https://doi.org/10.1037/0021-9010.74.3.462

Williams, L. J. & O'Boyle, E. H. (2015). Ideal, nonideal, and no-marker variables: The confirmatory factor analysis (CFA) marker technique works when it matters. *Journal of Applied Psychology*, 100 (5), 1579–1602. https://doi.org/10.1037/a0038855

Wang, J., Siegal, H. A., Falck, R. S., & Carlson, R. G. (2001). Factorial structure of Rosenberg's self-esteem scale among crack-cocaine drug users. Structural Equation Modeling, 8, 275–286.

Wu, C. H. (2008). An examination of the wording effect in the Rosenberg self-esteem scale among culturally Chinese people. Journal of Social Psychology, 148 (5), 535–552. https://doi.org/10.3200/SOCP.148.5.535-55

Table 1a

Results of three models under the equal SEV condition

			Bifactor Model				
Simulation models	FL	*n*	Convergent rate	Average AE	Width of Interval	Specificity	Sensitivity
Exact	.2	100	1065	.001	.498	.948	.386
		300	1426	.002	.284	.952	.817
		500	1607	.000	.225	.939	.939
		1000	1808	.001	.149	.958	.999
	.3	100	1639	.002	.464	.950	.369
		300	1954	-.004	.288	.926	.782
		500	1991	.001	.214	.960	.954
		1000	1999	.000	.162	.929	.998
	.4	100	1974	.003	.465	.959	.397
		300	2000	.002	.287	.932	.839
		500	2000	.001	.220	.946	.966
		1000	2000	.000	.146	.961	1.000
	.5	100	1999	-.003	.467	.959	.360
		300	2000	.000	.272	.957	.823
		500	2000	.000	.208	.959	.956
		1000	2000	.000	.148	.957	.999
Under	.2	100	1033	-.088	.498	.896	.148
		300	1069	-.092	.287	.761	.290
		500	1015	-.093	.209	.581	.526
		1000	1049	-.089	.160	.370	.799
	.3	100	1102	-.097	.513	.885	.130
		300	1217	-.097	.288	.738	.282
		500	1278	-.096	.240	.604	.475
		1000	1347	-.100	.162	.285	.729
	.4	100	1370	-.103	.544	.839	.126
		300	1559	-.104	.322	.706	.283
		500	1695	-.101	.244	.578	.449
		1000	1848	-.103	.176	.277	.710
	.5	100	1833	-.104	.561	.864	.117
		300	1987	-.107	.327	.657	.270
		500	1998	-.106	.255	.510	.404
		1000	2000	-.106	.179	.243	.681
Over	.2	100	1077	.001	.473	.957	.369
		300	1170	.002	.285	.940	.816
		500	1204	.001	.214	.954	.946
		1000	1216	.000	.140	.960	1.000
	.3	100	1202	.001	.504	.941	.365
		300	1343	-.001	.283	.945	.816
		500	1329	-.001	.208	.953	.938

			Bifactor Model				
Simulation models	FL	*n*	Convergent Results	Average AE	Width of Interval	Specificity	Sensitivity
Over	.3	1000	1341	.000	.153	.942	1.000
	.4	100	1334	-.006	.489	.949	.359
		300	1389	-.002	.271	.954	.801
		500	1400	.002	.210	.955	.955
		1000	1421	.000	.149	.956	1.000
	.5	100	1526	.003	.483	.944	.388
		300	1545	.001	.283	.945	.815
		500	1560	.000	.210	.946	.967
		1000	1591	.001	.146	.947	1.000
Interaction 1	.2	100	1051	.300	.595	.095	.920
		300	1280	.300	.430	.000	1.000
		500	1457	.304	.385	.000	1.000
		1000	1728	.300	.337	.000	1.000
	.3	100	1471	.295	.612	.104	.886
		300	1881	.300	.441	.000	1.000
		500	1949	.298	.390	.000	1.000
		1000	1997	.301	.330	.000	1.000
	.4	100	1824	.307	.617	.091	.910
		300	1981	.299	.438	.000	1.000
		500	1994	.302	.384	.000	1.000
		1000	2000	.300	.328	.000	1.000
	.5	100	1914	.306	.622	.085	.904
		300	2000	.298	.444	.000	1.000
		500	2000	.299	.397	.000	1.000
		1000	2000	.300	.329	.000	1.000
Interaction 2	.2	100	929	-.090	.506	.866	.165
		300	948	-.087	.277	.778	.323
		500	1050	-.086	.226	.617	.541
		1000	1178	-.081	.153	.433	.851
	.3	100	1071	-.088	.477	.885	.160
		300	1181	-.088	.289	.723	.355
		500	1165	-.086	.226	.638	.531
		1000	1150	-.085	.149	.392	.864
	.4	100	1400	-.082	.493	.885	.150
		300	1674	-.087	.271	.772	.358
		500	1808	-.088	.229	.612	.536
		1000	1933	-.086	.158	.405	.831
	.5	100	1887	-.091	.507	.872	.143
		300	1994	-.087	.299	.773	.335
		500	2000	-.088	.228	.631	.533
		1000	2000	-.087	.162	.384	.799

Note. FL= the average factor loadings which loaded on the error factor in simulation model of both groups; Average AE = average absolute error; Width of interval = width of 95% interval

Table 1b

Results of three models under the equal SEV condition

			Correlated Uniqueness Models				
Simulation models	FL	*n*	Convergent rate	Average AE	Width of Interval	Specificity	Sensitivity
Exact	.2	100	2000	-.043	.465	.942	.270
		300	2000	-.044	.266	.903	.623
		500	2000	-.045	.213	.858	.833
		1000	2000	-.043	.149	.805	.981
	.3	100	2000	-.060	.459	.931	.222
		300	2000	-.062	.276	.828	.528
		500	2000	-.057	.210	.813	.780
		1000	2000	-.058	.156	.659	.955
	.4	100	2000	-.065	.464	.914	.227
		300	2000	-.064	.280	.821	.523
		500	2000	-.065	.221	.750	.731
		1000	2000	-.066	.143	.581	.946
	.5	100	2000	-.071	.453	.922	.210
		300	2000	-.067	.265	.833	.517
		500	2000	-.067	.202	.778	.698
		1000	2000	-.067	.148	.573	.946
Under	.2	100	2000	-.100	.487	.857	.124
		300	2000	-.099	.283	.712	.281
		500	2000	-.101	.215	.532	.482
		1000	2000	-.098	.153	.273	.749
	.3	100	2000	-.132	.511	.820	.089
		300	2000	-.131	.284	.569	.186
		500	2000	-.131	.233	.369	.240
		1000	2000	-.132	.156	.086	.423
	.4	100	2000	-.164	.512	.741	.061
		300	2000	-.162	.312	.412	.096
		500	2000	-.160	.238	.198	.135
		1000	2000	-.161	.168	.017	.205
	.5	100	2000	-.188	.559	.696	.059
		300	2000	-.189	.315	.298	.060
		500	2000	-.188	.249	.099	.070
		1000	2000	-.188	.175	.006	.079
Over	.2	100	2000	-.033	.462	.945	.308
		300	2000	-.028	.271	.932	.715
		500	2000	-.029	.209	.911	.900
		1000	2000	-.031	.144	.860	.996
	.3	100	2000	-.046	.482	.930	.267
		300	2000	-.044	.268	.886	.623
		500	2000	-.044	.204	.877	.830

Simulation models	FL	*n*	Correlated Uniqueness Model Convergent Results	Average AE	Width of Interval	Specificity	Sensitivity
Over	.3	1000	2000	-.042	.154	.803	.989
	.4	100	2000	-.055	.474	.931	.245
		300	2000	-.051	.263	.874	.612
		500	2000	-.049	.205	.836	.829
		1000	2000	-.051	.146	.729	.985
	.5	100	2000	-.052	.466	.933	.250
		300	2000	-.055	.276	.856	.572
		500	2000	-.056	.207	.817	.785
		1000	2000	-.055	.147	.695	.983
Interaction 1	.2	100	2000	.255	.595	.142	.833
		300	2000	.256	.426	.000	.999
		500	2000	.254	.374	.000	1.000
		1000	2000	.255	.326	.000	1.000
	.3	100	2000	.232	.588	.191	.793
		300	2000	.233	.434	.001	.998
		500	2000	.234	.381	.000	1.000
		1000	2000	.236	.323	.000	1.000
	.4	100	2000	.225	.600	.222	.786
		300	2000	.220	.421	.003	1.000
		500	2000	.222	.371	.000	1.000
		1000	2000	.220	.321	.000	1.000
	.5	100	2000	.213	.573	.232	.756
		300	2000	.209	.428	.002	.999
		500	2000	.209	.387	.000	1.000
		1000	2000	.211	.319	.000	1.000
Interaction 2	.2	100	2000	-.097	.477	.865	.132
		300	2000	-.096	.269	.711	.314
		500	2000	-.097	.211	.552	.478
		1000	2000	-.095	.151	.277	.769
	.3	100	2000	-.119	.454	.841	.101
		300	2000	-.119	.272	.579	.235
		500	2000	-.120	.215	.412	.313
		1000	2000	-.119	.151	.117	.552
	.4	100	2000	-.133	.477	.786	.085
		300	2000	-.138	.274	.508	.138
		500	2000	-.139	.216	.289	.221
		1000	2000	-.138	.151	.052	.366
	.5	100	2000	-.156	.483	.746	.066
		300	2000	-.153	.292	.463	.102
		500	2000	-.154	.228	.220	.146
		1000	2000	-.153	.158	.026	.243

Note. FL= the average factor loadings which loaded on the error factor in simulation model of both groups; Average AE = average absolute error; Width of interval = width of 95% interval

Table 1c

Results of three models under the equal SEV condition

			Single Factor Model				
Simulation models	FL	*n*	Convergent rate	Average AE	Width of Interval	Specificity	Sensitivity
Exact	.2	100	2000	-.053	.469	.928	.243
		300	2000	-.054	.267	.874	.559
		500	2000	-.055	.214	.814	.786
		1000	2000	-.053	.151	.715	.970
	.3	100	2000	-.093	.477	.887	.167
		300	2000	-.095	.277	.721	.319
		500	2000	-.091	.211	.595	.544
		1000	2000	-.092	.158	.333	.802
	.4	100	2000	-.149	.479	.767	.076
		300	2000	-.149	.289	.431	.128
		500	2000	-.148	.230	.230	.183
		1000	2000	-.149	.153	.027	.280
	.5	100	2000	-.242	.491	.543	.067
		300	2000	-.238	.277	.092	.090
		500	2000	-.239	.220	.013	.113
		1000	2000	-.240	.158	.000	.186
Under	.2	100	2000	-.100	.486	.854	.122
		300	2000	-.099	.278	.710	.283
		500	2000	-.101	.213	.534	.481
		1000	2000	-.098	.153	.268	.753
	.3	100	2000	-.134	.509	.821	.086
		300	2000	-.133	.284	.559	.178
		500	2000	-.133	.234	.352	.232
		1000	2000	-.134	.158	.076	.403
	.4	100	2000	-.183	.520	.712	.054
		300	2000	-.180	.314	.347	.071
		500	2000	-.178	.245	.131	.085
		1000	2000	-.179	.171	.005	.100
	.5	100	2000	-.260	.555	.530	.067
		300	2000	-.262	.316	.085	.101
		500	2000	-.261	.252	.011	.143
		1000	2000	-.261	.176	.000	.221
Over	.2	100	2000	-.037	.467	.941	.295
		300	2000	-.031	.276	.921	.706
		500	2000	-.032	.211	.908	.883
		1000	2000	-.034	.145	.836	.994
	.3	100	2000	-.059	.491	.921	.227
		300	2000	-.057	.278	.845	.552
		500	2000	-.057	.209	.812	.769

Simulation models	FL	*n*	Single Factor Model Convergent Results	Average AE	Width of Interval	Specificity	Sensitivity
Over	.3	1000	2000	-.055	.155	.684	.973
	.4	100	2000	-.089	.496	.886	.169
		300	2000	-.086	.279	.747	.412
		500	2000	-.084	.219	.627	.589
		1000	2000	-.086	.153	.349	.867
	.5	100	2000	-.159	.538	.760	.068
		300	2000	-.164	.323	.396	.114
		500	2000	-.166	.239	.169	.105
		1000	2000	-.165	.170	.015	.201
Interaction 1	.2	100	2000	.251	.603	.158	.820
		300	2000	.252	.427	.000	.998
		500	2000	.250	.374	.000	1.000
		1000	2000	.251	.325	.000	1.000
	.3	100	2000	.217	.591	.240	.746
		300	2000	.219	.436	.003	.997
		500	2000	.220	.381	.000	1.000
		1000	2000	.222	.326	.000	1.000
	.4	100	2000	.188	.617	.352	.648
		300	2000	.183	.435	.025	.977
		500	2000	.184	.379	.000	1.000
		1000	2000	.183	.325	.000	1.000
	.5	100	2000	.122	.588	.591	.437
		300	2000	.118	.440	.141	.848
		500	2000	.118	.390	.023	.969
		1000	2000	.120	.326	.000	1.000
Interaction 2	.2	100	2000	-.098	.476	.868	.132
		300	2000	-.097	.268	.708	.307
		500	2000	-.098	.211	.547	.463
		1000	2000	-.096	.150	.267	.756
	.3	100	2000	-.129	.458	.826	.086
		300	2000	-.128	.275	.537	.200
		500	2000	-.129	.218	.348	.245
		1000	2000	-.128	.152	.078	.460
	.4	100	2000	-.167	.491	.718	.059
		300	2000	-.172	.275	.343	.074
		500	2000	-.173	.220	.132	.072
		1000	2000	-.172	.149	.008	.099
	.5	100	2000	-.248	.497	.512	.089
		300	2000	-.245	.296	.106	.110
		500	2000	-.247	.226	.008	.159
		1000	2000	-.246	.163	.000	.259

Note. FL= the average factor loadings which loaded on the error factor in simulation model of both groups; Average AE = average absolute error; Width of interval = width of 95% interval

Table 2a

Results of three models under the unequal SEV condition

			Bifactor Model				
Simulation models	FL	*n*	Convergent rate	Average AE	Width of Interval	Specificity	Sensitivity
Exact	.3	100	1487	.004	.484	.947	.391
		300	1911	-.002	.282	.946	.800
		500	1977	-.002	.217	.945	.943
		1000	2000	-.002	.152	.948	.999
	.4	100	1815	-.001	.489	.953	.396
		300	1991	-.003	.288	.950	.779
		500	2000	-.003	.211	.960	.944
		1000	2000	-.001	.158	.941	.997
	.5	100	1939	.000	.474	.955	.367
		300	1997	.000	.278	.954	.800
		500	2000	-.002	.213	.955	.945
		1000	2000	-.003	.156	.941	1.000
Under	.3	100	1110	-.089	.543	.853	.138
		300	1198	-.102	.314	.650	.290
		500	1237	-.094	.235	.583	.481
		1000	1314	-.095	.166	.342	.759
	.4	100	1417	-.098	.515	.879	.119
		300	1661	-.098	.302	.741	.311
		500	1782	-.098	.238	.555	.453
		1000	1939	-.100	.167	.307	.727
	.5	100	1844	-.100	.541	.862	.123
		300	1989	-.098	.320	.755	.292
		500	1997	-.099	.246	.581	.444
		1000	2000	-.098	.168	.291	.736
Over	.3	100	1163	.007	.489	.952	.387
		300	1295	-.001	.281	.954	.782
		500	1326	.000	.220	.943	.961
		1000	1333	.000	.150	.950	.999
	.4	100	1358	.003	.488	.947	.378
		300	1391	.003	.284	.959	.814
		500	1436	-.001	.213	.949	.963
		1000	1391	.000	.155	.934	1.000
	.5	100	1521	.008	.518	.928	.378
		300	1567	-.002	.287	.929	.799
		500	1570	.003	.214	.951	.964
		1000	1592	.000	.153	.942	.999
Interaction 2	.3	100	923	-.082	.492	.908	.189
		300	977	-.089	.286	.767	.369
		500	1017	-.090	.217	.645	.514

Simulation models	FL	*n*	Bifactor Model: Convergent Results	Average AE	Width of Interval	Specificity	Sensitivity
Interaction 2	.3	1000	1127	-.090	.150	.336	.823
	.4	100	1196	-.088	.491	.882	.157
		300	1563	-.093	.288	.719	.367
		500	1738	-.091	.228	.624	.484
		1000	1917	-.094	.163	.345	.775
	.5	100	1695	-.088	.491	.881	.146
		300	1944	-.097	.299	.718	.303
		500	1979	-.095	.232	.599	.485
		1000	2000	-.095	.153	.292	.779

Note. FL= the difference of the average factor loadings which loaded on the error factor in simulation model of two groups; Average AE = average absolute error; Width of interval = width of 95% interval

Table 2b

Results of three models under the unequal SEV condition

			Correlated Uniqueness Model				
Simulation models	FL	*n*	Convergent rate	Average AE	Width of Interval	Specificity	Sensitivity
Exact	.3	100	2000	-.059	.470	.916	.239
		300	2000	-.062	.281	.846	.543
		500	2000	-.061	.215	.788	.764
		1000	2000	-.061	.150	.613	.960
	.4	100	2000	-.075	.468	.902	.203
		300	2000	-.076	.275	.809	.432
		500	2000	-.075	.205	.683	.674
		1000	2000	-.073	.153	.486	.915
	.5	100	2000	-.083	.474	.891	.185
		300	2000	-.082	.274	.776	.405
		500	2000	-.084	.208	.664	.599
		1000	2000	-.085	.153	.402	.864
Under	.3	100	2000	-.127	.524	.808	.103
		300	2000	-.132	.295	.534	.167
		500	2000	-.128	.227	.344	.277
		1000	2000	-.128	.152	.089	.473
	.4	100	2000	-.157	.503	.751	.077
		300	2000	-.156	.289	.420	.112
		500	2000	-.157	.230	.184	.151
		1000	2000	-.159	.164	.014	.201
	.5	100	2000	-.182	.513	.685	.051
		300	2000	-.181	.308	.332	.066
		500	2000	-.182	.237	.100	.081
		1000	2000	-.182	.165	.002	.103
Over	.3	100	2000	-.041	.463	.940	.276
		300	2000	-.045	.265	.901	.619
		500	2000	-.044	.211	.850	.856
		1000	2000	-.045	.151	.782	.983
	.4	100	2000	-.056	.479	.917	.240
		300	2000	-.052	.270	.879	.557
		500	2000	-.056	.212	.828	.786
		1000	2000	-.055	.150	.679	.973
	.5	100	2000	-.060	.480	.895	.230
		300	2000	-.066	.276	.831	.512
		500	2000	-.062	.211	.790	.762
		1000	2000	-.063	.145	.621	.958
Interaction 2	.3	100	2000	-.116	.479	.841	.119
		300	2000	-.120	.268	.561	.216
		500	2000	-.121	.215	.410	.313

			Correlated Uniqueness Model				
Simulation models	FL	*n*	Convergent Results	Average AE	Width of Interval	Specificity	Sensitivity
Interaction 2	.3	1000	2000	-.122	.153	.097	.531
	.4	100	2000	-.144	.477	.755	.072
		300	2000	-.147	.285	.441	.142
		500	2000	-.145	.217	.251	.198
		1000	2000	-.146	.153	.038	.306
	.5	100	2000	-.163	.469	.724	.057
		300	2000	-.171	.281	.316	.084
		500	2000	-.168	.227	.138	.097
		1000	2000	-.169	.147	.008	.142

Note. FL= the difference of the average factor loadings which loaded on the error factor in simulation model of two groups; Average AE = average absolute error; Width of interval = width of 95% interval

Table 2c

Results of three models under the unequal SEV condition

			Single Factor Model				
Simulation models	FL	*n*	Convergent rate	Average AE	Width of Interval	Specificity	Sensitivity
Exact	.3	100	2000	-.083	.478	.878	.177
		300	2000	-.086	.282	.759	.397
		500	2000	-.085	.219	.630	.585
		1000	2000	-.086	.150	.353	.860
	.4	100	2000	-.127	.482	.816	.112
		300	2000	-.130	.282	.543	.180
		500	2000	-.128	.212	.350	.268
		1000	2000	-.126	.156	.083	.514
	.5	100	2000	-.186	.471	.703	.050
		300	2000	-.186	.284	.274	.052
		500	2000	-.188	.213	.073	.054
		1000	2000	-.188	.157	.006	.065
Under	.3	100	2000	-.127	.522	.804	.100
		300	2000	-.133	.295	.539	.168
		500	2000	-.129	.229	.338	.266
		1000	2000	-.129	.153	.089	.474
	.4	100	2000	-.165	.506	.745	.067
		300	2000	-.164	.293	.383	.091
		500	2000	-.165	.231	.164	.123
		1000	2000	-.167	.164	.009	.158
	.5	100	2000	-.214	.524	.626	.038
		300	2000	-.213	.314	.226	.055
		500	2000	-.214	.243	.048	.047
		1000	2000	-.215	.169	.001	.038
Over	.3	100	2000	-.050	.473	.938	.242
		300	2000	-.054	.267	.868	.567
		500	2000	-.053	.216	.806	.799
		1000	2000	-.054	.153	.702	.970
	.4	100	2000	-.078	.496	.891	.182
		300	2000	-.074	.282	.810	.446
		500	2000	-.078	.214	.704	.633
		1000	2000	-.076	.158	.467	.909
	.5	100	2000	-.111	.515	.818	.144
		300	2000	-.118	.300	.594	.245
		500	2000	-.114	.234	.419	.400
		1000	2000	-.115	.162	.154	.633
Interaction 2	.3	100	2000	-.120	.479	.832	.116
		300	2000	-.125	.266	.534	.199
		500	2000	-.125	.213	.384	.284

			Single Factor Model				
Simulation models	FL	*n*	Convergent Results	Average AE	Width of Interval	Specificity	Sensitivity
Interaction 2	.3	1000	2000	-.127	.152	.077	.487
	.4	100	2000	-.157	.476	.729	.055
		300	2000	-.160	.283	.377	.094
		500	2000	-.158	.216	.171	.129
		1000	2000	-.160	.151	.017	.186
	.5	100	2000	-.198	.465	.663	.046
		300	2000	-.207	.287	.172	.048
		500	2000	-.203	.223	.050	.053
		1000	2000	-.205	.147	.002	.041

Note. FL= the difference of the average factor loadings which loaded on the error factor in simulation model of two groups; Average AE = average absolute error; Width of interval = width of 95% interval

Table 3

CFA results after using seven models to fit the data of three groups separately

	χ^2	df	$\Delta\chi^2(df, p)$	SRMR	RMSEA	CFI
US						
1	169.19	35	NA	.06	.12	.89
2	62.29	29	106.9(6, <.000)	.03	.06	.97
3	103.07	29	66.12(6, <.000)	.05	.10	.94
4	55.46	25	113.73(10,<.000)	.03	.07	.97
5	86.67	25	82.52(10, <.000)	.05	.09	.95
6	NA	NA	NA	NA	NA	NA
7	33.94	15	21.52(10, <.000)*/ 52.73(10, <.000)**	.03	.07	.98
CN						
1	229.94	35	NA	.07	.12	.83
2	123.39	29	106.55(6,<.000)	.05	.09	.92
3	134.13	29	95.81(6, <.000)	.05	.10	.91
4	96.28	25	133.66(10, <.000)	.04	.09	.94
5	114.35	25	115.59(10, <.000)	.05	.10	.92
6	62.34	23	61.05(6, <.000)*/ 71.79(6, <.000)**	.03	.07	.97
7	34.12	15	62.16(10, <.000)*/ 80.22(10, <.000)**	.02	.06	.98
JP						
1	129.55	35	NA	.08	.12	.85
2	64.05	29	65.5(6, <.000)	.05	.08	.94
3	115.39	29	14.16(6,.028)	.08	.13	.86
4	47.62	25	81.93(10, <.000)	.05	.07	.96
5	110.32	25	19.23(10, .038)	.07	.14	.86
6	47.20	23	17.3(6, .008)*/ 68.19(6, <.000)**	.04	.08	.96
7	33.31	15	14.31(10, .159)*/ 77.01(10, <.000)**	.04	.08	.97

Note. 1 = model with no method effect; 2 = bifactor model with negative method effect; 3 = bifactor model with positive method effect; 4 = CU model with negative method effect; 5 = CU model with positive method effect.

* Comparing with the model specifies negative method effect
** Comparing with the model specified positive method effect

Figure 1

An example of ULMC (Bifactor) model

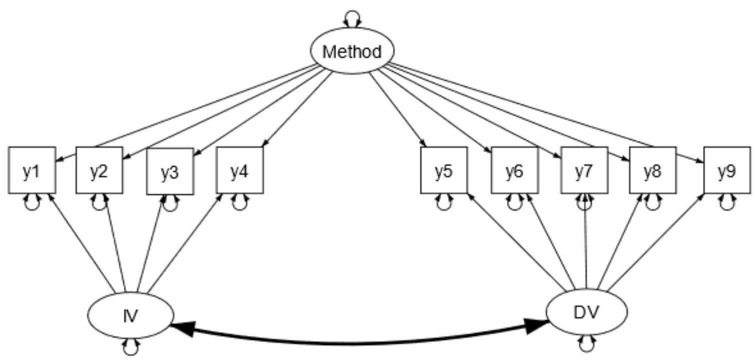
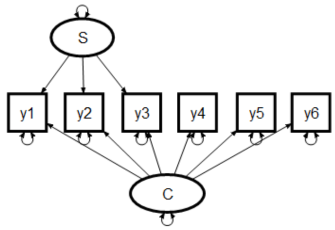
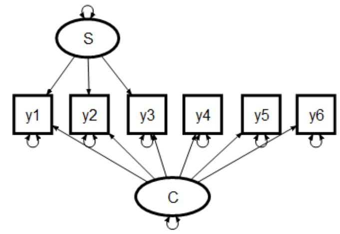
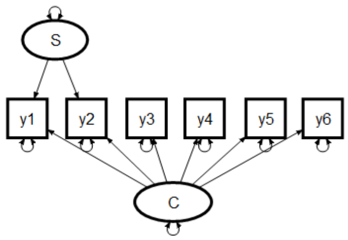
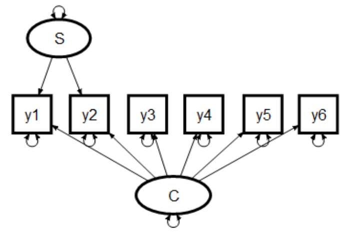
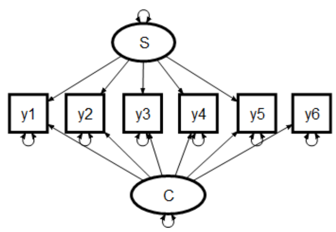
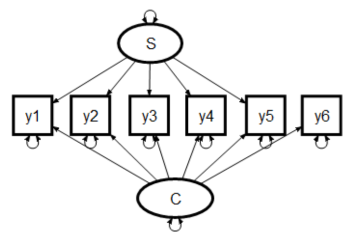
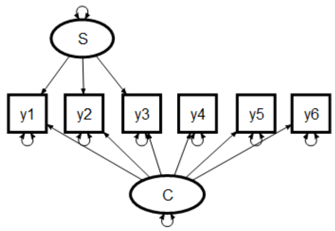
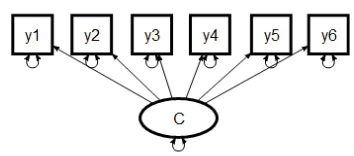
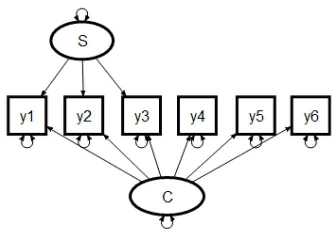
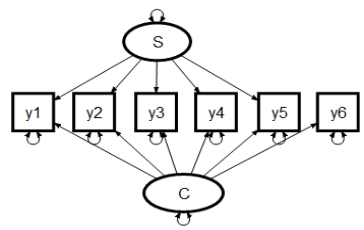


Figure 2

The configurations of models for data generation

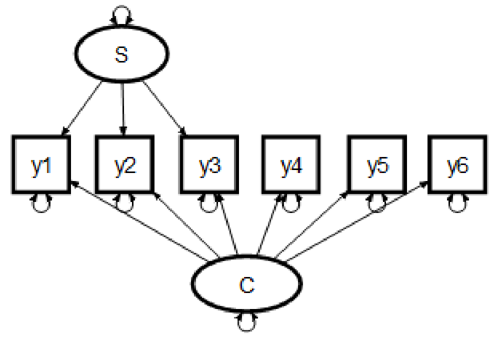
	Group1	Group 2
Exact Fitting		
Over Fitting		
Under Fitting		
Interaction 1		
Interaction 2		

Note. S factor represents SEV, C factor represents construct.

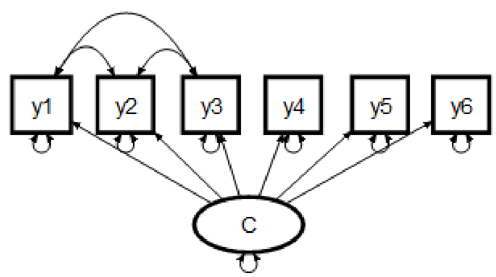
Figure 3

The configurations of models used to fit data

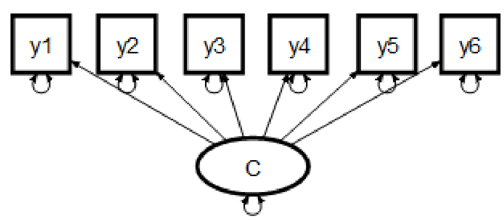
Bifactor



CU



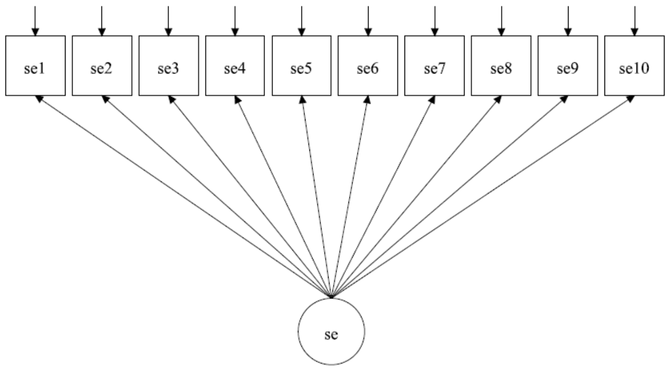
Single Factor



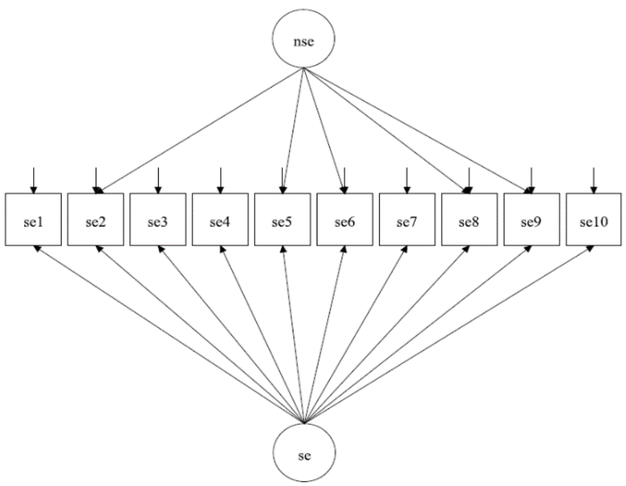
Note. S factor represents SEV, C factor represents construct.

Figure 4:

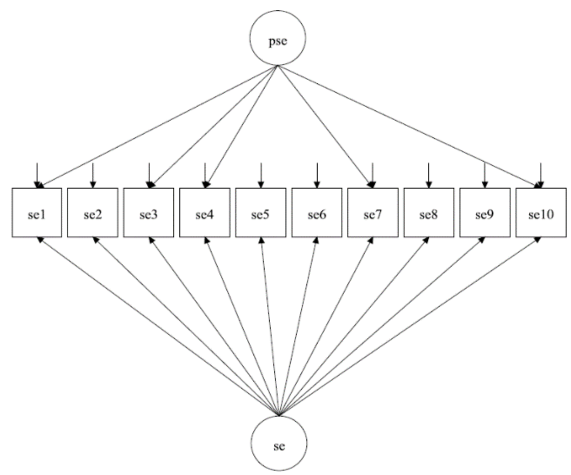
Theoretical Models for the Rosenberg Self-Esteem Sale



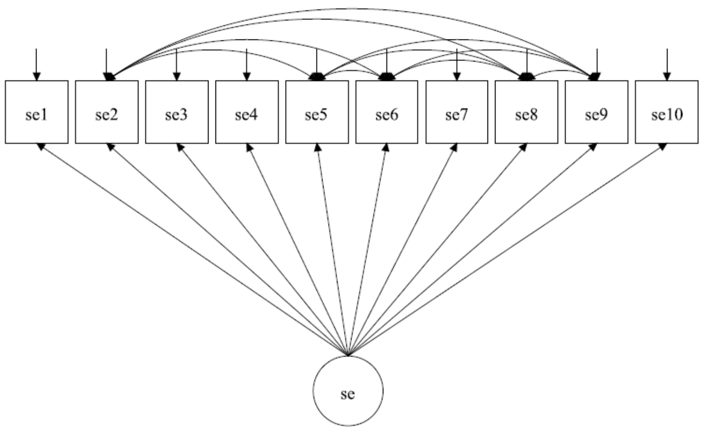
Model 1: Single Factor Model



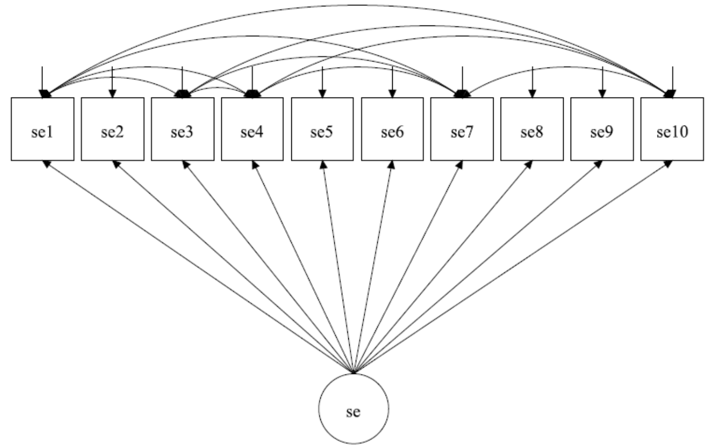
Model 2: Bifactor Model with Negative Method Factor



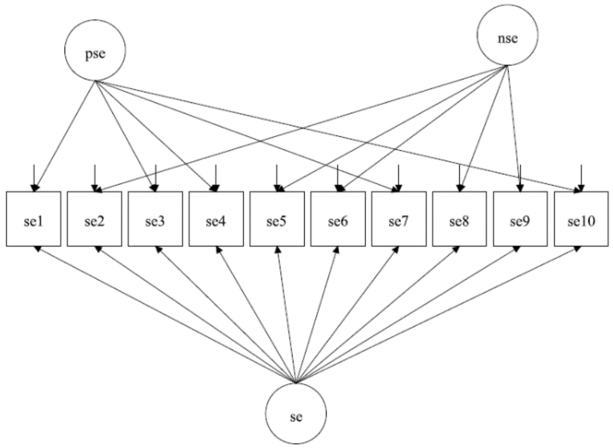
Model 3: Bifactor Model with Positive Method Factor



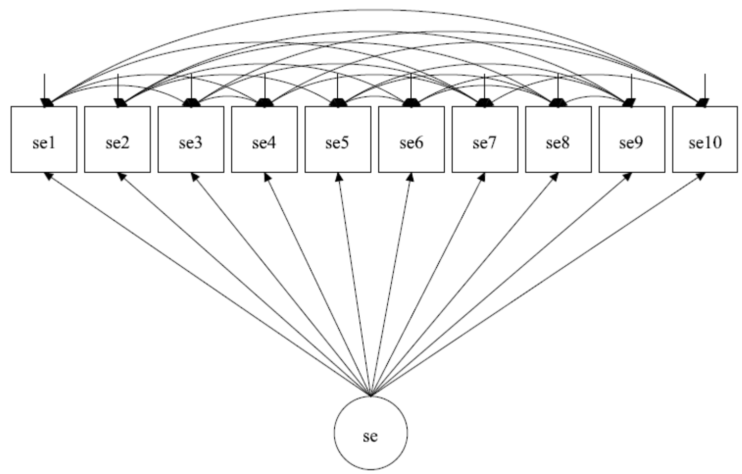
Model 4: CU Model with Negative Method Factor



Model 5: CU Model with Positive Method Factor



Model 6: Bifactor Model with Two Method Factors



Model 7: CU Model with Two Method Factors