IGNORING MULTILEVEL DATA STRUCTURE IN COMFIRMATORY FACTOR ANALYSIS WITH ORDINAL ITEMS

by

Li Zhou

B.A., East China Normal University, 2005

M.A., University of Pittsburgh, 2008

M.A., University of Pittsburgh, 2014

Submitted to the Graduate Faculty of

School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2016

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Li Zhou

It was defended on

November 14, 2016

and approved by

Clement A. Stone, Ph.D., Professor, Department of Psychology in Education

Lan Yu, Ph.D., Associate Professor, School of Medicine

Lauren Terhorst, Ph.D., Associate Professor, Department of Occupational Therapy

Dissertation Advisor: Feifei Ye, Ph.D., Assistant Professor, Department of Psychology in Education

IGNORING MULTILEVEL DATA STRUCTURE IN COMFIRMATORY FACTOR ANALYSIS WITH ORDINAL ITEMS

Li Zhou, PhD

University of Pittsburgh, 2016

This study used the Monte Carlo method to compare multilevel and single-level models in confirmatory factor analysis (CFA) of ordinal items with clustered data. Specifically, model fit indices, estimates of factor loading and standard error were compared among three models, two-level CFA, single-level CFA with adjusted standard error, and single-level CFA with normal standard error. Two different factorial structures were considered, 2 factors at both the within- and between-level (W2B2) and 2 factors at the within-level and 1 factor at the between-level (W2B1).

All model fit indices indicated that the two-level CFA model fitted the clustered data well. The model fit of the two-level CFA model was better than that of the single-level CFA model with adjusted standard error, which was better than that of the normal single-level CFA model. Chi-square *p* value and RMSEA were not as sensitive as CFI and TLI in the small sample size to the misspecification of factorial structure. When factor loadings across levels were the same in the true model, factor loadings estimated from the single-level models were acceptable. The standard error of the within-level factor loading estimated by the normal model was significantly smaller than the complex model, which was smaller than the two-level model,

suggesting that standard errors are underestimated when the single-level model is used to estimate the two-level data. The effect of design factors on the relative bias of the factor loading and standard errors between W2B2 model and W2B1 model were similar in most conditions.

These results suggest applied researchers consider the interest of the study first when selecting CFA models of clustered data. The single-level CFA with adjusted standard error is preferred when the interest of the study is at the individual level, while multilevel CFA is recommended when the interest is at the cluster level. However, in either case, the recommendation is to compare both models to prevent the spurious clustering effect. If there truly exists a multilevel data structure, standard errors estimated from the two-level CFA model are expected to be significantly larger than adjusted standard errors in the single-level model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENT

I would like to take this opportunity to express my sincere gratitude to all who encouraged and supported me to complete this dissertation and my doctoral study.

First, I would like to extend my greatest appreciation to my dissertation advisor, Dr . Feifei Ye, for her guidance and support throughout my doctoral study. I am thankful to Dr Ye, for her feedback and valuable advice to all of my questions at every stage of the dissertation. Dr Ye gave me lots of advice, instruction, and guidance in developing my knowledge and research skills in statistics and writing the dissertation. It was due to her paramount mentorship that I was able to complete this dissertation.

Second, I would like to thank other members in my dissertation committee, Dr. Clem Stone, Dr. Lauren Terhorst, and Dr. Yu Lan, for their valuable advice in helping me finish this dissertation.

Finally, I would dedicate my deepest gratitude to my parents. I would like to thank my parents who encourage me and support me to finish this dissertation with their enduring love, care, and support.

1.0 INTRODUCTION

1.1 STATEMENT OF THE PROBLEM

1.1.1 Multilevel data

In the educational and psychological setting, it is common that data has hierarchical structure. For example, to understand students' perception of math classroom engagement, questionnaires are usually handed out to students who are nested within classrooms while classrooms are nested within schools. Sampling units in the analysis can be students, classrooms, or schools. Data contain not only the information about individual characteristics but also the characteristics of classroom and schools.

Hierarchical data are often collected through the complex sampling design, such as cluster sampling, which randomly selects intact groups instead of individuals. Members of each selected group are considered to have more similar characteristics than members of different groups. Examples of clusters are school districts, schools, and classrooms. Cluster sampling can be carried out in stages, named multistage sampling design. For example, a district is selected, followed by schools in the district, and then classrooms in the schools are randomly selected (Gay, et al., 2006). There could be a various number of stages in the multistage sampling design. Two-stage sampling design is very common and as a result, the two-level model is the most

commonly used multilevel model to study the clustering effect (Asparouhov & Muthen, 2006). In the two-level model, variables at the cluster level can influence variables at the individual level. For example, in studying the individual behavior within the organization, the researcher need not only measure the individual characteristics but also measure the organizational factors that may influence the individual behavior (Hofmann, 1997).

1.1.2 Ignoring Clustering in Multilevel Data

Hierarchical data structure is a common phenomenon, but many studies still chose to ignore the multilevel structure, as pointed out by Pornprasertmanit et al. (2014). For example, in the study of Cassidy et al. (2005), the quality of child care environment was only measured at the class level while the clustering within the school and school district was not taken into account. The factor analysis was only performed on one single-level. The ignorance of multilevel data structure could be on purpose because researchers were only interested in the subjects at the lowest level in which clustering is only considered a nuisance effect. The ignorance of clustering could also be due to the difficulty in identifying the primary sampling units or the complexity of the data structure (Wu & Kwok, 2012).

When subjects are sampled using the simple random sampling, subjects are independent. However, when subjects are sampled using the complex sampling such as the cluster sampling, subjects in the same group are more likely to be similar than subjects from different groups. Therefore, ignoring clustering will result with inaccurate estimates of parameter and standard errors(*SE*) in, e.g., regression models (Moerbeek, 2004; Openakker & Van Damme,2000). It may be unable to find the correct relationship among variables if one level is ignored. For example, if the relationships between two variables are different across levels, researchers may be unable to

obtain the information about the relationship at the ignored level. In addition, the relationship at the retained level may also be distorted (Julian, 2001; Raudenbush & Bryk, 2002).

1.1.3 CFA with Multilevel Data

CFA has been used in identifying latent constructs underlying observed variables that can be on interval or ordinal scales. When subjects in CFA are cluster sampled, using the single-level CFA and ignoring clustering is problematic. Before examining the consequences of using the single-level CFA in the multilevel data, two traditional approaches for multilevel data are discussed: disaggregated and aggregated analysis.

Disaggregated analysis does not control for the clustering effect. Disaggregated analysis ignores the higher level data structure and only models observations at the lower level. In the CFA analysis, the bias introduced by ignoring clustering effect depends on the factorial structure across levels. According to Wu and Kwok (2012), factor loadings in the single-level CFA of multilevel data were more accurate when the factorial structures across levels were the same or when the factorial structure of the within-level was more complex than that of the between-level. As for the factor covariance, even when the between-level and within-level model were the same or when the within-level model was more complex than the between-level model, the factor covariance estimated using the single-level model was twice the factor covariance of the true model. Pornprasertmanit, et al. (2014) found that the factor loading was overestimated when the between-level communality was high and underestimated when the between-level communality was low. When the between-level communality was high, the standard error (SE) of factor loading estimated using disaggregated analysis was negatively biased. When the between-level communality was low, SE of factor loading estimated using disaggregated analysis was

positively biased. Also, the Standard error of the factor correlation tended to be high when the between-level communality was low.

The other traditional approach for multilevel data is the aggregated analysis. Aggregated analysis treats the parameter as the "marginal" parameter over clusters. Aggregated analysis ignores clustering and focuses on the variable averaged from each macro-level unit (Bollen, Tuller, & Oberski, 2013; Pornprasertmanit, et al., 2014; Wu & Kwok, 2012). According to Pornprasertmanit, et al., (2014), when the standardized factor loadings were not equal in two levels, the factor loadings estimated using aggregated analysis were biased especially when ICC was low and cluster size was small. Standard errors of factor loadings estimated by aggregated analysis were generally underestimated. The factor correlation and standard errors of the factor correlation estimated by aggregated analysis were also biased when the factor correlation was not equal across levels.

1.1.4 Model-Based Approach and Multilevel CFA

It has been stated that it is problematic to use aggregated analysis and disaggregated analysis in the multilevel data. Other approaches have been proposed to account for the phenomenon of clustering in the multilevel model. The common approaches are model-based approach and design-based approach. The model-based approach analyzes the multilevel data by using the hierarchical modeling and specifying the between-level and within-level relationship separately. The multilevel model can be used to find the effect of the cluster level variables on the individual outcome and the individual level variables on the individual outcome (Raudenbush & Bryk, 2002). Using a multilevel approach, within-group and between-group relations are modeled simultaneously.

Multilevel CFA is purposefully developed for CFA of multilevel data (Pornprasertmanit, Lee, & Preacher, 2014; Wu & Kwok, 2012). The two-level CFA model is the extension of single-level CFA to the two levels. It can investigate the factorial structure at both the within-level and the between-level. The multilevel covariance structure model is used to estimate the parameters and standard errors in the two-level CFA model. Different from single-level CFA, the total covariance matrix is separated into the within-level and between-level covariance matrix. Factor loadings, factor covariance matrix, and covariance matrix of error terms are estimated at both levels.

1.1.5 Design-Based Approach and CFA Adjusting for Standard Error

The design-based approach focuses on how a sample is drawn from a target population in nested data. This approach still uses the single-level model but adjusts the standard error so that subjects are like being selected using the simple random sampling. The selection probability is created for each unit at the between-level, and for each unit at the within-level accounting for the cluster size. Then, the selection probability is used to create the sampling weight for each unit at both levels (Stapleton, 2002).

In the design-based approaches, in order to adjust for standard errors, commonly used methods are design effect (DEFT) method, weighing method, and linearization method. The DEFT method is to multiply the standard error by square root of the mean of the design effect, the weighting approach is to create a new sampling weight by using different scaling methods, and the linearization method is to use the Taylor Series algorithm and obtain a weighted variation. There has been only one study comparing DEFT method, the weighting method, and

the linearization method in the complex sampling design (Stapleton, 2002). It is hard to make a conclusion about the advantage and disadvantage of these methods.

1.1.6 Design-Based Approach or Model-Based Approach

There is still debate about whether to use the design-based or model-based approach for multilevel data. Comparing the single level analysis adjusting for the clustering (i.e., design-based approach) and the multilevel analysis (i.e., model-based approach), the proportion of over-rejection of no significant effect of the parameter was reduced when the multilevel modeling was considered (Moulton, 1990). Even when the number of individuals within a cluster was reduced, the multilevel modeling still performed better than the single-level analysis using clustering correction. Chuah (2009) compared among single-level analysis using clustering correction and multilevel analysis using a small number of within-cluster observations. It was found that multilevel modeling outperformed models with clustered standard errors or normal standard errors.

Bollen, Tueller, and Oberski (2013) indicated that the model-based approach had high requirement for model specification while the design-based approach did not depend on correct model specification. But the design-based approach was sensitive to small sample size. They also indicated that whether design-based or model-based approach should be used in SEM to correct for the standard errors in the clustered data depended on various factors such as whether the cluster level model had a similar structure to the individual level model. Wu and Kwok (2012) claimed that design-based approach had its advantage in that only one single model need to be specified, but admitted that in the design-based approach, it was assumed that the within-level and between-level relationships were exactly the same. Wu and Kwok (2012) also thought that

single-level model might be preferred when the higher-level was not of the interest or the multilevel model was difficult to implement, while the two-level model might be preferred when the between-level relationship is of interest.

1.1.7 Simulation Studies Comparing CFA with MCFA in Clustered Data

The research about the impact of ignoring clustering in CFA was very limited. Julian (2001) and Pornprasertmanit et al. (2014) examined the impact of ignoring clustering for continuous indicators, while Stochl et al. (2015) studied the impact of ignoring clustering for ordinal indicators. According to the study of Pornprasertmanit et al. (2014), different factor loading or factor correlation across levels resulted with biased parameter estimates and standard error when the single-level CFA model was the analysis model. The study varied factor loading, factor correlation, ICCs, distribution of factor loading, distribution of ICCs, and sample size. This study provides good guidance about the factors that affect the accuracy of parameter estimates, but the study only used the continuous indicators. Julian (2001) also varied within-group sample size, between-group sample size, ICC, factor loading, and factor correlation. The model parameters were overestimated while standard errors were underestimated by the single-level CFA model. Only the continuous indicators were simulated in this study.

The limitation of Wu and Kwok (2012) was that the study only varied between-cluster and within-cluster sample size, ICCs, and number of factors at the within and between-level. The factor loading and factor correlation were fixed. In addition, the study treated the indicators as continuous and only used the MLR as the estimation method while mean and variance adjusted WLS (WLSMV) was an alternatively recommended estimation method for ordinal indicators (Flora & Curran, 2004). Stochl et al. (2015) mainly focused on comparing the estimation

methods and only varied ICC. It was thought that MCMC was more robust than FIML and WLSMV in estimating parameters and standard errors when the clustering was ignored.

1.1.8 Estimation methods in CFA

Single-level CFA has been applied to variables on an interval scale or an ordinal scale. For variables on an interval scale, ML (Maximum Likelihood) or MLR (Maximum Likelihood Robust) are the dominant estimation methods depending on whether there is concern with violation of multivariate normality. For variables on an ordinal scale, estimation methods include those for the continuous variable (ML and MLR) and those specifically developed for ordinal variables (e.g., WLSM, WLSMV). Researchers have studied the performance of ML, MLR and WLSMV for ordinal variables (Asparouhov & Muthen, 2007; & Herzberg, 2006; Flora & Curran, 2004; Lei, 2009; Li, 2010; Raykov, 2012; Rhemtulla, et al., 2012; Wirth & Edwards, 2007; Yu, 2002). The number of categories, sample size, and the shape of the distributions affect the performance of these estimation methods in the ordinal data.

When the number of categories is sufficiently large such as equal or greater than five, ML and MLR can be used to estimate the ordinal data although they were developed for the continuous data. WLSMV was found to more accurately estimate parameters when the number of categories were two, three, four, five, and seven (Beauducel & Herzberg, 2006; Orangje, 2003; Yang, Joreskog, & Luo,2010). When the threshold was symmetric, MLR could accurately estimate the factor loading of the ordinal data. But nonsymmetric thresholds and skewed distribution affected the accuracy of parameter estimates using MLR (Rhemtulla et al., 2012; Li,2010) . The accuracy of parameter estimates of MLR and WLSMV were affected by small

sample size (Lei,2009; Arnold-Berkovits, 2002). The accuracy of parameter estimates of ML did not seem to be affected by small sample size (Destifano, 2002; Bentler ,2006).

There have been very few studies comparing ML, MLR, and WLSMV in the multilevel CFA data, so it is unclear whether the performance difference of these three methods in the single-level CFA analysis can be extended to multilevel analysis.

This study aims to examine the consequence of ignoring clustering in CFA for ordinal indicator variables. There have been four simulated studies comparing MCFA with single-level CFA adjusting for standard error or using normal standard error, but none of them compared multilevel CFA, single-level CFA with adjusted standard error (complex model), and single-level CFA with normal standard error(normal model) at the same time. This study is the first simulation study to compare the accuracy of parameter estimates and related standard errors obtained from three methods simultaneously. This study aims to examine the factors that will affect the accuracy of the parameter estimates and standard errors when clustering is ignored. Survey and questionnaires are commonly used to collect data in educational, psychological and social sciences, thus the analysis of multilevel data cannot be limited in analyzing the continuous data. This study aims to compare three methods using the most commonly used five-category Likert-type scale. It will examine whether the findings from previous simulation studies using the continuous indicators can be extended to the ordinal items and whether findings from previous studies using ordinal indicators can be generalized to the current study when different simulation factors and different estimation methods are used. It will examine under which combination of the sample size, ICC, and factorial structure, the relative bias of the parameter estimates and standard errors are the smallest when the clustering structure is ignored.

1.2 PURPOSE OF THE STUDY

This study aims to examine the consequence of ignoring clustering in CFA for ordinal indicator variables. There have been four simulated studies comparing MCFA with complex single-level CFA model and normal single-level CFA model, but none of them compared multilevel CFA, single-level CFA with adjusted standard error, and single-level CFA with normal error at the same time. This study is the first simulation study to compare the accuracy of parameter estimates and related standard errors obtained from three methods simultaneously. This study aims to examine the factors that will affect the accuracy of the parameter estimates and standard errors when clustering is ignored. Survey and questionnaires are commonly used to collect data in educational, psychological and social sciences, thus the analysis of multilevel data cannot be limited in analyzing the continuous data. This study aims to compare three methods using the most commonly used five-category Likert-type scale. It will examine whether the findings from previous simulation studies using the continuous indicators can be extended to the ordinal items and whether findings from previous studies using ordinal indicators can be generalized to the current study when different simulation factors and different estimation models are used. It will examine under which combination of the sample size, ICC, and factorial structure, the relative bias of the parameter estimates and standard errors are the smallest when the clustering structure is ignored.

1.3 RESEARCH QUESTION

1. What is the difference in terms of model fit indices calculated from the two-level CFA model, single-level CFA model with normal standard error or complex standard error? What model fit indices, if any, are recommended in model selection?
2. How are three models compared in estimating the within-level factor loading and their standard errors? What design factors may impact the performance of these models?
3. What factors influence the performance of the two-level CFA model in recovering between-level factor loading?
4. What factors affect the performance of the two-level CFA model in recovering residual variance?

1.4 SIGNIFICANCE OF THE STUDY

Clustered data are commonly encountered in the educational and psychological setting, but researchers often missed the identification of clustering either by chance or on purpose. To find the consequences of misspecifying the multilevel data is important and meaningful. This study is the first simulation study to compare the accuracy of parameter estimates and related standard errors using MCFA, complex single-level CFA model, and normal single-level CFA model simultaneously. The study extended the simulation study of Julian (2001) and Pornprasertmanit et al. (2014) by incorporating the ordinal indicator variables, which made the result of the study applicable to the Likert-type questionnaires. Compared with the simulation study of Wu and Kwok (2012), this study added the factor loading and factor correlation as the simulation factor

and used WLSMV as an estimation method instead of MLR. From Table 1, it is easy to find the differences between my study and previous studies. In summary, this study aims to provide a comprehensive guideline in terms of comparison of single-level CFA models and multilevel CFA in a wide range of conditions commonly encountered in empirical multilevel CFA studies.

Table 1. Summary of previous studies and significance of the current study

	Type of Indicators	**Within-level /Between-Level Factor Number**	**Estimation Method**	**Sample size (number of clusters/number of cluster members)**	**ICC**	**Within /Between-Level Factor Loading**	**Within/ Between-Level Factor Correlation**
Julian (2001)	Continuous indicators	4/4 4/2 4/5	ML	100/5, 50/10, 25/20,10/50	0.05,0.15 0.45	Fixed to be 1	Fixed to be 0.5
Pornprasertmanit, et al. (2014)	Continuous indicators	2/2	ML	100/5, 10/50 400/20,40/200	0.05,0.25, 0.5,0.75, 0.95	0.7/0.49,0.7, 0.86	0.5/0.2, 0.5,0.8
Wu and Kwok (2012)	Continuous Indicators	3/1 3/3 1/3	MLR	50, 150, and 300/10, 50, and 200	0.1,0.5	Fixed to be 0.8	Fixed to be 0.3
Stochl (2015)	Ordinal Indicators	4/	FIML, MCMC, WLSMV	Fixed to be 1000	0.001-0.390 11 levels	Fixed to be 0.7	Fixed to be 0.4
Current Study	Ordinal Indicators	2/1 2/2	WLSMV	50/10, 20/25, 250/10, and 100/25	0.25, 0.45	0.8/0.5; 0.5/0.8; 0.5/0.5	0.3/0.3; 0.3/0.6;

2.0 LITERATURE REVIEW

The purpose of this chapter is to review (1) confirmatory factor analysis(CFA) in section 2.1; (2) multilevel confirmatory factor analysis (MCFA) in section 2.2; (3) complex sampling design and methods of adjusting the standard error when clustering is ignored in section 2.3;(4) findings of studies comparing single-level CFA with MCFA in section 2.4.

2.1 CONFIRMATORY FACTOR ANALYSIS

CFA is an important statistical method in SEM. This section will focus on the measurement model of the CFA, the covariance structure of CFA, the theory of CFA using categorical variables, the comparison of estimation methods in the CFA, and the model fit of CFA.

2.1.1 Introduction to CFA and CCFA

2.1.1.1 CFA

Structural equation modeling is a statistical modeling technique that can be described as a combination of common factor analysis and a set of multiple regressions. It has the ability to handle the latent variable, observed variables, and measurement errors at the same time (Hox, 1998). It is not just a single analysis but a collection of different techniques including path

analysis, confirmatory factor analysis (CFA), latent growth curve modeling, multilevel SEM for clustered data, and multi-group SEM. In this study, CFA is the focus.

CFA is a statistical tool to examine the latent common factors underlying a set of observed variables (Kline, 2005; Muthen & Muthen ,1998-2002). CFA is commonly used for assessing the construct validity, developing and improving measurement instruments, and evaluating factor invariance across time or groups (Jackson, Gillaspy, & Purc-Stephenson, 2009).

CFA aims to explain the covariance/ correlations among variables using the specified model. Technically, it is to test the hypothesis that the observed covariance matrix is equal to the model implied covariance matrix. It can be specified as the following:

$$\sum = \sum(\theta) \qquad (1)$$

where $\sum$ is population covariance matrix and $\sum(\theta)$ is model implied covariance matrix using population parameters. In CFA, the model implied covariance matrix turns to be:

$$\sum(\theta) = \Lambda_x \Psi \Lambda_x' + \Theta_\delta \qquad (2)$$

where Ψ is covariance matrix of the latent factors , Λ_x is the matrix of factor loadings, and Θ_δ is the covariance matrix of measurement errors.

The measurement model for CFA is the regression model that describes the relationship between observed variables and latent continuous variables. For the continuous factor indicators, the relationship among variables are expressed by linear regression model; for the binary or ordered categorical factor indicators, the relationship among variables are expressed by probit or logistic regression model; for unordered categorical factor indicators, the relationship among variables are expressed by multinomial logistic regression model. CFA can detect the

relationships among factors, and the relationship between factors and observed variables (Muthen & Muthen, 1998-2002).

In the model of linear CFA, the relationship between the latent factor and the continuous indicators are specified as the following:

$$x_{ij} = \tau_j + \lambda_j \xi_i + \delta_{ij} \qquad (3)$$

where x_{ij} is an observed variable *i* loading on factor *j*, τ_j is an intercept, λ_j is a factor loading, ξ_i is a specified factor, and δ_{ij} is the residual of the item. The model linearly relates the response variables to the specific factor. The ordinary CFA assumes the observed variables are continuous and normally distributed (Kim & Yoon, 2011).

In order to perform CFA, the researcher needs to know the number of factors and on which factor each item loads. CFA can be performed using either correlation or covariance matrix. CFA provides a test of significance for factor loadings. Significant indicators are retained in the model while insignificant indicators are candidates for dropping. To identify a CFA model, the number of unique elements in a covariance matrix must be greater than or equal to the number of parameters.

2.1.1.2 CCFA

The ordinary CFA treats dichotomous or polytomous responses as continuous variables and ignores the categorical nature of the data, which may lead to biased parameter estimates. CCFA assumes that ordered-categorical item responses are discrete representation of continuous latent responses. The latent response variables are manifested as discrete scores with a set of thresholds. The distribution of categorical response of a particular variable is determined by the latent response distribution with a set of threshold parameters.

$$x_{ij} = c, \text{if } \tau_{jc} < x_{ij}^{*} < \tau_{jc+1} \quad (4)$$

where τ_{jc} is the *c* ordered response of the *jth* item and c=0,1,….,C-1. C is the number of categories of an ordinal variable. It assumes that individuals possess a latent score, x_{ij}^{*}, of individual *i* on item j. The distribution of categorical responses to a particular item is reflected by the latent score distribution of x_{ij}^{*} corresponding to that item. The threshold parameter τ is the point on the continuous latent response scale that separates the manifest discrete responses (Wirth & Edwards, 2007). Even the ordinal variables are discrete in nature, it is still assumed that the underlying latent responses are continuous (Kim & Yoon, 2011).

2.1.2 CFA Model Fit and Model Fit Indices

Model fit indices are important in detecting whether the specified model fit the sample data. An initial model specified according to a theory could be inappropriate in reproducing a sample covariance matrix, producing large difference between the sample covariance matrix and model specified covariance matrix. The chi-square test and other model fit indices will suggest to reject the model. Under these circumstances, the model can be modified to improve the model fit. Model modification involves freeing fixed parameters or fixing free parameters. Free a fixed parameter will increase the number of parameters in the model and decrease the degree of freedom while fixing a free parameter will decrease the number of parameters in the model and increase the degree of freedom. In the single-level CFA model, the relationship among the factors should not be too high. When factors are highly correlated with each other, there could be an existence of a higher order factor.

The common model fit indices are chi-square test, CFI (Comparative Fit Index), TLI (Tucker-Lewis Index), SRMR (Standardized Root Mean-square Residual), and RMSEA (Root Mean Square Error of Approximation). CFI is an incremental fit index. A model with CFI ≥ 0.95 is considered with good fit. Tucker-Lewis Index (TLI) is another incremental fit index, which ranges from 0-1. A model with TLI ≥ 0.95 is a good model. A higher value indicates better model fit. RMSEA is an absolute fit index. A model with RMSEA ≤ 0.06 is considered a good fit while RMSEA ≤ 0.08 is considered an acceptable fit. SRMR is another absolute fit index, representing geometric mean of residuals. An SRMR ≤ 0.08 is considered a good fit. CFI, TLI, RMSEA, and SRMR provide different information about model fit, thus it is beneficial to use the combination of them.

2.1.3 Estimation Methods of CFA with ordinal variables

In SEM, when factor analysis is performed on ordinal variables, treating the observed variable as continuous leads to biased parameter estimates. Thus, the ordinal nature of the variables should be taken into account. From the previous studies, the most commonly adopted methods to estimate the CFA for ordinal variables are ML, MLR, and WLSMV. These methods are specified as the following.

2.1.3.1 Maximum Likelihood (ML)

The most common estimation method for CFA is maximum likelihood which assumes that the indicators are continuous and normally distributed. It produces asymptotically unbiased, consistent estimators of parameters (Bollen, 1989). Maximum likelihood maximizes the likelihood of the observed data. This is equivalent to minimize the discrepancy function F_{ML}:

$$F_{ML} = \ln|\sum(\theta)| + trace[S\sum^{-1}(\theta)] - \ln|S| - p \quad (5)$$

where θ is the vector of the model parameters, $\sum(\theta)$ is the model implied covariance matrix, S is the sample covariance, and p is the total number of observed variables (Bollen, 1989).Based on the continuous and normal distribution, the sample covariance matrix is computed. When there is adequate sample size, multivariate normal distribution, and correct model specification, ML provides consistent, efficient, and unbiased parameter estimates, asymptotic standard error, and good model fit.

ML is not appropriate for the categorical data. It is mainly because that the sample product moment relationship such as Pearson correlation or polychoric correlations does not perform well with ordinal variables using ML. The chi-square statistic is inflated and parameters and related standard errors are negatively biased (Flora & Curran, 2004).

2.1.3.2 Robust maximum Likelihood (MLR)

MLR produces maximum likelihood estimation with robust standard error. The corrected standard errors are obtained using a sandwich-type estimator, which incorporates an observed Fished information matrix, $\Delta' I_{obs} \Delta$, into the asymptotical covariance matrix of the estimated parameter vector $\hat{\theta}$. The corrected standard error estimates are calculated by taking the square root of the diagonal of the estimated asymptotic covariance matrix. The estimated asymptotic covariance matrix is formulated as the following:

$$aCov(\hat{\theta}) = [N^{-1}(\hat{\Delta}' \hat{I}_{obs} \hat{\Delta})^{-1} \hat{\Delta} \hat{I}_{obs} \hat{\Gamma} \hat{I}_{obs} \hat{\Delta} (\hat{\Delta}' \hat{I}_{obs} \hat{\Delta})^{-1}] \quad (6)$$

Where $\Delta' I_{obs} \Delta$ is the observed Fisher information matrix, $\hat{\Delta}$ is calculated by taking the first derivative of the covariance matrix with respective to θ, and N is the number of observation.

$\hat{\Gamma}$ is taken as *W* in the weighted least squares fitting function when variables are continuous (Satorra, 1992; Satorra & Bentler, 1994; Muthen & Satorra, 1995).

MLR considers the distribution properties of the items. MLR can be used to deal with both continuous and categorical data. MLR can be used to deal with nonnormal data and missing data. MLR produces both a rescaled chi-square test statistics and standard errors that are robust to non-normality (Satorra, 1992; Satorra & Bentler, 1994). Parameter estimates are still obtained using asymptotically unbiased estimator, but standard errors and chi-square statistics are corrected to enhance the robustness of ML to nonnormality. Thus, the parameter estimates by MLR are the same as those estimated by ML. Only the standard errors and chi-square tests are different. The mean- and variance-adjusted chi-square statistic in Mplus is also known as the Satorra-Bentler scaled X^2.

2.1.3.3 WLS and WLSMV

The weighted least square approach was proposed to estimate a weight matrix based on the asymptotic variances and covariance matrices. Flora and Curran (2004) and Wirth and Edwards (2007) discussed the use of WLS in the categorical data. The WLS function for categorical variable was defined as

$$F_{WLSC} = (r-\rho)' W^{-1} (r-\rho) \qquad (7)$$

where W^{-1} was the inverse of a weight matrix that is positive definite, ρ was a $p \times p$ model correlation matrix and r was a $p \times p$ sample correlation matrix which could be tetrachoric or polychoric correlation matrix. The weakness of the WLSc was mentioned by previous researchers (Wirth and Edwards, 2007; Flora and Curran, 2004). The number of unique variances and covariances grew rapidly as the number of indicators increased. In the large model, W was

often nonpostive definite and could not be inverted. WLS_C requires a sufficiently large sample to estimate an accurate weight matrix. Wirth and Edwards (2007) pointed out that there was no closed form solution to the asymptotic covariance matrix of categorical data in the above equation. The computational burden of using full weight matrix was another issue. Thus, it was proposed to use only the diagonal elements of the weight matrix instead of using a full weight matrix.

According to Wirth and Edwards (2007), modified WLS for ordered-categorical indictors was defined as

$$F_{M\,WLSC} = (r-\rho)' W_D^{-1} (r-\rho) \qquad (8)$$

W_D^{-1} contained only the diagonal elements of the full weight matrix. This modification greatly reduced the number of nonzero elements and therefore reduced the computational burden. Two MWLSc estimators were mean adjusted WLS (WLSM) and mean and variance adjusted WLS (WLSMV). Because of the removal of off-diagonal elements, weight matrix was not the optimal weight matrix. MWLSc had the biased standard errors and test statistics. One way to correct inaccuracies was to use the Satorra-Bentler scaled chi-square and robust standard errors. It adjusted the chi-square test statistic and standard errors of the parameters but did not adjust the model degrees of freedom (Satorra & Bentler, 1994; Yuan & Bentler ,1998). Another method proposed by Muthen et.al.(1997) adjusted the chi-square test statistics, standard errors, and the model degrees of freedom. WLSMV was appropriate in the small to moderate sample size. Chi-square of the WLSMV is computed using the second-order correction of the fit function while chi-square of the WLSM uses the first-order correction (Asparouhov & Muthen, 2007).

2.1.4 Using Estimators for Continuous Variable in Categorical Data

Although previous researchers have used ML, MLR, and WLSMV in estimating ordinal variable, WLSMV was specifically proposed to estimate ordinal data while ML and MLR was specifically developed to estimate continuous data. WLSMV makes no distributional assumption about the variables. Studies have been performed to examine the performance of ML, MLR, WLS, and WLSMV on the categorical data and the advantage of WLSMV over ML and WLS was obvious. The performance of MLR and WLSMV in the ordinal data need to be examined and compared, as well as the impact from the number of indicators, the number of categories, sample size, and the shape of distributions.

2.1.4.1 Effect of the number of indicators, the number of categories, sample size, and the shape of distributions on WLSMV

The number of indicators did not have effect on WLSMV (Flora & Curran, 2004; Lei, 2010). Flora and Curran (2004) found that increasing the number of indicators did not have effect on WLSMV. Lei (2009) also found that the number of indicators did not seem to have a significant effect on the relative bias of parameter estimates and convergence rate comparing six variable model with nine variable model.

WLSMV was typically developed for ordinal data. With the increasing of number of categories, the performance of WLSMV might not be as good as ML or MLR. However, there is no strict rule for the maximum number of the categories for the WLSMV. It was found that WLSMV performed well in two, three, and four responses (Beauducel & Herzberg, 2006). Oranje (2003) found that WLSMV performed equally well in the parameter estimates for the two, three, and five-category responses. Yang, Joreskog, and Luo (2010) found that factor

loadings and factor correlations obtained by WLSMV was unbiased regardless of the number of categories (two, five, or seven).

Small sample size affected the parameter estimates of WLSMV (Li, 2010; Lei, 2009; Arnold-Berkovits, 2002). Li (2010) found that WLSMV produced moderate overestimation of the interfactor correlation when the sample size was as small as 200. Lei (2009) found that WLSMV was sensitive to small sample size when the sample size was as small as 100. The standard error estimated from WLSMV became more negatively biased when the sample size became small. But Beauducel and Herzberg (2006) found that small sample size did not have effect on the parameter estimates since WLSMV performed well even in small sample, large model with moderate loadings.

The shape of the distribution affected the parameter estimates of WLSMV (Li, 2010; Flora & Curran, 2004). Li (2010) found that WLSMV produced moderate overestimation of the interfactor correlation when the distributions were moderately nonnormal. Flora and Curran (2004) simulated categorical data with skewness up to 1.25 and kurtosis up to 3.75. It was found that increasing the skewness and kurtosis generally increased the relative bias of factor loadings and factor correlations across the sample sizes. However, Lei (2010) and Yang, Joreskog, and Luo (2010) found that factor loadings and factor correlations obtained by WLSMV was unbiased regardless of the shape of distribution (symmetric vs asymmetric).

2.1.4.2 Effect of the number of indicators, the number of categories, sample size, and the shape of distributions on MLR

The number of indicators did not seem to affect parameter estimates of MLR. Lei (2009) found that the number of indicators did not seem to have a significant effect on the relative bias of

parameter estimates and convergence rate comparing six variable model with nine variable model.

It was still debate whether MLR should be used for the number of categories larger than five (Li, 2010; Raykov, 2012; Rhemtulla, et al. ,2012; Beauducel & Herzberg,2006). Raykov (2012) and Rhemtulla, et al. (2012) thought that MLR was preferred for the number of categories equal to or larger than five. In the study of Rhemtulla, Brosseau-Liard, and Savalei (2012), it was found that that the relative bias of MLR was not larger than 10% with five or more categories in any of the conditions. It seems that the estimation method developed for the continuous variable is also appropriate. When category threshold was generally symmetric, MLR was as good as or better than WLSMV. However, Bequducel and Herzberg (2006) suggested that WLSMV should be better than the continuous method for the number of categories equal to five, six, and seven. Also, in the simulation study of Li (2012), it was disappointing that relative bias of the factor loading was large for the number of category larger than five and there was even substantially negative bias in parameter estimates and standard errors and low rate of the coverage of factor loadings using MLR for four- category response.

The shape of the distribution affected the MLR estimates (Li, 2010). Li (2010) simulated the slightly nonnormal distribution with skewness of 0.5 and kurtosis of 1.5 and the moderately nonnormal distribution with the skewness of 1.5 and kurtosis of 3.0. MLR underestimated the factor loadings. However, Lei (2009) found that MLR estimates were unbiased across the shape of the distribution (symmetric, mildly skewed, and moderately skewed).

2.1.4.3 Compare MLR and WLSMV in the Categorical Confirmatory Factor Analysis

In conclusion, non-normal distribution affected the performance of both MLR and WLSMV (Li, 2010). Li (2010) found that WLSMV was more accurate in estimating the factor loading while

MLR was more accurate in estimating the factor correlation under non-normality. However, Lei (2009) found that the shape of distribution did not seem to affect the relative bias of parameter estimates for both MLR and WLSMV. Small sample size affected the accuracy of parameter estimates for both MLR and WLSMV. Lei (2009) and Arnold-Berkovits (2002) found that WLSMV was more sensitive to small sample size than MLR when the sample size was as small as 100. The minimum sample size for WLSMV should be 250. With the combination of small sample size and non-normal distribution, Li (2010) found that MLR outperformed WLSMV in estimating the standard error of factor loadings and factor correlations.

The number of indicators did not seem to significantly affect the relative bias of parameter estimates and convergence rate for MLR and WLSMV (Flora and Curran ,2004; Lei, 2009). Even there was impact of model size on the standard error of parameter estimates, there was slight difference between two methods. WLSMV is more appropriate than MLR for number of categories less than five (Li, 2010). Although WLSMV was developed for categorical data, Li (2010) found that WLSMV outperformed MLR in estimating factor loadings even for the number of category of 6, 8, and 10. Studies have been performed on comparing ML and WLSMV. However, limited studies have been performed on comparing MLR with WLSMV.

2.2 MULTILEVEL CONFIRMATORY FACTOR ANALYSIS

This section focuses on the CFA under the circumstance of the multilevel data. MCFA is a commonly recommended model-based approach in analyzing the relationship among variables in the multilevel data. MCFA has been demonstrated to be more accurate in the parameter estimates (factor loading and factor correlation), their standard errors, and item thresholds than

the classical CFA when there is existence of clustering. Fit indices (i.e. CFI, RMSEA, and chi-square) have also been proved to be more reasonable when clustering is acknowledged (Stochl et al., 2015; Pornprasertmanit, Lee, & Preacher, 2014; Wu & Kwok, 2012). The following section will first introduce the background in which MCFA is applied. Then the multilevel covariance structure modeling, estimation methods of MCFA, and the definition of the construct and parameters in the two-level model will be described.

2.2.1 Research Background about Organizational Effects in MCFA

In analyzing the multilevel data in the organizational research, the common difficulties are the aggregation bias, misestimated standard errors, and heterogeneity of regression. Aggregation bias occurs when a variable may have different effects at different levels of the organization. In the area of education, for instance, the average socioeconomic status (SES) may predict student achievement above and beyond the individual's SES. Misestimated standard errors occur in the multilevel data when it fails into considering the dependence among individual responses within the same organization. This dependence may occur because of the ways individuals are selected or the same characteristics shared within the organization. Heterogeneity of regression occurs when the prediction of individual characteristics on the outcome measure varies across the organization. The problem of how organizations affect the individuals within the organization can be investigated using multilevel model. In the organizational research, at level-1, the units are individuals and each individual's outcome is measured by a set of individual characteristics; at level-2, the units are organizations such as schools (Raudenbush &Byrk, 2002).

2.2.2 Introduction of MCFA

2.2.2.1 Theory of MCFA

Multilevel CFA (MCFA) is merely an extension of CFA to include various levels in the model. Multilevel covariance structure models can account for the variability for the data collected through the procedure of complex sampling. Multilevel covariance structure modeling allows for the different model structures at the between and within-level (Julian,2001; Muthen, 1994).

The multilevel factor modeling assumes a conventional factor analysis covariance structure at both levels. The level-1 subjects can be expressed in terms of the multilevel linear factor model as:

$$x_{ig} = \nu + \Lambda_w \eta_{wig} + \Lambda_b \eta_{bg} + \varepsilon_{wig} + \varepsilon_{bg} \qquad (9)$$

where x_{ig} is the response for student *i* in group *g*, ν is the grand mean, Λ_w and Λ_b are factor loading matrices for within and between-group, η_{wig} is the random factor varying over students within the respective schools, η_{bg} is the factor varying randomly across groups, ε_{wig} and ε_{bg} are within and between-group errors, and Var(x_{ig})= $\sum_B + \sum_W$. From the above equation, it is clear that the factor loading matrices are allowed to differ across levels (Kaplan,2009; Muthen, 1994).

The single-level CFA uses the total variance-covariance matrix. MCFA separates the between-level and within-level variability and allows for the simultaneous estimation of covariance matrices at both levels.

$$\sum_B = \Lambda_B \Psi_B \Lambda_B^{'} + \Theta_B \qquad (10)$$

and

$$\sum_W = \Lambda_W \Psi_W \Lambda_W^{'} + \Theta_W \qquad (11)$$

where Λ is a vector of factor loadings, Ψ is a covariance matrix of the factor, Θ is the covariance matrix of error terms , and subscripts *B* and *W* refer to between-group and within-group parameters. Thus, when the linear factor analysis is extended to the multilevel, the total sample covariance matrix can be expressed in terms of the factor model parameters (Muthen, 1994; Kaplan, 2009):

$$\Sigma_T = \Lambda_B \Psi_B \Lambda_B^{'} + \Theta_B + \Lambda_W \Psi_W \Lambda_W^{'} + \Theta_W \qquad (12)$$

Multilevel covariance structure modeling can estimate the parameters and related statistics such as ICC accurately.

The estimation algorithm of the multilevel covariance structure analysis with ML is demonstrated as the following. The total sample covariance matrix S_T is a consistent estimator for the total population covariance matrix Σ_T. The pooled-within sample matrix S_{PW} is an unbiased and consistent estimator of population covariance matrix Σ_W. The between sample matrix S_B is an unbiased and consistent estimator of $\Sigma_W + c\Sigma_B$, where c is common group size in the balanced data and close to the mean of the groups sizes in the unbalanced data. Thus, ML estimate of Σ_B is $c^{-1}(S_B - S_{PW})$. It can be told that Σ_B is a function of both S_B and S_{PW}. The estimation procedure of multilevel covariance structure analysis is shown in four steps. Step 1 is to estimate the total sample covariance matrix; step 2 is to estimate the between sample covariance matrix and ICC; step 3 is to estimate the parameters at the within-level; and step 4 is to estimate the parameters at the between-level (Muthen,1994).

2.2.2.2 Definition of construct at the within and between-level

Latent variable at the within-group level represents the properties of the within-group units. Latent variable at the between-group level represents the properties of the between-group units

and it reflects the collective properties of the within-level latent variables such as means or variances of the within-level latent variables (Pornprasertmanit, Lee, & Preacher,2014).

In the single-level CFA, standardized factor loadings for an observed variable represent the correlation between an observed variable and a latent variable without controlling for clustering. Factor loading can be considered as regression coefficient of an observed variable predicted by a factor. The squared correlation is the proportion of variation of an observed variable explained by a latent variable without considering cluster membership (Pornprasertmanit, Lee, & Preacher, 2014).

In MCFA, within-group parameter estimates represent the relationship among variables controlling for the effect of clusters. Between-group parameter estimates in MCFA represent the relationship among variables at the between- level. For example, the between-level factor loading represents the correlation between a between-level indicator and a between-level latent factor.

When the regression coefficients of the within-level and between-level are the same, they will be the regression coefficients for the single-level model. However, when the regression coefficient differs across levels, the regression coefficient calculated from a single-level model is a weighted average of the effects if a regression coefficient is computed from a two-level model (Raudenbush & Bryk, 2002). In the MCFA model, factor loading represents the relationship between the observed indicator and the latent factor. But it is not clear whether the factor loading in the single-level model is the weighted average of the factor loading calculated from the two-level model. It was suggested to compare the factor loading of the single-level CFA model to the factor loading of the within-level of the MCFA model (Pornprasertmanit, Lee, & Preacher,2014).

At the between-level, the squared value of standardized factor loading corresponds to the proportion of between-level variations of an indicator explained by a between-level latent variable. At the within-level, the squared value of standardized factor loading corresponds to the proportion of within-level variance of an indicator explained by a within-level latent variable considering the clustering membership. On the standardized scale, the residual variance of an indicator is calculated by subtracting the communality from 1 at both the within and between-level.

2.2.3 Model Fit in MCFA

The model fit in MCFA is assessed similar to model fit in the single-level CFA. But the model fit in MCFA need to be performed at both the between-level and within-level. According to Hsu et al. (2014) and Ryu and West (2009), RMSEA, CFI, and TLI could only detect the model misspecification of the within-level model but not the model misspecification of the between-level model or the entire model. Hsu et al. (2014) manipulated the number of clusters, number of cluster members, and item ICCs in the two-level CFA model and found that SRMR was the only fit index that could be used to evaluate the within-level model fit regardless of the model complexity. Furthermore, RMSEA, CFI, and TLI were more sensitive to the misspecification of the factor loading while SRMR was more sensitive to the misspecification of the factor covariance. Item-level ICC did not have influence on the performance of the model fit indices.

The limitation of the study of Hsu et al. (2014) was that the model was misspecified through constraining the factor correlations between factors across levels to be 0 or constraining the factor loadings to be 0. In fact, there were other types of model misspecification such as the misspecified factorial structure. Another limitation was that the data was generated using the

multivariate normal distribution, which might make the result of the study unable to be generalized to the multilevel CFA model with categorical indicators. Last, this study only simulated the model with the equal number of indicators per factor and the equal number of subjects within the group.

2.2.4 Estimation method of MCFA

2.2.4.1 ML

The computational algorithm of ML in the two-level CFA has been specified in section 2.2.2.1.

2.2.4.2 MLR

MLR calculates the robust standard errors robust to nonnormality and rescaled chi-square test of model fit. In the multilevel data, robust chi-squares and standard errors provide certain protection against the heterogeneity of subjects and the misspecification of the multilevel model (Hox, et al., 2010).

2.2.4.3 WLSMV

WLSMV in estimating the multilevel SEM of ordinal variables was proposed. At the first step, univariate ML is used to estimate the vector of means at the between-level and the diagonal elements of Σ_B and Σ_W. At the second step, the off diagonal elements of Σ_B and Σ_W are estimated using the bivariate ML. At this step, the model parameters of two levels are estimated by WLSMV. At last, the asymptotic covariance matrix is obtained. WLSMV is the mean-and-variance corrected estimator with robust chi-square. In the multilevel analysis, the number of

parameters in the between-level often tends to be larger than the number of groups, so it is preferred to use only the diagonal elements of the matrix (Hox, et al., 2010).

2.2.5 Compare the Estimation Methods in the Multilevel CFA Studies

Several studies compared estimation methods in multilevel CFA. Hox et al. (2010) simulated a multivariate normal distribution and found that ML performed better than MLR and WLSMV. MLR and WLSMV only accurately estimated the factor loadings when the number of group was large. MLR need the larger number of group than WLSMV to make the estimation of the factor loading accurate. Standard errors of factor loadings estimated from the WLSMV were as accurate as or more accurate than those estimated from ML. MAAS and Hox (2004) found that MLR was more accurate than ML in the multilevel data when the distribution was non-normal. But the large sample size was still required.

For polytomous items, several studies adopted ML or FIML by treating ordinal variables as continuous, especially in those studies with number of categories larger than five (Whitton & Fletcher, 2014; Brondino et al.,2013). Stochl et al. (2015) compared FIML, WLSMV, and MCMC for ordinal variables. Stochl et al. (2015) found that the factor loading and factor correlation estimated by WLSMV using the multilevel model were unbiased regardless of ICC. MCMC did not show obvious advantage over FIML and WLSMV when clustering was incorporated into the model. MCMC was more robust than FIML and WLSMV in the estimation of parameters (factor loading, correlation, and thresholds) when the clustering was ignored. However, SE was underestimated regardless of the estimation methods. Without considering the clustering, the factor loading estimated by WLSMV was underestimated while the correlation was still almost unbiased.

Among previous studies about multilevel CFA, maximum likelihood (ML) (Whitton & Fletcher,2014; Brondino et al.,2013; Ryu, 2014; Greenbaum, Wang, & Boothroyd, 2011; Leonardo Grilli & Carla Rampichini, 2007), maximum likelihood robust (MLR) (Zimprich, Perren, & Horung,2005; Haenens, Damme, & Onghena,2012), Muthen maximum likelihood(MUML) (Wu, 2009; Dyer et al., 2005; Ryu, 2014),and WLSMV (Little, J., 2013; Pornprasertmanit, Lee, & Preacher, 2014; Stochl et al.,2015) were the common methods. Although WLSMV was preferred over other methods for single-level study of the ordinal variables, there have been more empirical MCFA studies using the ML and MLR. There have been very limited studies comparing the estimation methods of MCFA. It is not certain whether the advantage of WLSMV over ML and MLR in the single-level CFA of categorical data can be extended to MCFA.

2.3 COMPLEX SAMPLING DESIGN AND METHODS OF ADJUSTING STANDARD ERROR

This section will first introduce the complex sampling and clustering followed by three methods to adjust for the standard errors in the single-level CFA.

2.3.1 Complex Sampling Design and Clustering

It has been specified that hierarchical data are often collected through cluster sampling. In traditional SEM analysis, assumption is that observations are independent and identically distributed. The large survey data are usually collected through multistage sampling or cluster

sampling. Intact groups rather than individual subjects are randomly selected. The intact group can be school districts, schools, or classrooms. The example of stratified cluster sampling is that schools are selected by stratifying at the appropriate level and then choosing an appropriate proportion of schools in each stratum. Then an appropriate proportion of students are selected in each stratum. The underestimation of the sample variance in the complex sampling design are the common consideration in previous studies (Thomas & Heck, 2001; Stapleton, 2006; Asparouhov & Muthen, 2006; Wu & Kwok, 2013).

Because of the cluster sampling, subjects obtained from cluster sampling are homogeneous in nature. The sample variance estimated from the clustered data is smaller than those estimated from the traditional method assuming independent data. If the study ignores the complex sampling design and the unequal probability of selection, parameters estimated assuming the simple random sampling may depart from the true value. It will cause the inflation of Type I error rate and the incorrect assertion of the significant relationship.

2.3.2 Adjusting Standard Error in the Complex Sampling Design

The consequences of ignoring the clustering mainly reflect in the downward bias of standard error and the inflation of Type I error rate. When the single-level analysis is performed in the multilevel data, the default standard errors are lower than the true standard errors in the clustered data. Except for using the multilevel modeling, another reasonable method is to apply the statistical method of adjusting the standard errors for clustering or to adjust variances to account for homogeneity within clusters. It is understandable that with the increase of ICC, the cluster of homogeneity increases and the standard error is more negatively biased. The method of adjusting standard errors with the single-level model is displayed as the following.

2.3.2.1 Make Adjustment by Using Design Effect

There are various approaches to adjust the variance estimates in a single-level analysis using complex sampling design. The first approach to adjust the standard error is to incorporate an inflation factor, the design effect. The design effect (DEFT) is an expected effect of the complex sampling design on the sampling variance:

$$\text{DEFT}=\frac{SE^2_{complex}}{SE^2_{SRS}} \qquad (13)$$

From the above equation, DEFT is the ratio of the sampling variance obtained using the complex sampling design to the sampling variance that would have been obtained if the simple random sampling is used. It is to inflate the standard error by multiplying the square root of the mean design effect of a variable. Since the DEFT is associated with ICC, it was thought that for those ICC smaller than 0.05, there is little need in applying the DEFT (Stapleton, 2002). This method is conservative in estimating the sampling error in the complex model with a large number of parameters (Stapleton, 2006). Concerning the conservativeness of this approach, the parameter should be evaluated at a more liberal level such as 0.05 rather than 0.01 (Thomas & Heck,2001).

2.3.2.2 Make Adjustment by Using Sampling Weight

The second approach is to create a design-effect adjusted sampling weight. In the complex sampling design, a subgroup may have a higher probability of being selected and thus more weight is given to a certain group. Not using the weighting in the complex sample design causes the underestimation of the true variance of the population. It was found that the standard error was negatively biased without using the sampling weight in the complex sampling design

(Stapleton, 2006). Make adjustment to the weight will make the relative frequency of the observations in the sample in congruence with those in the population (Walker & Young, 2003).

This adjusted sampling weight is created through dividing the normalized sampling weight by the average design effect of a variable. Most commonly used weights are raw weight and relative weight. Sum the raw weight across all observations yields the effective sample size *N*: $\sum_{i=1} w_i = N$, where w_i is the raw weight for individual *i*. The observation of a higher probability of selection has a small raw weight. The weighted mean is calculated as sum of product of raw weight and sample statistic x_i divided by the sum of the raw weight:

$$\hat{\mu} = \frac{\sum_{i=1} w_i x_i}{\sum w_i} \tag{14}$$

And the variance of the weighted mean is calculated as:

$$\text{var}(\hat{\mu}) = \frac{\sum_{i=1} w_i (x_i - \hat{\mu})^2}{\sum_{i=1} w_i (\sum_{i=1} w_i - 1)} \tag{15}$$

The relative weight is calculated as $w_i / \bar{w}$. The relative weight is preferred over the raw weight in the complex sampling design because the relative weight can yield the effect sample size while still adjusting for oversampling (Walker & Young, 2003).

In the multistage stratified sampling, sampling weights are assigned to one of the levels or to both levels. The lower level of the clustering had the greater effect on the parameter estimates than the higher-level (Asparouhov & Muthen, 2006). The sampling weight on the between-cluster level is $w_j = 1/p_j$, where p_j is the probability that cluster j is included in the sample. The sampling weight on the within-cluster level is $w_{i/j} = 1/p_{i/j}$, where $p_{i/j}$ is the

probability for individual *i* in cluster *j* of being selected given the cluster *j* is selected. The within-cluster weights are commonly scaled to improve the estimation. For example, weights are standardized so that they can be summed to the sample size of the cluster. Scaling to the cluster sample size will give the most robust performance.

The scaled within-level weight for individual *i* in the cluster *j* is calculated as

$$w_{ji}^{*} = w_{ji} \frac{n_j}{\sum_{i=1} w_{ji}}, \tag{16}$$

where n_j is the size of cluster j and $\sum_{i=1} w_{ji}^{*} = n_j$. For the *jth* cluster, the scaled weight is calculated as

$$w_{j}^{*} = w_{j} \frac{G}{\sum_{j=1} w_{j}} \tag{17}$$

where w_j is the raw weight for the cluster *j* and *G* is the number of clusters (Stapleton,2002). The accuracy of the weight method depends on the cluster size, informativeness of the within-level weights, ICC, and the unequal weighting effect (Asparouhov & Muthen, 2006). The advantage of the weighting method is the simple calculation of the weight. In addition, it is simple to adjust all standard errors simultaneously by applying just one change in the weight instead of manually multiply each standard error by an inflation factor(Stapleton,2006).

2.3.2.3 Linearization Method

The third approach is to estimate the sampling variance by using linearization. Using the Taylor Series linearization method, the variance is calculated as a weighted combination of the variation measured by the first-order derivatives across the primary sampling units within the same

stratum (Kalton, 1983a). Muthen and Satorra (1995) extended this linearization method to model the covariance structure in the complex survey data. The mean is a $p \times 1$ vector of simple weighted means. To calculate the asymptotic sample covariance matrix across all stratum and primary sampling units is equivalent to calculate a weighted covariance matrix across all elements in the data. The vector of parameter estimates is calculated by minimizing the likelihood function:

$$\ln L = \sum_{i=1}^{II} \sum_{j=1}^{hi} \sum_{k=1}^{nij} \sum_{l=1}^{nijk} w_{ijkl} f(x_{ijkl} | \hat{\theta}) \qquad (18)$$

where x_{ijkl} is the observation for the *lth* student in the *kth* school within the *ith* strata and *jth* primary sampling units and $f(y_{ijkl} | \hat{\theta})$ is the distribution for x_{ijkl} given the parameter $\hat{\theta}$. The standard error for the complex sampling design data is calculated via the asymptotic covariance matrix, $a\text{cov}(\hat{\theta}) = I^{-1} \Gamma I^{-1}$. I is the information matrix and Γ is a measure of the pooled variability across *ith* strata and *jth* primary sampling units. If there is no effect due to complex sampling, the elements on the diagonal of the final resulting matrix will be one and the scaling factor will also become one. This method mainly includes replacing sampling covariance matrix with the weighted sample covariance matrix and replacing fisher information with a sandwich estimator of variance (Bollen, Uueller, & Oberski, 2013).

Stapleton (2006) has been the only study that compared three methods. It is hard to make a conclusion about the advantage and disadvantage of the methods based on one study. Except for the above methods, balanced repeated replication, jack-knife and bootstrapping techniques are applicable (Thomas & Heck,2001).

2.4 STUDIES COMPARING SINGLE-LEVEL CFA WITH MCFA

There have been very limited simulation studies comparing the single-level CFA with multilevel CFA in the clustering data. The results from these studies are briefly summarized as the following: Pornprasertmanit et al. (2014) compared single-level CFA with two correlated factors and two-level CFA with two correlated factors at both levels. The two-level CFA was demonstrated to fit the data better when the data was simulated with clustering structure. It was found that the standardized factor loadings were not biased when the factor loadings across levels were simulated to be the same. When the factor loadings were simulated to be different across levels, the factor loading estimated from two-level CFA model and single-level CFA model differed. When the average of the item ICC was less than 0.25, the absolute difference of the factor loading was within a reasonable range. The factor loading was overestimated when the between-level communality was high and underestimated when the between-level communality was low. When the between-level communality was high, the standard error (SE) of factor loading estimated using disaggregated analysis was negatively biased. When the between-level communality was low, SE of factor loading estimated using disaggregated analysis was positively biased. When the factor correlations across levels were simulated to be the same or ICC was smaller than 0.25, the absolute bias was within 0.05. Also, the standard error of the factor correlation tended to be higher when the between-level communality was lower, and vice versa.

Julian (2001) compared the single-level CFA model with four factors, the two-level CFA model with four factors at both levels, the model with four factors at the within-level and two factors at the between-level, and the model with four factors at the within-level and five factors at the between-level. It was demonstrated that model parameters (i.e. factor loading, variance,

and covariance) tended to be overestimated, corresponding standard errors tended to be underestimated, and chi-square statistics was inflated when the two-level CFA was estimated using the single-level CFA. The relative bias of factor loading and standard errors were not affected by different factorial structures. The relative bias of standard error increased as the ratio of groups to group members decreased when ICC was larger than 0.05. Except when ICC was as large as 0.45, relative bias of parameter estimates was not affected by the ratio of groups to group members. When ICC was 0.45, the relative bias of the factor loading in three models increased from 0.02 to 0.14, from 0.02 to 0.10, and from 0.04 to 0.13 when the ratio of groups to group members decreased from 20 to 0.2. The ignoring of clustering could be neglected when ICC was smaller than 0.05 and group size was small. The relative bias of factor covariance was affected by different factorial structure. The relative bias of factor covariance was smaller when the factorial structures across levels were the same or when the between-level was simpler than the within-level; the relative bias of the factor covariance was larger when the between-level was more complex than the within-level.

Stochl et al. (2015) simulated a five-factor model at the within-level and it was proved that ignoring the clustering would underestimate the factor loading, related standard error, and item threshold for the multilevel data using WLSMV and the relative bias increased with ICC. But the estimate of factor correlation was almost unbiased even when ICC was large.

Wu and Kwok (2012) compared the single-level design-based approach with model-based two-level approach in the clustering data. It was found that factor loadings estimated from the single-level CFA model were accurate when the factorial structures were the same at the within and between-level but were seriously biased when the factorial structures were different across levels. Factor loadings on the single item were still unbiased when the within-level

factorial structure was more complex than that of the between-level. But factor loadings were seriously biased when the between-level factorial structure was more complex than that of the within-level and the degree of bias increased with the increase of ICC. The two-level model accurately estimated the factor variance, covariance, and residual variances in three factorial structures. Single-level CFA model could not accurately estimate the factor variances and residual variances when the between-level model was more complex than the within-level model. Factor variances were underestimated and residual variances were overestimated using the single-level CFA model. Even when the between-level and within-level model were the same or when the within-level model was more complex than the between-level model, the factor covariance estimated using the single-level model was twice the factor covariance of the true model.

Based on the limited simulation studies, the following conclusions could be made: First, the accuracy of the disaggregated parameter estimates (factor loading, factor correlation, and residual variances) and related standard errors were affected by the factorial structure of the true model. Second, the disaggregated factor loading, factor correlation, and related standard errors could be overestimated or underestimated depending on the between-level communality of the true model. Third, the accuracy of parameter estimates and related standard errors were affected by ICC. When the item-level ICC increased, the relative bias of parameters estimated from the single-level CFA using adjusted standard error increased. Fourth, the ratio of the number of groups to number of group members affected the accuracy of standard errors of parameter estimates. The effect from sample size was not as large as the effect from ICC.

In addition to the previous simulation studies, Stapleton, Yang, and Hancock (2016) stated that when a construct measured had cluster-level dependency, the single-level model with

a correction of standard error was also appropriate. Complex single-level CFA model and MCFA model were compared. The study used $\chi 2$, CFI, RMSEA, and SRMR to examine the model fit. It was found that two-level model did not perform well when the ICC was low or the within-cluster persons were smaller than 50 based on the $\chi 2$ while the single-level model with designed-based adjusted standard error performed well.

2.5 LITERATURE REVIEW OF APPLIED RESEARCH OF MCFA

Using the database of PsycINFO, among 15 empirical studies about MCFA using likert-type questionnaires (Breevaart, 2012;Brondino, et al., 2013; Dedrick & Greenbaum, 2010; Dyer, 2005;Grilli & Rampichini, 2007; Gajewski & Boyle,2013; Greenbaum et al., 2011; Haenens, et al., 2012 ; Klangphahol et al., 2010; Little, 2013;Ryu,2013; Whitton & Fletcher, 2014; Wu, 2009; Zimprich et al., 2005;Zhang & Wang, 2005), the following findings were obtained:

1. The between-level factorial structure was the same as or simpler than the within-level factorial structure.
2. The item-level ICC ranged from 0.028 to 0.55. Most of item-level ICCs were larger than 0.10.
3. In majority of the studies, the factor loadings at the between-level were higher than those of the within-level.
4. The factor correlation at the between-level was substantially higher than that of the within-level in most models. Dedrick and Greebaum(2010) suggested to use the one-factor model at level-2 when the factor correlation at level-2 was as high as 0.9.

3.0 METHODOLOGY

The main goal of the study is to examine and compare the performance of three approaches for CFA with multilevel data: model-based approach using two-level CFA, single-level CFA with standard errors adjusting for clustering effect, and the single-level CFA with normal standard error. A Monte Carlo study was conducted to examine the conditions under which the bias resulting from using the single-level model in the multilevel data is consequential. The manipulated simulation design factors include: 1) number of cluster members, 2) number of clusters, 3) factor correlation, 4) factor loading, and 5) Item ICC. According to previous empirical studies, the within-level model could be more complex, simpler or the same as the between-level model. In this study, two models with different factorial structures were examined: the two-level model with correlated factors at both levels and the two-level model with the correlated factor at the within-level and one factor at the between-level. MPLUS is used to generate the data based on the multilevel covariance structure modeling. Data analysis are also performed in MPLUS.

The following sections will first introduce manipulated simulation design factors including factorial structure (number of factors at two levels, factor loading, and factor correlation at two levels), ICC, and sample size (number of cluster members and number of clusters), and then present evaluation criteria, data generation model and data generation validation.

3.1 SIMULATION DESIGN FACTORS

3.1.1 Factorial Structure

For the studies comparing the single-level CFA with MCFA, previous simulation studies adopted same or different factorial structures at the between and within-level. The between-level factorial structure can be more complex or simpler than the within-level factorial structure. Julian (2001) incorporated the four-factor model at the within-level and four factor or two factor or five factor model at the between-level. All parameters were simulated on the unstandardized scale. Wu and Kwok (2012) used a three-factor model either at the within or between-level or both. The factor correlations among factors were set to be 0.3, the single item factor loadings were all set to be 0.8, and the complex item factor loadings were all set to be 0.4. The study constrained the factor loadings across the levels to be the same. Pornprasertmanit, Lee, and Preacher (2014) simulated the two-factor model with the same factorial structure across levels. At the between-level, the factor correlation was simulated to be 0.2, 0.5, or 0.8. At the within-level, the factor correlation was fixed to be 0.5. The factor loadings were all set to be 0.7 at the within-level and were all set to be 0.49, 0.7, or 0.86 at the between-level. Stochl et al. (2015) simulated data using a five-factor CFA model at the within-level and no common factors at the between-level. All factor loadings were set to be 0.7 and factor correlations were set to be 0.4.

Pornprasertmanit et al. (2014) has been the only study to simulate the two correlated factor model at the within-level. The model with two correlated factors is the simplest form in the educational and psychological setting. Therefore, this study used the two correlated factors at the within-level or both levels, and each factor has five ordinal item indicators. From the study of Wu and Kwok (2012), the parameter estimates were accurate if the clustering was ignored

when the factorial structures were the same across levels. In order to compare the effect of different factorial structures on the relative bias of the parameters in ignoring the clustering, the data generation models of the current study were the 1) two-level CFA model with two correlated factors at both levels (W2B2); 2) two-level CFA model with two correlated factors at the within-level and the one factor at the between-level (W2B1). The factor correlation at the within and/or between-level were set to be 0.3 or 0.6 to represent small and medium correlation respectively. In Model 1 (W2B2), the factor correlation at the within level was lower (.3 vs .6) or the same (.3 vs .3).

According to Pornprasertmanit et al. (2014), the distribution of the factor loading did not significantly contribute to the accuracy of parameter estimates, and thus all items were constrained to have the same factor loading at the within- or between-level. In applied research, factor loading of 0.6 is considered moderate and factor loading of 0.8 is considered high. For Model 1 (W2B2), factor loadings at the within-level was set to be higher, lower, or equal to those at the between-level. More specifically, factor loading was set to be 0.8 at within-level versus 0.5 at the between-level, or 0.5 at the within-level versus 0.8 at the between-level, or 0.5 at two levels.

3.1.2 Item ICC

Item ICC measures the proportion of variance in an item that is due to the between-group clustering, which can be calculated from parameters in two-level CFA. According to Julian (2001), the item-level ICC was calculated as the following:

$$\rho_{jj} = \frac{[\sum_B]_{jj}}{[\sum_B]_{jj} + [\sum_W]_{jj}}$$

$$\rho_{jj} = \frac{[\lambda_j^2 * \Psi_B + \Theta_B]}{[\lambda_j^2 * \Psi_B + \Theta_B] + [\lambda_j^2 * \Psi_W + \Theta_W]} \qquad (21)$$

where $\sum_B$ and $\sum_W$ are between-level and within-level covariance, respectively. *jj* refers to the *jth* diagonal element of the covariance matrix of variable *j* . λ_j is the factor loading of the *jth* variable, Ψ_B and Ψ_W are the between-level and within-level covariance matrix of the factors, and Θ_B and Θ_W are the between-level and within-level matrix of residual variances (Julian,2001). Since ICC is to assess the homogeneity within the cluster, it is an important factor. The large ICC states that the degree of clustering is high.

From the previous empirical studies of MCFA, item ICC ranged from 0.028 to 0.55. ICC smaller than 0.05 was considered negligible in studying clustering effect (Stapleton, 2002). In the simulation studies, Julian (2001) used ICC of 0.05, 0.15, and 0.45; Pornprasertmanit, et al. (2014) used ICC of 0.05, 0.25, 0.50 , 0.75, and 0.95; Wu and Kwok(2012) used ICC of 0.1 and 0.5; Stochl et al. (2015) simulated 11 levels of ICCs ranging from 0.001 to 0.390. Based on the national representative sample, Hedges and Hedberg (2007) stated that the average ICC was 0.22 for all schools and 0.19 for schools of low socioeconomic status. Hox and Maas (2001) stated that ICCs were below 0.2 in most educational research while ICCs were above 0.33 when group characteristics such as socio-econometric status was studied. According to Pornprasertmanit, et al. (2014), ICCs of indicators of the same factor could be same or different.

Julian (2001), Pornprasertmanit, et al. (2014) , and Wu and Kwok (2012) all manipulated item ICC in the simulation study instead of the latent factor ICC. In the current study, two

conditions of ICCs were set: 0.25 and 0.45. According to Pornprasertmanit et al. (2014), the distribution of ICCs did not significantly contribute to the accuracy of parameter estimates, thus the distribution of ICCs was not varied for this study.

3.1.3 Sample Size

From previous empirical studies of MCFA, within-level sample size ranged from 6 to 72,899 and between-level sample size ranged from 25 to 4,783. In the study of Julian (2001), the ratio of the number of clusters to the number of cluster members was set to be 100/5, 50/10, 25/20, and 10/50. In the study of Pornprasertmanit, Lee, and Preacher (2014), the ratio of the number of clusters to the number of cluster members was set to be 100/5, 10/50, 400/20, and 40/200. Wu and Kwok (2012) selected the cluster sizes at 10, 50, and 200, and the cluster number at 50,150, and 300. In the MCFA study of Hox and Maas (2001), it was found that the smallest sample size for the accurate parameter estimate was 10 observations within 50 clusters. Based on the results of previous studies, the ratio of the size of between-cluster to the within-cluster in the current study was 50/10, 20/25,250/10, and 100/25. When the within-cluster sample sizes or the total sample sizes are the same, the effect of sample size on the accuracy of parameter estimates can be better investigated. In the current study, the total sample size for the small sample was 500 and the total sample size for the large sample was 2500.

3.1.4 Model Estimation

Based on Section 2.0, WLSMV estimates parameters more accurately than ML and MLR in CFA of ordinal items considering the robustness of its performance to the two, three, and four-

category items. The number of indicators did not affect WLSMV's performance. The performance of WLSMV was not affected by the distribution of the thresholds while the performance of MLR was affected by the distribution of the thresholds in the ordinal data. This study intends to examine the accuracy of parameter estimation of factor loading, factor correlation, and residual variances using WLSMV.

In summary, there were 4 (combination of within-level and between-level sample size) ×2 (combination of between-level and within-level correlation) ×3 (combination of between-level and within-level factor loading) ×2 (ICC) =48 conditions for the first MCFA model in the original design. To make all residual variances of the between-level positive, the combination of within-level factor loading of 0.5 and between-level factor loading of 0.8 could not be set up with ICC of 0.25. Therefore, finally there were 40 conditions for W2B2 model. Similarly, there were 4 (combination of within-level and between-level sample size) ×2 (within-level correlation) ×3 (combination of between-level and within-level factor loading) ×2 (ICC)=48 conditions for the second MCFA model in the original design. But finally 40 conditions were adopted for W2B1 model for the same reason as in W2B2 model. In each condition, 100 datasets were generated, and each generated dataset was analyzed with WLSMV by three models: two-level CFA model, single-level CFA model using normal error, and single-level CFA model using complex error. Table 2 lists the simulation factors for two data generation models of different factorial structures.

Table 2. Simulation factors for three models of different factorial structures

	Within-Cluster /Between-Cluster Factor Correlation(2)	within-cluster/ between-cluster sample size (4)	ICC (2)	within-cluster and between-cluster factor loading(3)	Estimation Method(1)
W2B2 model	0.3/ 0.6; 0.3/ 0.3;	10/50, 25/20 ,10/250 , and 25/100	0.25 and 0.45	0.5 and 0.8; 0.5 and 0.5; 0.8 and 0.5	WLSMV
W2B1 model	0.3/; 0.6/	10/50, 25/20 ,10/250 , and 25/100	0.25 and 0.45	0.5 and 0.8; 0.5 and 0.5; 0.8 and 0.5	WLSMV

3.2 EVALUATION CRITERIA

Before examining the outcome variables, it is necessary to look at the rates of improper solutions across simulation conditions. An improper solution means nonconvergence or a solution that converges but there exits one or more out-of-bound parameters. The improper solutions need to be removed from the final analysis (Flora & Curran, 2004). In this study, the improper solution could be that the residual variance is negative.

The outcome variables to be investigated include model fit indices, the relative bias of the factor loading, residual variance, and their related standard errors (Hoogland & Boomsma,

1998). Chi-square, RMSEA, CFI, TLI, and SRMR are used to evaluate the model fit. Relative bias of parameter estimate (including factor loading and residual variance) is calculated as:

$$B(\hat{\theta}_i) = \frac{\bar{\hat{\theta}}_i - \theta_i}{\theta_i} \qquad (24)$$

where θ_i is the true value of the ith parameter, and $\bar{\hat{\theta}}_i$ is the mean of the *ith* parameter estimates across the 100 replications. Relative bias less than 5% is the trivial bias, between 5% and 10% is the moderate bias, and greater than 10% is the substantial bias (Wallentin, Joreskog, & Luo,2010).

Following Flora and Curran (2004), the pooled mean of the factor loading at each level is examined instead of examining the factor loading of each individual item.

$$\text{Pooled Mean} = n^{-1}\sum_{i=1}^{n}\overline{\hat{\lambda}_i} \qquad (22)$$

where n is the number of indicators and $\overline{\hat{\lambda}_i}$ is the mean across replications of each factor loading. It is to first calculate the mean of factor loading across replications of each cell. Then the pooled mean of the factor loading of all items are calculated. In this study, for the within-level and between-level, the pooled mean of the factor loading is calculated across 10 items, respectively. The pooled standard deviation of the factor loading is calculated.

$$\text{Pooled SD} = \sqrt{n^{-1}\sum_{i=1}^{n} VAR(\hat{\lambda}_i)} \qquad (23)$$

where $VAR(\hat{\lambda}_i)$ is the sample variance of each factor loading across replications. In this study, the pooled standard deviation is calculated across 10 items for the within-level and between-level, respectively.

3.3 DATA GENERATION

The first data generation model was two-level CFA model (see Figure 1) with correlated factors at both levels. The continuous item scores were first generated based on a two-level CFA model with specifically defined ICC, factor loading, factor variance, and item residual variance. The continuous item scores were then transformed into five-category ordinal data by incorporating threshold parameters. For the five-category data, four thresholds were needed.

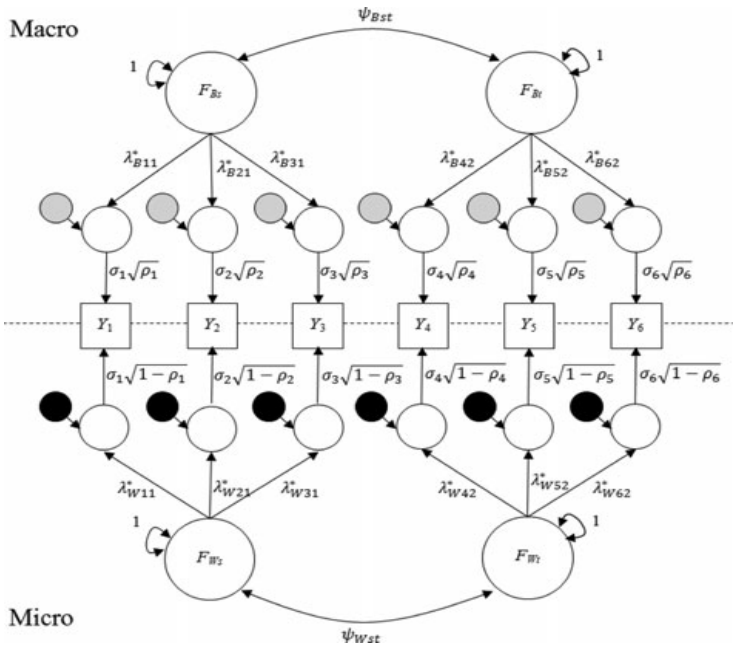

Figure 1. Two-Level CFA Model with Correlated Factors at Both Levels

Figure 1 is a two-level CFA model with three indicators loading on each factor of each level. According to Figure 1, it can be told that the latent variable at the within-level is explained by a factor at the within-level and the circle represents the residual variance of the variable that cannot be explained by the factor. The similar rule mechanism is applied to the between-level.

Item thresholds were adopted from the empirical data on the school engagement scale (Wang, et al., under review), which has 38 items that are all ordinal on a 5-point Likert scale:

ranging from 1 (no at all like me) to 5 (very much like me). Each student reported on his or her own (a) behavioral engagement, using four items expressing their effort and investment while participating in school activities; (b) behavioral disengagement, using eight items expressing their avoidance, defiance, opting out of school activities; (c) emotional engagement, using five items expressing their positive emotions in school; (d) emotional disengagement, using 5 items expressing their negative emotions in school; (e) cognitive engagement, using five items expressing their persistence, planning, and strategy use during school activities; (f) cognitive disengagement, using two items expressing their lack of perseverance and withdrawal from school activities; (g) social engagement, using five items expressing their collaboration with peers and engagement in relationships in school; and (h) social disengagement, using four items expressing their lack of interest and involvement in social interactions and relationship with others in school. School engagement was conceptualized to consist of school engagement and disengagement dimensions. Five items with high factor loading on the engagement factor and disengagement factor were selected respectively. The factor loading on the general factor ranged from 0.404 to 0.792. The correlation between the two dimensions was .4. Table 3 presents the item thresholds of the ten selected items, with the first five on school engagement, and the last five on school disengagement.

Table 3. Item thresholds of 10 items

	b1	b2	b3	b4
Item 1	-1.464	-1.298	-0.258	0.682
Item 2	-1.410	-1.273	-0.252	0.719
Item 3	-1.829	-1.647	-0.632	0.261
Item 4	-1.591	-1.422	-0.332	0.605
Item 5	-1.150	-0.739	-0.135	0.470
Item 6	-2.081	-1.884	-1.063	-0.262
Item 7	-1.274	-0.980	0.069	0.929
Item 8	-1.384	-1.145	-0.082	0.886
Item 9	-0.509	0.049	1.051	1.653
Item 10	-0.584	0.561	1.319	1.793

The minimum number of items loading on one factor was three to make a factor identifiable. Flora and Curran (2004) stated that a factor with 5 item indicators was commonly seen in practice. In this study, the number of items loading on the first factor and second factor were 5 at both the within and between-level.

In the first data generation model, the within-level factor loading was set to be 0.5 and the between-level factor loading was set to be 0.8. The factor correlations of both levels were set to be 0.3. The setup of factor loading and factor correlation were not affected by the item-level ICC. According to equation (10), the between-level covariance matrix was:

$$\Sigma_B = \begin{bmatrix} 0.8 & 0 \\ 0.8 & 0 \\ 0.8 & 0 \\ 0.8 & 0 \\ 0.8 & 0 \\ 0 & 0.8 \\ 0 & 0.8 \\ 0 & 0.8 \\ 0 & 0.8 \\ 0 & 0.8 \end{bmatrix} \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \begin{bmatrix} 0.8 & 0.8 & 0.8 & 0.8 & 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.8 & 0.8 & 0.8 & 0.8 & 0.8 \end{bmatrix} + \begin{bmatrix} 0.383 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 0.383 \end{bmatrix}$$

(26)

According to equation (11), the within-level covariance matrix was:

$$\Sigma_W = \begin{bmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0.5 & 0 \\ 0.5 & 0 \\ 0.5 & 0 \\ 0 & 0.5 \\ 0 & 0.5 \\ 0 & 0.5 \\ 0 & 0.5 \\ 0 & 0.5 \end{bmatrix} \begin{bmatrix} 1 & 0.3 \\ 0.3 & 1 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 & 0.5 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix} + \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix}$$

(27)

The between-level residual variances depended on the item ICC. According to equation (21), when the between-level residual variance of item *j* was 0.383, the within-level residual variance of item j was 1, the factor loading of the within-level was 0.5, the factor loading of the between-level was 0.8, the factor variance of within- and between-level were both 1, the ICC was calculated as the following:

$$\rho_{jj} = \frac{[\lambda_j^2 * \Psi_B + \Theta_B]}{[\lambda_j^2 * \Psi_B + \Theta_B] + [\lambda_j^2 * \Psi_W + \Theta_W]}$$
$$= \frac{[0.8^2 * 1 + 0.383]}{[0.8^2 * 1 + 0.383] + [0.5^2 * 1 + 1]} \qquad (28)$$
$$= 0.45$$

In Mplus, the probit link is available for WLSMV. In the probit regression, residual is normally distributed with variance of 1. Thus, the within-level residual variance , Θ_W, was 1. The within-level factor variance and between-level factor variance were set to be 1 so that the factor loadings are directly the correlation between factors and items, following the recommendation of the use of standardized loading in Pornprasertmanit, et. al. (2014).

The second data generation model was two-level CFA model with two factors at the within-level and one factor at the between-level. The item thresholds were the same as those in the first simulation model. The factor correlation between two factors at the within-level was 0.3 or 0.6. The factor loading, factor variances, and residual variances were set the same as those in W2B2 model.

3.4 DATA VALIDATION

3.4.1 W2B2 model: Two-level CFA with correlated factors at both levels

In the data validation part, the data generation model was the model with correlation of 0.3 at both levels and the factor loading of 0.5 at the within-level and 0.8 at the between-level. The item ICCs were all set to be 0.45. The number of clusters was set to be 100 with 25 subjects within the cluster. Table 4 presents relative frequency of the items among 2500 observations,

several items were right skewed such as item 1 and several items were left skewed such as item 9 and item 10. Most items had smallest frequency at the second category.

Table 4. Item frequency distribution

Relative Frequency	Category 1	Category 2	Category 3	Category 4	Category 5
Item 1	0.156	0.033	0.241	0.252	0.318
Item 2	0.160	0.028	0.250	0.263	0.299
Item 3	0.102	0.026	0.220	0.262	0.390
Item 4	0.134	0.025	0.235	0.263	0.343
Item 5	0.179	0.087	0.148	0.163	0.423
Item 6	0.076	0.028	0.131	0.201	0.564
Item 7	0.176	0.066	0.264	0.246	0.248
Item 8	0.174	0.044	0.267	0.239	0.275
Item 9	0.363	0.149	0.253	0.116	0.119
Item 10	0.337	0.325	0.166	0.070	0.102

A two-level CFA was conducted. The chi-square *p* value was 0.884, CFI was 1.00, TLI was 1.00, RMSEA was 0, SRMR was 0.02 for within-level, and 0.085 for the between-level. The ICCs of items ranged from 0.342 to 0.462. The model fit was generally good. The SRMR_B was a little above the cutoff criterion value. The unstandardized factor loadings of the within-level ranged from 0.454 to 0.597 and the average was 0.52, which was quite close to the true value of 0.5. The within-level correlation was 0.274.

The between-level factor loadings ranged from 0.509 to 0.953 and the average was 0.747, which did not deviate a lot from the true value of 0.8. The between-level factor correlation was 0.229 and it deviated from the true value of 0.3. The average of the between-level residual variance was 0.389, which was quite close to the true value of 0.383. The recovery of the factor loadings and the factor correlation at the within-level was better than that of the between-level.

Table 5. Parameter Estimates from the two-level model

	Within-level factor correlation True: 0.3	Within-level Factor loading True: 0.5	Between-level factor correlation True : 0.3	Between-level factor loading True: 0.8	Between-level residual Variance True: 0.383	ICC True: 0.45
W2B2 model:						
Average	0.274	0.52	0.229	0.747	0.389	
Relative Bias:	-0.086	0.04	-0.236	-0.066	0.016	

4.0 RESULTS

The results were presented in the following order: 1) the rate of improper solutions, 2) model fit indices including the significance level (*p* value) of chi-square statistic, SRMR at the within level (SRMR_W) and the between level (SRMR_B), RMSEA, CFI, and TLI for each model, 3) parameter estimates for each model 4) a comparison of three models in terms of model parameter estimates and standard errors. In each section, results were presented for the W2B2 model first and then the W2B1 model.

4.1 RATES OF IMPROPER SOLUTIONS

Table 6 presents the rates of improper solutions obtained with WLSMV. All replications converged in all conditions. Improper solutions resulted from the negative residual variances at the between-level in the two-level CFA. In W2B2 model, totally 498 cases had the negative residual variances and the improper solutions mostly occurred in the small sample sizes especially when the between-level sample size is small. Negative residual variances appeared most when the factor loadings were 0.5 at both levels and ICC was 0.25.

W2B1 model had obviously fewer cases with the negative residuals variances than W2B2 model. In W2B1 model, there were totally 58 cases had the negative residual variances and most occurred in the small sample sizes regardless of the small within-level sample size or small

between-level sample size. The condition that had the most cases of the negative residual variances was when the within-level and between-level factor loading were both 0.5, ICC was 0.25, the within-level sample size was 25 and the between-level sample size was 20. 7 and 10 cases had the negative residual variances for different factor correlations, respectively. When the within-level and between-level factor loading were both 0.5, ICC was 0.25, the within-level sample size was 10, and the between-level sample size was 50, 4 and 6 cases had the negative residual variances, respectively.

Table 6. Rates of improper solutions for cases of negative residuals

	W2B2 Model					W2B1 Model	
	Within-level / Between-level Factor Loading	ICC	Sample Size	Within-level /Between-level Factor Correlation	Cases with Negative Residuals	Within-level Factor Correlation	Cases with Negative Residuals
1	0.5/ 0.8	0.45	50(10)	0.3/0.3	8	0.3	0
2				0.3/0.6	4	0.6	0
3			20(25)	0.3/0.3	38	0.3	0
4				0.3/0.6	34	0.6	2
5			250(10)	0.3/0.3	0	0.3	0
6				0.3/0.6	0	0.6	0
7			100(25)	0.3/0.3	0	0.3	0
8				0.3/0.6	0	0.6	0
9	0.5/0.5	0.45	50(10)	0.3/0.3	13	0.3	1
10				0.3/0.6	6	0.6	0
11			20(25)	0.3/0.3	33	0.3	4
12				0.3/0.6	27	0.6	4
13			250(10)	0.3/0.3	0	0.3	0
14				0.3/0.6	0	0.6	0
15			100(25)	0.3/0.3	1	0.3	0
16				0.3/0.6	0	0.6	0
17	0.5/0.5	0.25	50(10)	0.3/0.3	26	0.3	4
18				0.3/0.6	21	0.6	6
19			20(25)	0.3/0.3	57	0.3	7
20				0.3/0.6	51	0.6	10
21			250(10)	0.3/0.3	0	0.3	0
22				0.3/0.6	0	0.6	0
23			100(25)	0.3/0.3	0	0.3	0
24				0.3/0.6	0	0.6	0
25	0.8/0.5	0.45	50(10)	0.3/0.3	24	0.3	1
26				0.3/0.6	17	0.6	1
27			20(25)	0.3/0.3	32	0.3	4
28				0.3/0.6	23	0.6	1
29			250(10)	0.3/0.3	0	0.3	0
30				0.3/0.6	0	0.6	0
31			100(25)	0.3/0.3	2	0.3	0
32				0.3/0.6	0	0.6	0
33	0.8/0.5	0.25	50(10)	0.3/0.3	10	0.3	2
34				0.3/0.6	7	0.6	2
35			20(25)	0.3/0.3	40	0.3	4
36				0.3/0.6	24	0.6	5
37			250(10)	0.3/0.3	0	0.3	0
38				0.3/0.6	0	0.6	0
39			100(25)	0.3/0.3	0	0.3	0
40				0.3/0.6	0	0.6	0

4.2 EVALUATION OF MODEL FIT

4.2.1 W2B2 Model

Table 7 presented the proportion of the model fit statistics that met the cut-off criteria for the two-level model, complex model, and normal model in the W2B2 model. According to the recommended cut-off criteria, chi-square *p* value larger than 0.05 is considered good fit. According to the recommended cut-off criteria, the two-level model is considered good fit when SRMR_W and SRMR_B are smaller than 0.08. According to recommend cut-off criteria, the model is considered good fit when RMSEA is smaller than 0.06, CFI and TLI are larger than 0.95.

4.2.1.1 Chi-square test statistics

Looking at *p* values of the chi-square of the two-level model and complex model, at least 90% of the *p* value in the two-level model and complex model were above 0.05, indicating the good fit of the models. When the normal model was used to estimate the clustered data, proportion of the *p* value that was above 0.05 was very low especially under the condition of high ICC. The influence of the high ICC was larger than the influence of the small sample size.

4.2.1.2 SRMR_W and SRMR_B

Looking at the SRMR_W and SRMR_B of the two-level model, all SRMR_Ws were smaller than 0.08, indicating the good model fit of the within-level model. At least 81.6% of the SRMR_B were smaller than 0.08 when sample size was large. The between-level model had

poorer fit when sample size was small. In most conditions with the small sample size, SRMR_B was smaller than 0.08 in less than 5% of the replications.

SRMR_W and SRMR_B were good fit indices to evaluate the fit of the two-level model while other fit indices previously developed for the single-level model were not as good as them. In this study, SRMR_W and SRMR_B found that the within-level fit the data well across all conditions while the between-level was affected by the small sample size. This phenomenon could not be detected by the chi-square p value. p value of the chi-square statistics seemed to be more affected by the ICC.

4.2.1.3 RMSEA, CFI, and TLI

Looking at RMSEA, the two-level model fit the data well. Looking at CFI and TLI of the two-level model, the proportion that indices meeting the criteria was not as high as that of the chi-square *p* value in the small sample size. But generally model still fit well looking at these indices. Looking at RMSEA of the complex model, the complex model fit the data well. Looking at CFI and TLI of the complex model, the complex model fit the data well when sample size was large. When sample size was small, the proportion that the index was within the cut-off criteria was as low as 63.8% when the sample size was 50(10) and as low as 43.3% when sample size was 20(25). When the sample size was small, the complex model fit the data better when ICC was low than when ICC was high.

Looking at RMSEA, CFI, and TLI of the normal model, the normal data fit the data better when the sample size was large than when the sample size was small. When the sample size was small, the normal model fit to the data a little better when the between-level sample size was 50 than when the between-level sample size was 20. The proportion of the indices within the cut-off criteria was lower in the 0.8 0.5 0.45 and 0.5 0.5 0.45 than in the 0.5 0.8 0.45

especially when the sample size was 20(25). Also, when the sample size was small, the normal model fit the data better when ICC was low than when ICC was high. Similar to the findings from the chi-square *p* value, the normal model did not fit the data as well as the two-level model and the complex model.

In general, it was found that when the complex model was used to estimate the clustered data, the model still fit the data looking at *p* value and RMSEA and model fit better with the large sample size looking at CFI and TLI. When the normal model was used to estimate the two-level data, the model fit was not good as good as the complex model especially when ICC was high and sample size was small (Table 7).

Table 7. Proportion of the model fit statistics meeting the cut-off criteria for two-level model, complex model, and the normal model in W2B2 model

	Sample Size	**ICC**	**Within-level and Between-level Factor Loading**	**Within-level and Between-level Factor Correlation**	**Model**	**Chi-square *p***	**RMSEA**	**CFI**	**TLI**	**SRMR_W**	**SRMR_B**
1	50(10)	0.45	0.5/0.8	0.3/0.3	1	0.989	1	0.967	0.902	1	0.326
					2	0.967	1	0.989	0.957		
					3	0.359	0.989	0.967	0.924		
2				0.3/0.6	1	1	1	0.448	0.396	1	0.781
					2	0.938	1	1	0.958		
					3	0.292	1	1	0.938		
3	50(10)	0.45	0.5/0.5	0.3/0.3	1	0.989	1	0.851	0.690	1	0
					2	0.954	1	0.759	0.678		
					3	0.057	0.908	0.138	0.069		
4				0.3/0.6	1	0.979	1	0.809	0.681	1	0
					2	0.915	1	0.638	0.638		
					3	0.021	0.830	0.106	0.043		
5	50(10)	0.25	0.5/0.5	0.3/0.3	1	0.946	1	0.838	0.743	1	0.014
					2	0.973	1	0.986	0.959		
					3	0.824	1	1	0.959		
6				0.3/0.6	1	0.899	1	0.759	0.633	1	0.089
					2	0.975	1	0.962	0.962		
					3	0.810	1	0.987	0.975		
7	50(10)	0.45	0.8/0.5	0.3/0.3	1	0.987	1	1	1	1	0
					2	0.974	1	0.947	0.855		
					3	0.053	0.868	0.395	0.237		

Table 7 (continued)

	Sample Size	ICC	Within-level and Between-level Factor Loading	Within-level and Between-level Factor Correlation	Model	Chi-square *p*	RMSEA	CFI	TLI	SRMR_W	SRMR_B
8				0.3/0.6	1	1	1	1	1	1	0
					2	0.952	1	0.964	0.880		
					3	0.072	0.867	0.446	0.253		
9	50(10)	0.25	0.8/0.5	0.3/0.3	1	0.956	1	1	1	1	0
					2	0.967	1	1	1		
					3	0.711	1	1	1		
10				0.3/0.6	1	0.957	1	1	0.978	1	0.011
					2	0.957	1	1	1		
					3	0.688	1	1	1		
11	20(25)	0.45	0.5/0.8	0.3/0.3	1	1	1	1	1	1	0.081
					2	0.935	1	0.790	0.742		
					3	0.016	0.677	0.645	0.435		
12				0.3/0.6	1	1	1	1	1	1	0.212
					2	0.955	1.000	0.848	0.803		
					3	0.030	0.742	0.818	0.652		
13	20(25)	0.45	0.5/0.5	0.3/0.3	1	1	1	1	1	1	0
					2	0.955	1	0.478	0.433		
					3	0	0.119	0	0		
14				0.3/0.6	1	1	1	1	1	1	0
					2	0.918	1	0.507	0.438		
					3	0	0.151	0	0		

Table 7 (continued)

	Sample Size	ICC	Within-level and Between-level Factor Loading	Within-level and Between-level Factor Correlation	Model	Chi-square *p*	RMSEA	CFI	TLI	SRMR_W	SRMR_B
15	20(25)	0.25	0.5/0.5	0.3/0.3	1	1	1	0.907	0.884	1	0
					2	0.953	1	0.884	0.814		
					3	0.651	1	0.907	0.791		
16				0.3/0.6	1	1	1	0.837	0.816	1	0.082
					2	0.939	1	0.878	0.837		
					3	0.571	1	0.878	0.816		
17	20(25)	0.45	0.8/0.5	0.3/0.3	1	1	1	1	1	1	0
					2	0.941	1	0.574	0.500		
					3	0	0.044	0	0		
18				0.3/0.6	1	1	1	1	1	1	0
					2	0.948	1	0.571	0.519		
					3	0	0.052	0.013	0		
19	20(25)	0.25	0.8/0.5	0.3/0.3	1	1	1	1	1	1	0
					2	0.967	1	1	0.967		
					3	0.250	0.983	0.983	0.933		
20				0.3/0.6	1	0.974	1	1	1	1	0

4.2.2 W2B1 Model

Table 8 presented the proportion of the model fit statistics that met the cut-off criteria for the two-level model, complex model, and normal model in the W2B1 model.

4.2.2.1 Chi-square test statistics

Looking at *p* values of the chi-square of the two-level model and complex model, all *p* values of two-level model and complex model were far above 0.05, indicating the good fit of the two-level model and complex model. When the normal model was used to estimate the clustered data, all *p* values of chi-square were larger than 0.05 when ICC was 0.25. It indicated that when ICC was low, the normal model could still fit the data that had clustering. When ICC was 0.45, all *p* values were smaller than 0.05 with the within/between level factor loading of 0.5 0.5 and 0.8 0.5.

4.2.2.2 SRMR_W and SRMR_B

Looking at the SRMR_W and SRMR_B of the two-level model, all SRMR_Ws were smaller than 0.08, indicating the good model fit of the within-level model. At least 87% of the replications had SRMS_Bs smaller than 0.08 for the large sample size, indicating the good model fit of the between-level model when the sample size was as large as 2500. However, the between-level model did not fit the data that well when the sample size was small.

SRMR_W and SRMR_B were good fit indices to evaluate the fit of the two-level model while other fit indices previously developed for the single-level model were not as good as them. In this study, SRMR_W and SRMR_B found that the within-level fit the data well across all conditions while the between-level was affected by the small sample size. This phenomenon could not be detected by the chi-square p value.

4.2.2.3 RMSEA, CFI, and TLI

Looking at RMSEA, CFI, and TLI of the two-level model, two-level model fit the data well in most conditions, which was similar to the chi-square *p* value.

Looking at RMSEA, CFI, and TLI of the complex model, all RMSEAs were smaller than 0.06 regardless of the sample size. Almost all CFI and TLI were larger than 0.95 when the sample size was large. However, when sample size was 20(25) and ICC was 0.45, CFI and TLI was typically low, indicating the bad fit of the complex model under this condition.

Looking at RMSEA, CFI, and TLI of the normal model, when ICC was low, the normal model still fit the data, but when ICC was high, the fit indices indicated the bad fit of the model. The result from the p value and result from RMSEA, CFI, and TLI generally agreed with each other.

In general, it was found that when the complex model was used to estimate the two-level data, the model fit the data in all conditions looking at *p* value and RMSEA and the model fit the data better with the large sample size looking at CFI, and TLI. It was also found that when the normal model was used to estimate the two-level data, the model fit was not good especially when ICC was high (Table 8).

Table 8. Proportion of the model fit statistics meeting the cut-off criteria for the two-level model, complex model, and normal model in W2B1 model

	Sample Size	**ICC**	**Within-level and Between-level Factor Loading**	**Within-level and Between-level Factor Correlation**	**Model**	**Chi-square *p***	**RMSEA**	**CFI**	**TLI**	**SRMR_W**	**SRMR_B**
1	50(10)	0.45	0.5/0.8	0.3/0.3	1	0.95	1	1	0.78	1	0.97
					2	0.960	1	0.990	0.950		
					3	0.540	1	0.990	0.990		
2				0.3/0.6	1	0.98	1	1	0.89	1	0.97
					2	0.970	1	1	0.980		
					3	0.530	1	1	1		
3	50(10)	0.45	0.5/0.5	0.3/0.3	1	0.979	1	1	0.727	1	0.030
					2	0.969	1	0.797	0.727		
					3	0.080	0.949	0.797	0.191		
4				0.3/0.6	1	0.990	1	1	0.780	1	0.040
					2	0.960	1	0.840	0.770		
					3	0.090	0.920	0.840	0.250		
5	50(10)	0.25	0.5/0.5	0.3/0.3	1	0.927	1	1	0.739	1	0.260
					2	0.947	1	0.968	0.968		
					3	0.864	1	0.968	0.989		

Table 8 (continued)

	Sample Size	ICC	Within-level and Between-level Factor Loading	Within-level and Between-level Factor Correlation	Model	Chi-square *p*	RMSEA	CFI	TLI	SRMR_W	SRMR_B
6				0.3/0.6	1	0.914	1	1	0.777	1	0.276
					2	0.936	1	0.978	0.968		
					3	0.808	1	0.978	1		
7	50(10)	0.45	0.8/0.5	0.3/0.3	1	0.989	1	1	0.989	1	0
					2	0.949	1	0.959	0.888		
					3	0.050	0.878	0.959	0.333		
8				0.3/0.6	1	0.989	1	1	1	1	0.020
					2	0.949	1	0.969	0.939		
					3	0.050	0.818	0.969	0.394		
9	50(10)	0.25	0.8/0.5	0.3/0.3	1	0.928	1	1	0.979	1	0.051
					2	0.948	1	1	1		
					3	0.775	1	1	1		
10				0.3/0.6	1	0.969	1	1	0.979	1	0.061
					2	0.939	1	1	1		
					3	0.714	1	1	1		

Table 8 (continued)

	Sample Size	ICC	Within-level and Between-level Factor Loading	Within-level and Between-level Factor Correlation	Model	Chi-square *p*	RMSEA	CFI	TLI	SRMR_W	SRMR_B
11	20(25)	0.45	0.5/0.8	0.3/0.3	1	1	1	1	0.99	1	0.480
					2	0.970	1	0.880	0.780		
					3	0.110	0.890	0.880	0.840		
12				0.3/0.6	1	1	1	1	0.989	1	0.438
					2	0.969	1	0.918	0.846		
					3	0.102	0.897	0.918	0.857		
13	20(25)	0.45	0.5/0.5	0.3/0.3	1	1	1	1	1	1	0
					2	0.938	1	0.489	0.427		
					3	0	0.197	0.489	0		
14				0.3/0.6	1	1	1	1	1	1	0
					2	0.958	1	0.614	0.521		
					3	0	0.229	0.614	0		
15	20(25)	0.25	0.5/0.5	0.3/0.3	1	0.989	1	1	0.784	1	0.204
					2	0.903	1	0.903	0.860		
					3	0.580	1	0.903	0.946		

Table 8 (continued)

16				0.3/0.6	1	0.977	1	1	0.844	1	0.188
					2	0.911	1	0.944	0.889		
					3	0.622	1	0.944	0.956		
17	20(25)	0.45	0.8/0.5	0.3/0.3	1	1	1	1	1	1	0
					2	0.947	1	0.635	0.541		
					3	0	0.062	0.635	0		
18				0.3/0.6	1	1	1	1	1	1	0
					2	0.979	1	0.767	0.686		
					3	0	0.040	0.767	0.010		
19	20(25)	0.25	0.8/0.5	0.3/0.3	1	0.989	1	1	1	1	0.010
					2	0.917	1	1	0.968		
					3	0.343	0.989	1	0.958		
20				0.3/0.6	1	1	1	1	1	1	0.021
					2	0.947	1	1	1		
					3	0.305	1	1	0.989		

4.3 PARAMETER ESTIMATES

This section examines the effect of the factors on the relative bias of the mean of the within-level factor loading, between-level factor loading, and the residual variance in the two-level model, and effect of factors on the relative bias of the factor loading in the complex model, and normal model. A 5×4×2 between-subjects ANOVA was first performed on the relative bias of mean of within-level factor loading, between-level factor loading, residual variance as a function of factor, ICC, sample size, and correlation. The combination of factor and ICC (FactorICC) was a between-subject variable with five levels (0.5 0.8 0.45, 0.5 0.5 0.45, 0.5 0.5 0.25, 0.8 0.5 0.45, 0.8 0.5 0.25); sample size was a between-subject variable with four levels (50(10), 20(25), 250(10), 100(25)); between-level and within-level factor correlation was a between-subject variable with two levels (0.3/ 0.3 and 0.3/0.6).

Next, a 5×4×2×3 mixed ANOVA was performed on relative bias of mean of within-level factor loading as a function of FactorICC , sample size, correlation, and different models. The within-subjects independent variable was modeled with 3 levels (two-level model, complex model, and normal model). It intended to examine the effect of factors on the relative bias of within-level factor loading in three models. The analysis was first performed on W2B2 model followed by W2B1 model.

4.3.1 W2B2 Model

For outcome variables that have significant interaction or main effects, simple effect analysis was conducted. The effect is considered significant and warrants further investigation when the *p* value was smaller than .05 and $\eta_p^2 > .01$. Table 9 provides a summary of ANOVA results for the

relative bias of within-level mean factor loading (RB_meanW) and between-level mean factor loading (RB_meanB) of two-level model, residual variance (RB_res), and within-level factor loading in the complex (RB_cmeanW) and normal model (RB_nmeanW).

Table 9. Summary of η_p^2 for the RB_meanW, RB_meanB, RB_res , RB_cmeanW, and RB_nmeanW from Between-subjects ANOVA

Source	Two-Level Model RB_meanW	Two-Level Model RB_meanB	Two-Level Model RB_res	Complex Model RB_cmeanW	Normal Model RB_nmeanW
sample	.033*	.028*	.044*	.001	.001
corr	.000	.001	0	.001	.001
FactorICC	.003	.012*	.007	.959*	.959*
sample * corr	.000	.002	.001	.001	.001
sample * FactorICC	.003	.011*	.005	.002	.002
corr * FactorICC	.000	.000	.000	.000	.000
sample * corr * FactorICC	.000	.000	.001	.000	.000
factor*icc	0	0	0	.053*	.053*
factor	.002	.009	.004	/	/
icc	.002	.006	.005	/	/

4.3.1.1 Relative bias of mean of within-level factor loading

Only sample size had significant effect on the relative bias of the mean of the within-level factor loading (Table 9). From the result of ANOVA, relative bias of the small sample size was significantly larger than the large sample size, but there were no significant differences between two small sample sizes and two large sample sizes. The mean and standard error of the relative bias of the within-level factor loading among sample sizes was reported in Table 10. In general, the relative bias of the within-level factor loadings was trivial. As shown in Figure 2, the

relative bias of the within-level factor loading was within 0.02 across all conditions. The relative bias of the within-level factor loading decreased when the sample size increased from 500 to 2500.

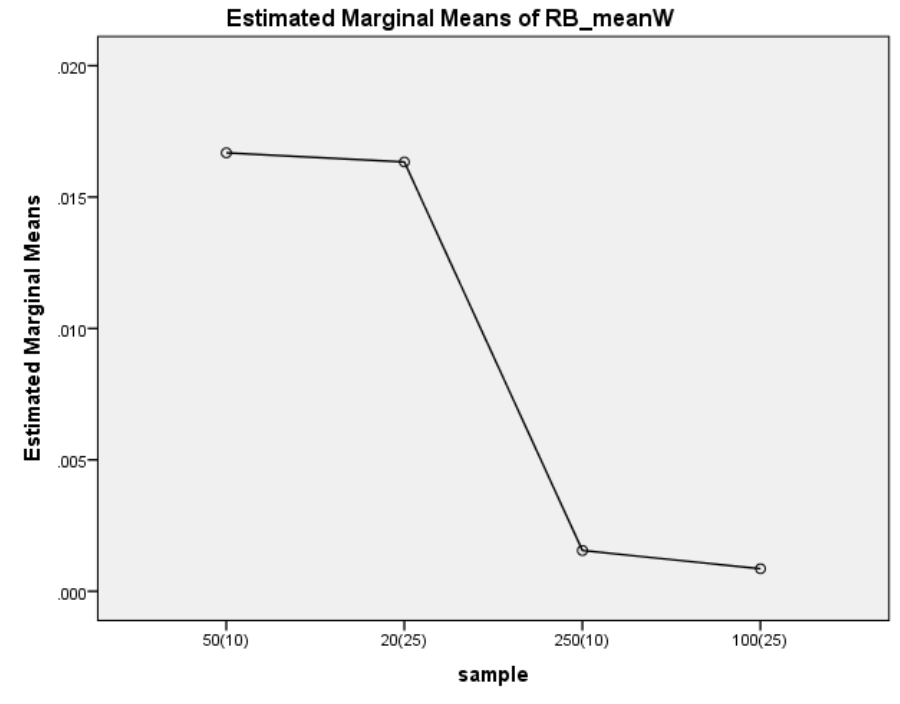

Figure 2. Relative bias of within-level factor loading as a function of sample size

Table 10. The relative bias of the within-level factor loading as a function of the sample sizes

Sample	*N*	*M*	*SE*
50(10)	864	.017	.001
20(25)	641	.016	.002
250(10)	1000	.002	.001
100(25)	997	.001	.001

4.3.1.2 Relative bias of mean of between-level factor loading

The interaction of sample and FactorICC, sample, and FactorICC all had significant effect on the relative bias of the mean of the between-level factor loading (Table 9), but none of the following marginal comparison and simple comparison was significant. Looking at the Figure 3, the

relative bias of the between-level factor loading was all within 0.05 except when the within-level sample size was 25 and the between-level sample size was 20. Negative relative bias exceeded -0.1 when the sample size was 20(25), the ICC was 0.45, and the factor loadings were 0.5 0.5 or 0.8 0.5. It indicated that the estimation of the true model could not be good when the ICC was high and the between-level sample size was small (Figure 3).

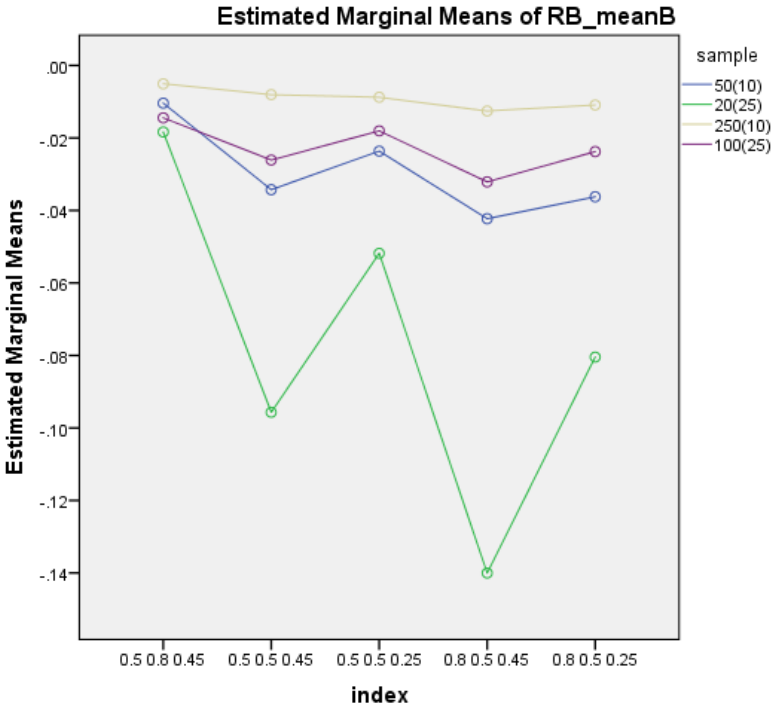

Figure 3. Relative bias of between-level factor loading as a function of FactorICC and sample size

Table 11. Mean and standard errors of relative bias of mean of between-level factor loading by sample size and FactorICC

sample	FactorICC	*M*	*SE*
50(10)	0.5 0.8 0.45	-.010	.010
	0.5 0.5 0.45	-.034	.010
	0.5 0.5 0.25	-.024	.011
	0.8 0.5 0.45	-.042	.011
	0.8 0.5 0.25	-.036	.010
20(25)	0.5 0.8 0.45	-.018	.012
	0.5 0.5 0.45	-.096	.012
	0.5 0.5 0.25	-.052	.014
	0.8 0.5 0.45	-.140	.011
	0.8 0.5 0.25	-.080	.012
250(10)	0.5 0.8 0.45	-.005	.010
	0.5 0.5 0.45	-.008	.010
	0.5 0.5 0.25	-.009	.010
	0.8 0.5 0.45	-.013	.010
	0.8 0.5 0.25	-.011	.010
100(25)	0.5 0.8 0.45	-.014	.010
	0.5 0.5 0.45	-.026	.010
	0.5 0.5 0.25	-.018	.010
	0.8 0.5 0.45	-.032	.010
	0.8 0.5 0.25	-.024	.010

4.3.1.3 Relative bias of the residual variance

Only the sample size had the significant effect on the relative bias of the residual variance (Table 9). The relative bias of the residual variances was negatively biased and it was within -0.1(Table 12). Looking at the plot, the relative bias looked larger when the sample size was small than when the sample size was large. The relative bias with 20(25) was significantly larger than relative bias with other sample sizes, $p<.001$, respectively. The relative bias with 50(10) was significantly larger than that with 250(10), $p<.001$. However, there was no significant difference in the relative bias between 50(10) and 100(25). The relative bias with 250(10) was significantly

smaller than that with 100(25), $p<.001$. The relative bias looked larger when the ICC was high than when the ICC was low, however, the ICC did not have significant effect on the relative bias of the residual variance in terms of the statistical test (Figure 4).

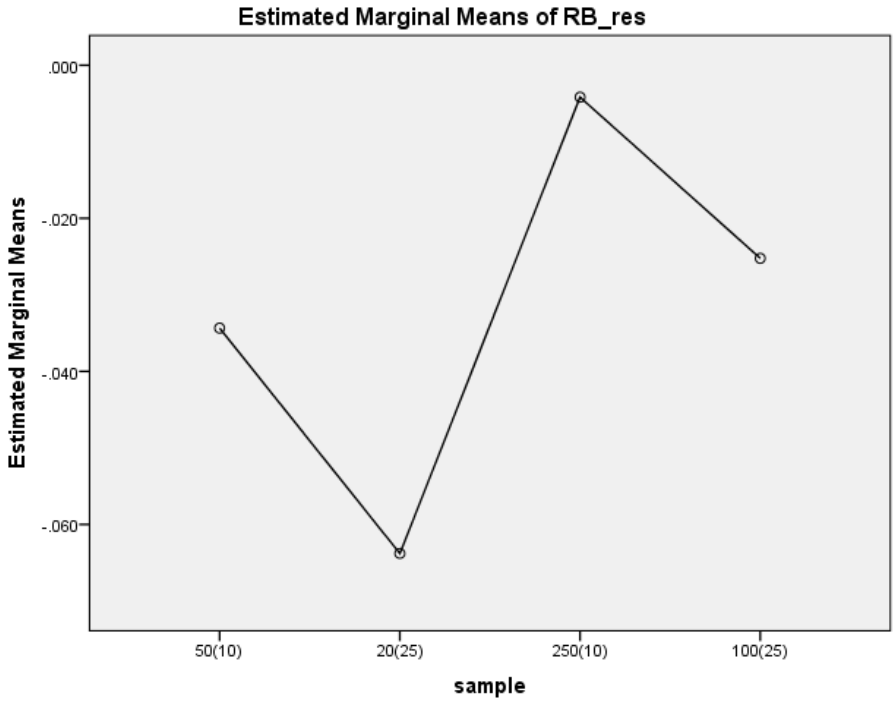

Figure 4. Relative bias of residual variance as a function of FactorICC and sample size

Table 12. Mean and standard errors of the relative bias of the residual variances by the sample sizes

sample	*M*	*SE*
50(10)	-.034	.003
20(25)	-.064	.004
250(10)	-.004	.003
100(25)	-.025	.003

4.3.1.4 Relative bias of mean of factor loading in the complex model

The interaction of factor loading and ICC had significant effect on the relative bias of the factor loading averaged across other factors (Table 9). When the factor loadings were estimated from

the single-level complex model, the relative bias of the factor loadings was within 0.1 when the factor loadings across levels were the same. As shown in Figure 5, when the factor loadings across levels were different, the relative bias of the factor loadings were larger than 0.2.

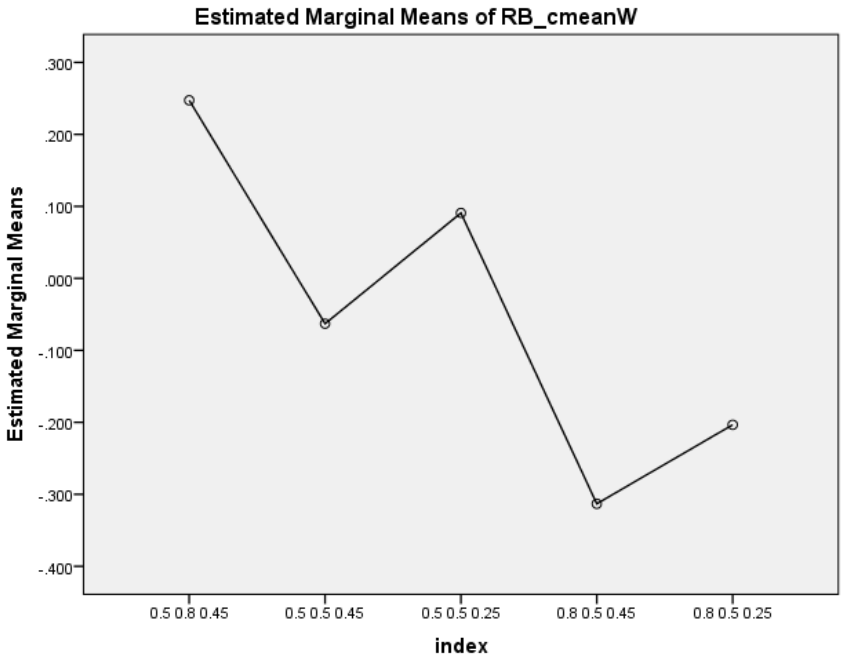

Figure 5. Relative bias of factor loading as a function of FactorICC in the complex model

Concerning the result of the ANOVA, the relative bias with 0.5 0.5 was significantly smaller than that with 0.8 0.5 for each ICC averaged across sample size and factor correlation (Figure 6). The mean and standard error of the relative bias of the factor loading by different ICC and factor loading was reported in Table 13.

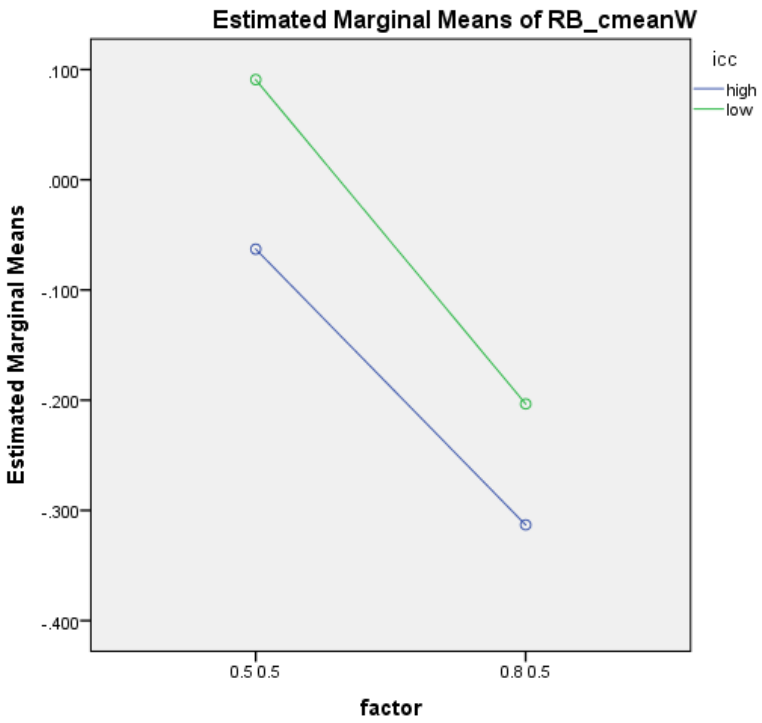

Figure 6. Relative bias of factor loading as a function of factor loading and ICC in the complex model

Table 13. Mean and standard error of the mean of the factor loading by FactorICC in the complex model

ICC	factor loading	*M*	*SE*	*N*
0.45	0.5 0.5	-.063	.002	720
	0.8 0.5	-.313	.002	702
0.25	0.5 0.5	.091	.002	645
	0.8 0.5	-.203	.002	719

4.3.1.5 Relative bias of mean of factor loading in the normal model

The interaction of factor loading and ICC had significant effect on the relative bias of the factor loading averaged across other factors (Table 9). When the factor loadings were estimated from the single-level normal model, the relative bias of the factor loadings was within 0.1 when the factor loadings across levels were the same. As shown in Figure 7, when the factor loadings across levels were different, the relative bias of the factor loadings were larger than 0.2.

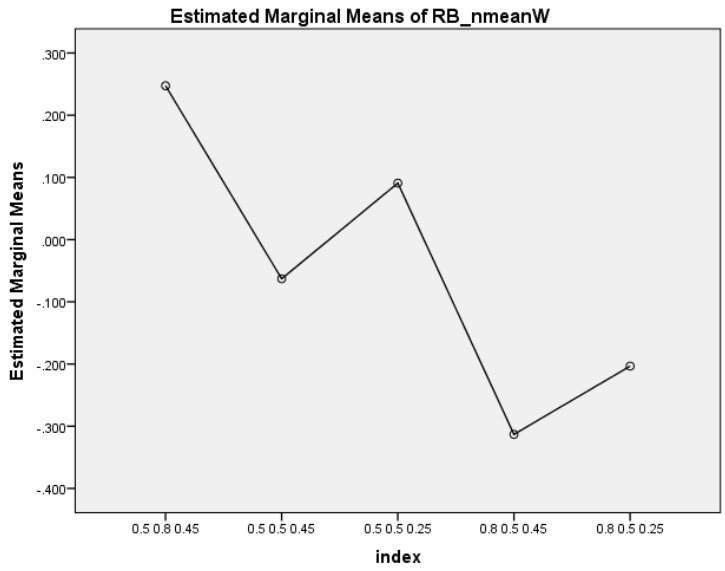

Figure 7. Relative bias of factor loading as a function of FactorICC and sample size in the normal model

Concerning the result of the ANOVA, the relative bias with 0.5 0.5 was significantly smaller than that with 0.8 0.5 for each ICC averaged across sample size and factor correlation (Figure 8). The mean and standard error of the relative bias of the factor loading by different ICC and factor loading was reported in Table 14.

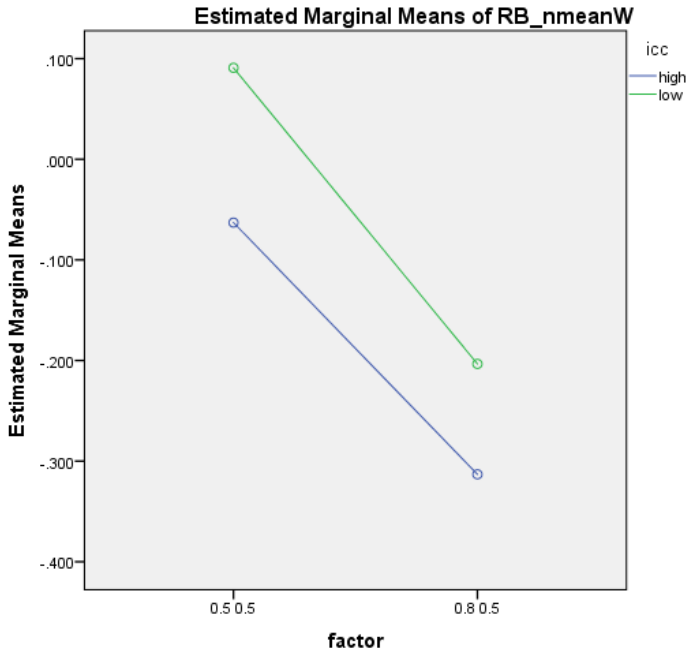

Figure 8. Relative bias of factor loading as a function of factor loading and ICC in the normal model

Table 14. Mean and standard error of the mean of the factor loading by FactorICC in the normal model

ICC	factor loading	*M*	*SE*	*N*
0.45	0.5 0.5	-.065	.002	720
	0.8 0.5	-.315	.002	702
0.25	0.5 0.5	.090	.002	645
	0.8 0.5	-.204	.002	719

4.3.1.6 Compare the relative bias of mean of within-level factor loading in three models

A 5×4×2×3 mixed analysis of variance was performed on relative bias of mean of within-level factor loading as a function of FactorICC, sample size, correlation, and different models. The within-subjects independent variable was model with 3 levels (two-level model, complex model, and normal model). The assumption of homogeneity of variance and homogeneity of covariance were not met, Box' *M*=11378.205, F (234, 2742364.109)=48.005, *p*<.001, Mauchly's *W*=.022.

The pattern of difference on relative bias of mean of within-level factor loading among models was significantly different among sample sizes, F(3.068,3540.172)=24.919, *p*<.001, η_p^2=.021. The pattern of difference on relative bias of mean of within-level factor loading among models was significantly different among FactorICCs, F(4.090, 3540.172)=15270.344, *p*<.001, η_p^2=.946. There was a significant difference on the relative bias of the mean of within-level factor loading among models averaged across FactorICCs, sample sizes and correlations, F(1.023, 3540.172)=5032.482, *p*<.001, η_p^2=.592. There was a significant difference on relative bias of the mean of within-level factor loading among FactorICCs averaged across models , sample sizes, and factor correlations, F(4,3462)=12804.584, *p*<.001, η_p^2=.937. However, there

was no significant difference on relative bias of the mean of within-level factor loading among sample sizes averaged across models, FactorICCs, and factor correlations, F(3, 3462)=9.146, p<.001, η_p^2=.008. Table 15 reported the partial effect size for the relative bias of the within-level factor loading for three models.

Table 15. Summary of η_p^2 for the relative bias of the within-level factor loading from mixed ANOVA

Source	RB_meanW
model*sample*corr*FactorICC	0
model*corr*FactorICC	0
model*sample*FactorICC	.004
model*sample*corr	0
model*FactorICC	.946*
model*corr	0
model*sample	.021*
model	.592*
sample	.008
corr	.001
FactorICC	.937*
factor*icc*model	.040*
factor*model	.471*

The pattern of difference among three models on the relative bias of the mean of the within-level factor loading among factor loadings and between low ICC and high ICC was significantly different averaged across sample size, F(2, 3461)=72.369, p<.001, η_p^2=.040. The

pattern of difference among models among factor loadings was significantly different averaged across ICC, factor correlation, and sample size, F(2,3461)=10973.807, p<.001, η_p^2 =. 864 (Figure 9) (Table 16).

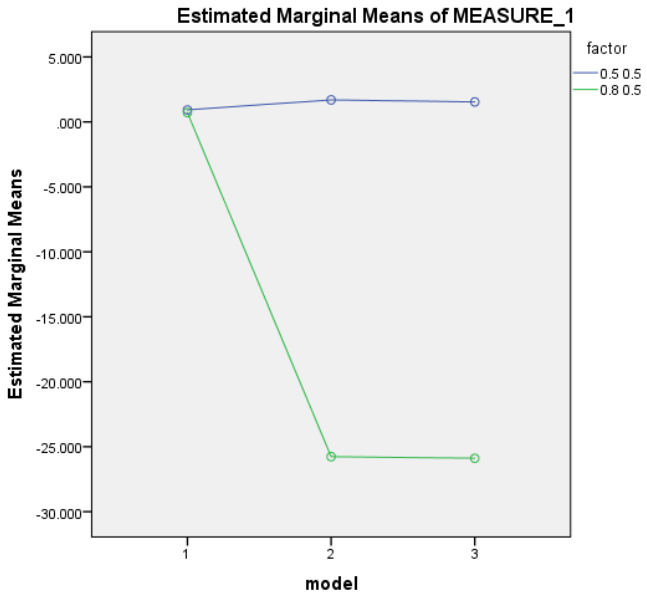

Figure 9. Relative bias of factor loading as a function of model and factor loading averaged across ICC, sample sizes, and factor correlation

Table 16. Mean and the standard error of the mean of the within-level factor loading by FactorICCs

FactorICC	*M*	*SE*	*N*
0.5 0.8 0.45	16.845	.132	716
0.5 0.5 0.45	-3.897	.130	720
0.5 0.5 0.25	6.312	.143	645
0.8 0.5 0.45	-20.641	.132	702
0.8 0.5 0.25	-13.404	.131	719

There was a significant difference among models averaged across ICC and sample size for 0.5 0.5 , F(2, 3461)=1796.356, p<.001, η_p^2 =.509. The relative bias of the two-level model was significantly smaller than that of the complex model and normal model, F(1, 3462) =

3548.265, p<.001, η_p^2=.506; F(1, 3462)=3591.231, p<.001, η_p^2=.509, respectively. However, the relative bias estimated from the complex model was not significantly smaller than that estimated by the normal model, F(1, 3462)=30.034, p<.001, η_p^2=.009. There was a significant difference among models averaged across ICC and sample size for 0.8 0.5 , F(2, 3461) = 1016.744, p<.001, η_p^2=. 370. The relative bias of the two-level model was significantly smaller than that of the complex model and normal model, F(1, 3462)=2011.914, p<.001, η_p^2=.368; F(1, 3462)=2033.393, p<.001, η_p^2=.370, respectively. However, the relative bias estimated from the complex model was not significantly smaller than that estimated by the normal model, F(1, 3462)=14.158, p<.001, η_p^2=.004 . The mean and standard error of the relative bias of the within-level factor loading among three models at different factor loading and ICC was reported in Table 17.

Table 17. Mean and standard error of the mean of the within-level factor loading by FactorICC and model

FactorICC	Model	*M*	*SE*
0.5 0.8 0.45	1	1.158	.156
	2	24.736	.155
	3	24.640	.157
0.5 0.5 0.45	1	1.080	.155
	2	-6.296	.154
	3	-6.476	.155
0.5 0.5 0.25	1	.824	.170
	2	9.084	.169
	3	9.029	.171
0.8 0.5 0.45	1	.844	.157
	2	-31.314	.156
	3	-31.452	.157
0.8 0.5 0.25	1	.522	.156
	2	-20.339	.155
	3	-20.394	.156

There was a significant difference in the relative bias among models for each sample size averaged across other factors, F(1, 863.246)=73.422, p<.001, η_p^2=.078; F(1.003, 641.646) =96.229, p<.001, η_p^2=.131;F(1,999.012)= 62.094, p<.001, η_p^2=.059; F(1,996.048)= 60.432, p<.001, η_p^2=.057. The relative bias of mean of within-level factor loading of two-level model was significantly smaller than that of complex model for each sample size, p<.001, respectively. The relative bias of mean of within-level factor loading of two-level model was significantly smaller than that of normal model for each sample size, p<.001, respectively. The relative bias of mean of within-level factor loading of complex model was significantly smaller than that of normal model for each sample size, p<.001, respectively (Figure 10) (Table 18).

Table 18. Mean and standard error of the mean of the within-level factor loading by sample size and model

Sample	Model	*M*	*SE*
50(10)	1	1.668	.141
	2	-4.659	.140
	3	-4.753	.141
20(25)	1	1.633	.165
	2	-4.779	.164
	3	-5.050	.165
250(10)	1	.155	.130
	2	-4.918	.129
	3	-4.932	.130
100(25)	1	.086	.130
	2	-4.947	.129
	3	-4.987	.130

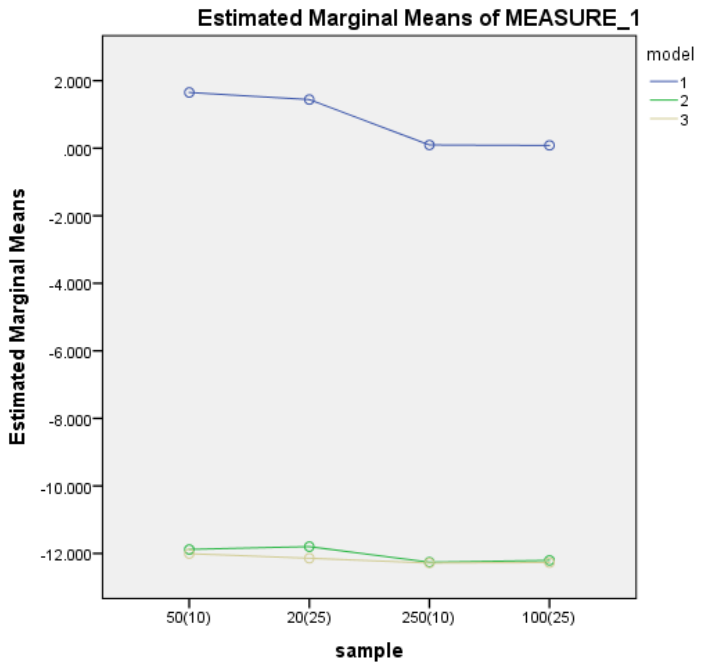

Figure 10. Relative bias of factor loading as a function of model and sample sizes averaged across ICC, factor loading, and factor correlation

From the result of analysis of variance, it was found that the relative bias of the within-level factor loading estimated by the two-level model was significantly smaller than that estimated by the complex model and the normal model for each factor loading averaged across ICC, sample sizes, and factor correlations. However, the relative bias of the within-level factor loading estimated by the complex model was not significantly smaller than that estimated by the normal model for each factor loading averaged across ICC, sample sizes, and factor correlations.

In conclusion, when the two-level model was used to estimate the parameters, the large relative bias of the within-level factor loading was affected by the small total sample size. There was no difference whether the within-level or the between-level sample size was large or small. As for the single-level complex model and normal model, sample size did not have any effect on the relative bias of the factor loading.

When the factor loadings across levels were the same, the factor loadings estimated from the single-level complex model and single-level normal model were all within 0.1, which were still acceptable. However, the factor loadings estimated from the single-level model when factor loadings across levels in the true model were different were much larger than those when factor loadings across levels in the true model were the same (Figure 9). This conclusion had been proved by the study of Porprasertmanit et al.(2014) and Wu and Kwok(2012).

Julian found that ignoring the clustering overestimated the factor loading using ML. Stochl et al. (2015) found that ignoring the clustering underestimated the factor loading using WLSMV. The conclusion from this study was different from Stochl et al. (2015) and Julian (2001). When the single-level model was used to estimate the clustered data, the factor loading was positively biased and above 0.1 when the between-level factor loading was higher than the within-level factor loading in the true model. The factor loading was negatively biased and above -0.1 when the within-level factor loading was higher than the between-level factor loading regardless of the ICC and sample size. Under the condition of the same factor loadings across levels, the higher ICC resulted in the higher relative bias in the single-level model, which was also proved by Pornprasertmanit et al. (2014) and Stochl et al. (2015).

4.3.1.7 Compare the standard error among three models

A 5×4×2×3 mixed analysis of variance was performed on the standard error of the mean of within-level factor loading as a function of FactorICC, sample size, correlation, and different models. Table 19 reports the partial effect size of the standard errors of the within-level factor loading in three models. The within-subjects independent variable was model with 3 levels (two-level model, complex model, and normal model). Table 19 reported the partial effect size for the standard errors of the within-level factor loading for three models.

Table 19. Summary of η_p^2 of the standard error of the within-level factor loading in three models

Source	η_p^2
model	.931*
model * sample	.750*
model * FactorICC	.548*
model * sample * FactorICC	.375*
sample	.969*
FactorICC	.819*
sample * FactorICC	.565*
model*factor	.292*
corr	0
corr * FactorICC	.002
sample * corr * FactorICC	.001
Model*corr	.001
Model*sample*corr	0
Model*sample*FactorICC*corr	0

The pattern of difference among samples, models, and FactorICCs was significantly different averaged across factor correlation, F(14.590, 4209.122)=173.152, p<.001, η_p^2=.375. The pattern of the difference among different models among different samples was significantly different averaged across other factors, F(3.647, 4209.122)=3470.479, p<.001, η_p^2=.750(Table 20).

Table 20. Mean and standard error of the mean of standard error of the factor loading by sample and model

sample	model	*M*	*SE*
50(10)	1	9.838	.030
	2	6.706	.015
	3	4.968	.008
20(25)	1	11.807	.036
	2	8.284	.017
	3	4.910	.009
250(10)	1	4.573	.028
	2	3.161	.013
	3	2.264	.007
100(25)	1	4.610	.028
	2	4.148	.013
	3	2.261	.007

The pattern of the difference among different FactorICCs among models was significantly different averaged across other factors, F(4.863, 4209.122)=1049.930, p<.001, η_p^2=.548. There was a significant difference in the SE among different models, F(1.216, 4209.122)=46789.749, p<.001, η_p^2=.931. There was a significant difference among different samples averaged across other factors, F(3, 3462)=36521.363, p<.001, η_p^2=.969. There was a significant difference among different FactorICCs averaged across other factors, F(4,3462)=3911.927, p<.001, η_p^2=.819. The pattern of the difference among different samples and FactorICCs was significantly different averaged across other factors, F(12,3462)=374.585, p<.001, η_p^2=.565 .

Since the pattern of difference among FactorICC and models averaged across sample size and factor correlation was significantly different, F(2,3461)=80.496, p<.001, η_p^2=.044, the two-

way interaction of the model and factor loading was performed. The pattern of difference among models and factor loadings was significantly different averaged across ICC, factor correlation, and sample size, F(2,3461)=713.013, p<.001, η_p^2=.292. There was a significant difference among models averaged across ICC and sample size for 0.5 0.5 , F(2,3461)= 7457.889, p<.001, η_p^2=.812. The standard error of the two-level model was significantly larger than that of the complex model and the normal model, F(1,3462)= 963,884, p<.001, η_p^2=.146; F(1,3462) =715.340, p<.001, η_p^2=.104, respectively. The standard error of the complex model was significantly larger than that of the normal model, F(1,3462)=14612.412, p<.001, η_p^2=.808. There was a significant difference among models averaged across ICC and sample size for 0.8 0.5 , F(2,3461)= 5968.333, p<.001, η_p^2=.775. The standard error of the two-level model was significantly larger than that of the complex model and the normal model, F(1,3462)= 351.228, p<.001, η_p^2=.092; F(1,3462)=441.010, p<.001, η_p^2=.113, respectively. The standard error of the complex model was significantly larger than that of the normal model, F(1,3462)=11780.439, p<.001, η_p^2=.773 (Figure 11)(Table 21).

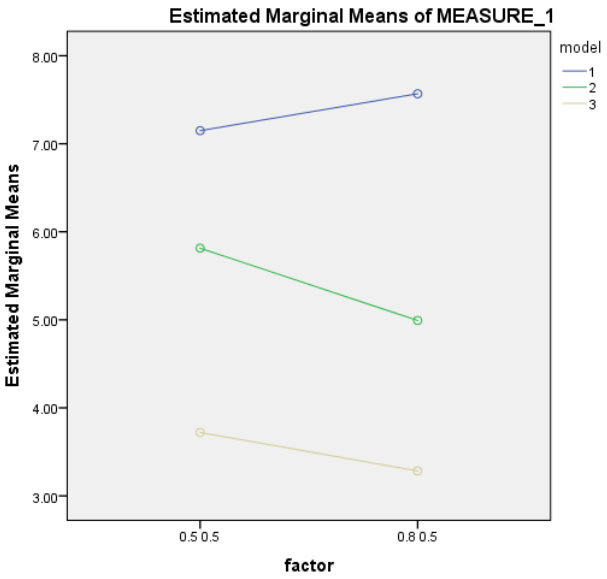

Figure 11. Standard Error as a function of factor loading and model averaged across sample size, ICC, and factor correlation

Table 21. Mean and standard error of the mean of standard error of the factor loading by model and FactorICC

FactorICC	model	*M*	*SE*
0.5 0.8 0.45	1	8.198	.034
	2	5.296	.016
	3	3.135	.009
0.5 0.5 0.45	1	8.295	.033
	2	7.604	.016
	3	4.335	.009
0.5 0.5 0.25	1	6.620	.037
	2	4.705	.018
	3	3.721	.010
0.8 0.5 0.45	1	8.590	.034
	2	6.467	.016
	3	3.753	.009
0.8 0.5 0.25	1	6.832	.034
	2	3.803	.016
	3	3.058	.009

Therefore, standard errors estimated from the two-level model were the largest, followed by the complex model and the normal model. For three models, the standard errors in the condition of factor loadings of 0.5 across levels and ICC of 0.25 were generally the smallest. In the study of Porprasertmanit et al. (2014), it was found that SE was affected by the between-level standardized factor loading. When the between-level standardized factor loading was low, SEs was higher; when the between-level standardized factor loading was high, SEs was lower. In this study, it was also found that when ICC was the same, the lower between-level factor loading resulted in the higher SE and the higher between-level factor loading resulted in the lower SE in the single-level model (Figure 12).

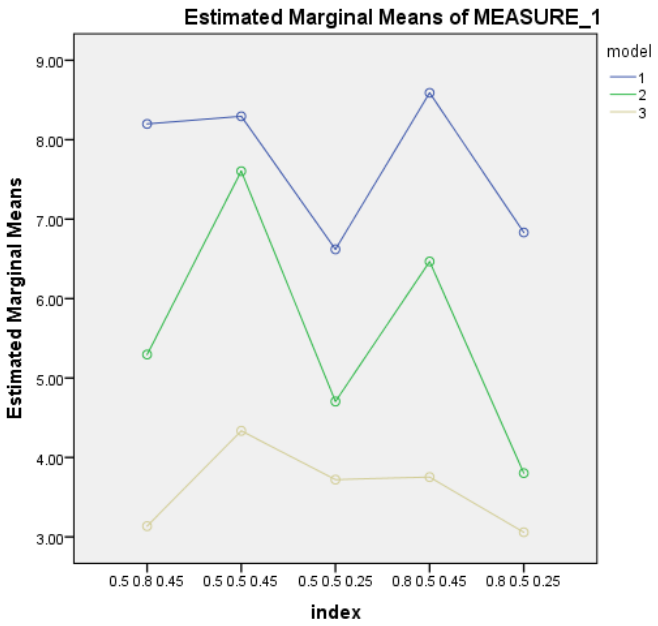

Figure 12. Standard error as a function of FactorICC and model averaged across sample size and factor correlation

Julian (2001) and Stapleon (2006) stated that the standard errors were underestimated if the clustering structure in the data was not accounted for. Our ANOVA results corroborated their

findings with the standard error estimated from the two-level model consistently larger than the complex model and normal model across different levels of FactorICCs. The standard error estimated from the complex model was consistently larger than the normal model across the FactorICCs.

4.3.2 W2B1 Model

For outcome variables that have significant interaction or main effects, simple effect analysis were conducted. The effect is considered significant and warrants further investigation when the *p* value was smaller than .05 and η_p^2 >.01. Table 22 provides a summary of ANOVA results for RB_meanW, RB_meanB, RB_res, RB_cmeanW, and RB_nmeanW.

Table 22. Summary of η_p^2 for the RB_meanW, RB_meanB, RB_res , RB_cmeanW, and RB_nmeanW from Between-subjects ANOVA in the W2B1 Model

Source	Two-Level Model	Two-Level Model	Two-Level Model	Complex Model	Normal Model
	RB_meanW	RB_meanB	RB_res	RB_cmeanW	RB_nmeanW
sample	.029*	.008	.073*	.001	.001
corr	.000	.000	.000	.000	.000
FactorICC	.002	.002	.002	.946*	.946*
sample * corr	.000	.000	.000	.000	.000
sample * FactorICC	.002	.002	.004	.002	.002
corr * FactorICC	.000	.000	.000	.000	.000
sample * corr * FactorICC	.000	.000	.000	.000	.000
Factor*icc	0	0	0	.038*	.039*
factor	.001	.001	.002	/	/
icc	.001	.001	.001	/	/

4.3.2.1 Relative bias of mean of within-level factor loading

Only sample size had significant effect on the relative bias of the mean of the within-level factor loading (Table 23). From the result of ANOVA, relative bias of the small sample size was significantly larger than that of the large sample size, but there were no significant differences between two small sample sizes and two large sample sizes (Table 23). In general, the relative bias of the within-level factor loadings was trivial. The relative bias of the within-level factor loading was within 0.02 across all conditions. The relative bias of the within-level factor loading decreased when the sample size increased from 500 to 2500 (Figure 13).

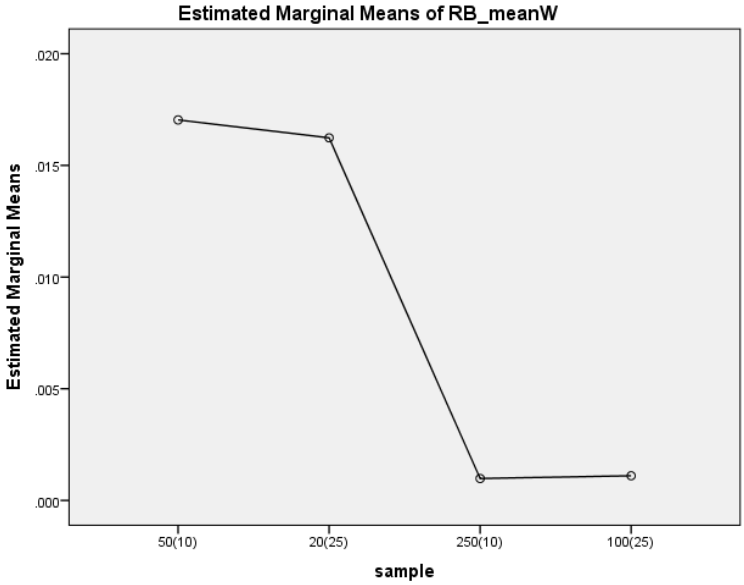

Figure 13. Relative bias of within-level factor loading as a function of sample size

Table 23. Mean and standard errors of relative bias of mean of within-level factor loading by sample size

Sample	*M*	*SE*	*N*
50(10)	.017	.001	983
20(25)	.016	.001	959
250(10)	.001	.001	1000
100(25)	.001	.001	1000

4.3.2.2 Relative bias of mean of between-level factor loading

ANOVA demonstrated that none of the factors had significant effect on the relative bias of the between-level factor loading.

4.3.2.3 Relative bias of the residual variance

Only the sample size had the significant effect on the relative bias of the residual variance (Table 24). The relative bias of the residual variances was negatively biased and it was within -0.1. Looking at the plot, the relative bias looked larger when the sample size was small than when the sample size was large. The relative bias with 20(25) was significantly larger than relative bias with other sample sizes, $p<.001$, respectively. The relative bias with 50(10) was significantly larger than relative bias of 250(10), $p<.001$. However, there was no significant difference in the relative bias of the residual between 50(10) and 100(25). The relative bias with 250(10) was significantly smaller than that with 100(25), $p<.001$. The relative bias looked larger when the ICC was high than when the ICC was low, however, the ICC did not have significant effect on the relative bias of the residual variance in terms of the statistical test (Figure 14) (Table 24).

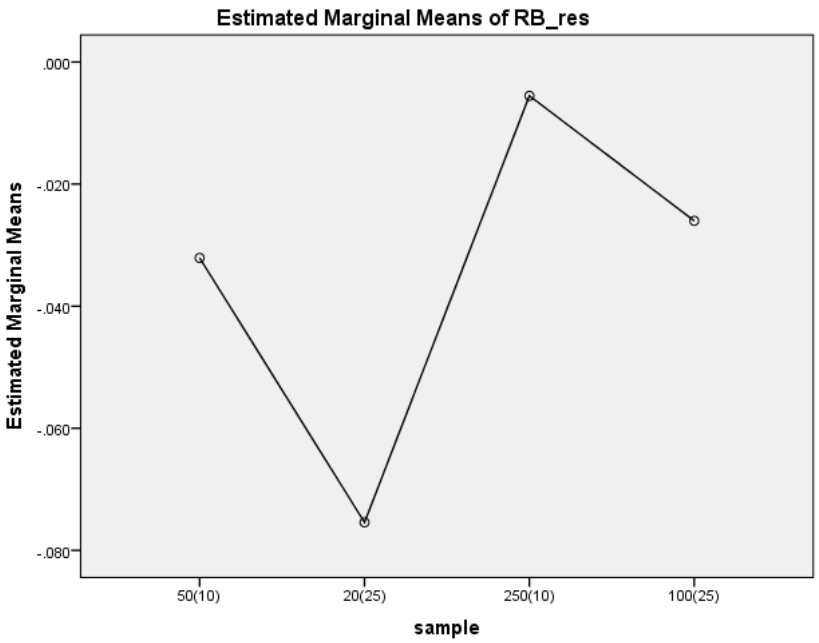

Figure 14. Relative bias of residual variance as a function of sample size

Table 24. Mean and the standard errors of the relative bias of residual variances by the samples

Sample	*M*	*SE*	*N*
50(10)	-.032	.003	983
20(25)	-.075	.003	959
250(10)	-.006	.003	1000
100(25)	-.026	.003	1000

4.3.2.4 Relative bias of mean of factor loading in the complex model

The interaction of factor loading and ICC had significant effect on the relative bias of the factor loading averaged across other factors (Table 23). When the factor loadings were estimated from the single-level complex model, the relative bias of the factor loadings was within 0.1 when the factor loadings across levels were the same. As shown in Figure 15, when the factor loadings across levels were different, the relative bias of the factor loadings were larger than 0.2.

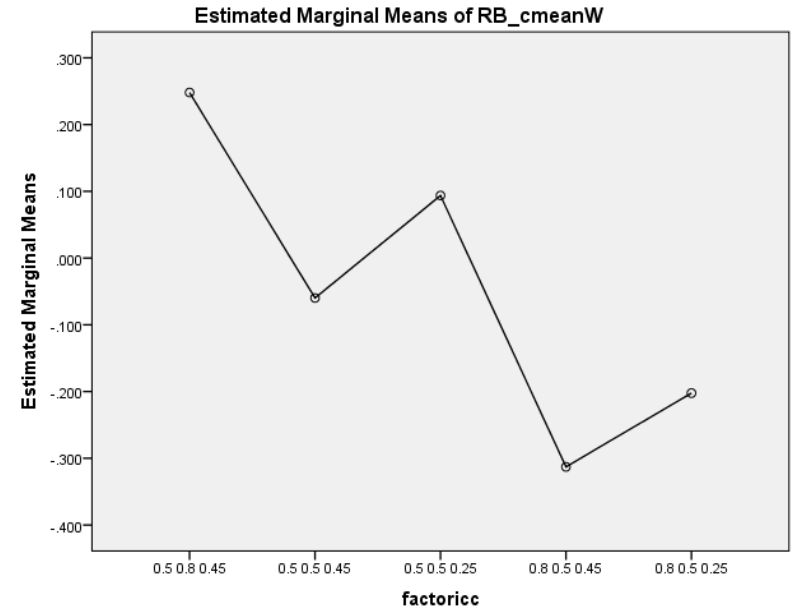

Figure 15. Relative bias of factor loading as a function of FactorICC in the complex model

Concerning the result of the ANOVA, the relative bias with 0.5 0.5 was significantly smaller than that with 0.8 0.5 for each ICC averaged across sample size and factor correlation (Figure 16) (Table 25).

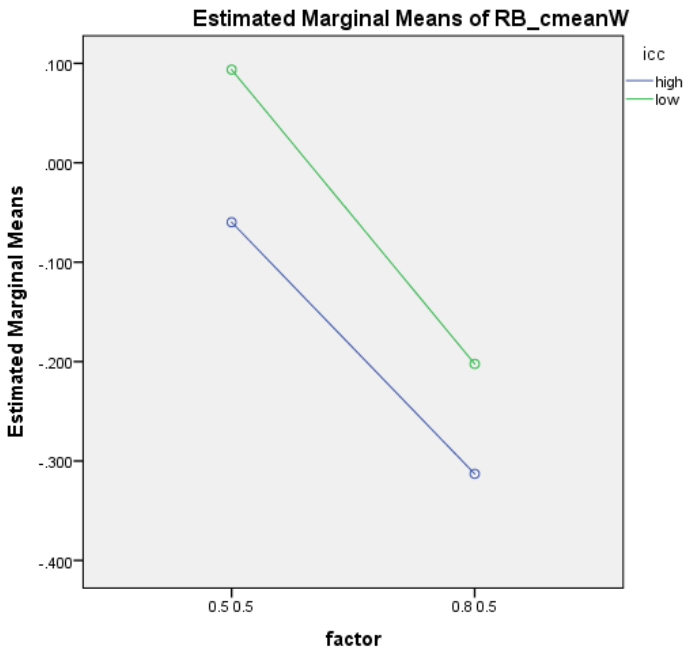

Figure 16. Relative bias of factor loading as a function of factor loading and ICC in the complex model

Table 25. Mean and standard error of the mean of the factor loading by factor loading and ICC in the complex model

ICC	Factor Loading	*M*	*SE*	*N*
0.45	0.5 0.5	-.060	.002	791
	0.8 0.5	-.313	.002	793
0.25	0.5 0.5	.094	.002	773
	0.8 0.5	-.202	.002	787

4.3.2.5 Relative bias of mean of factor loading in the normal model

The interaction of factor loading and ICC had significant effect on the relative bias of the factor loading averaged across other factors (Table 23). When the factor loadings were estimated from the single-level normal model, the relative bias of the factor loadings was within 0.1 when the

factor loadings across levels were the same. As indicated in Figure 17, when the factor loadings across levels were different, the relative bias of the factor loadings were larger than 0.2.

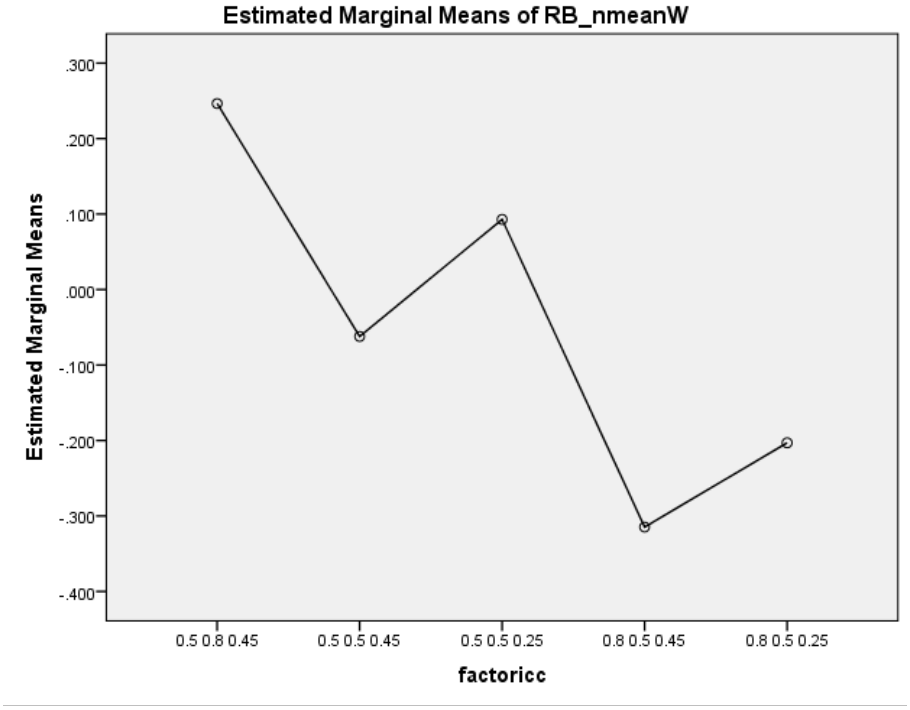

Figure 17. Relative bias of factor loading as a function of FactorICC in the complex model

Concerning the result of the ANOVA, the relative bias with 0.5 0.5 was significantly smaller than that with 0.8 0.5 for each ICC averaged across sample size and factor correlation (Figure 18) (Table 26).

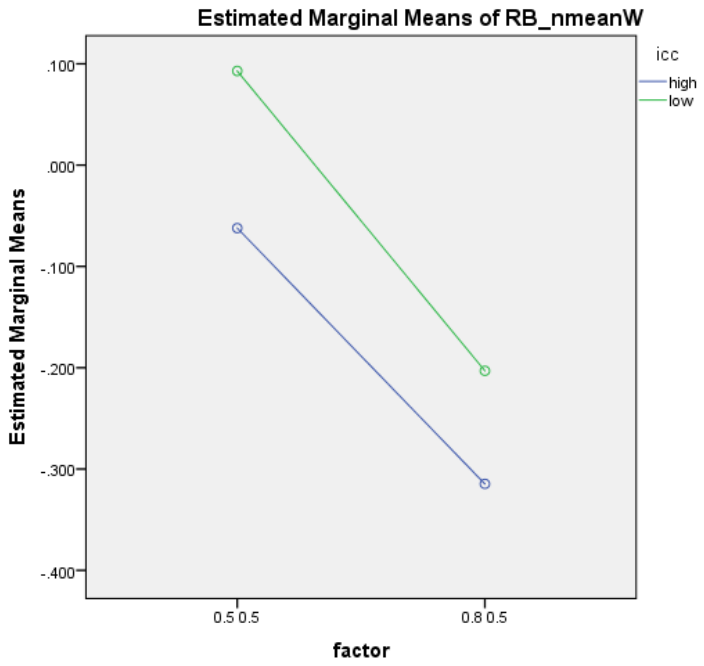

Figure 18. Relative bias of factor loading as a function of factor loading and ICC in the normal model

Table 26. Mean and standard error of the mean of the factor loading by factor and ICC in the normal model

ICC	factor loading	*M*	*SE*	*N*
0.45	0.5 0.5	-.062	.002	791
	0.8 0.5	-.315	.002	793
0.25	0.5 0.5	.093	.002	773
	0.8 0.5	-.203	.002	787

4.3.2.6 Compare the relative bias of mean of within-level factor loading in three models

A 5×4×2×3 mixed analysis of variance was performed on relative bias of mean of within-level factor loading as a function of FactorICC, sample size, correlation, and different models. The within-subjects independent variable was model with 3 levels (two-level model, complex model, and normal model). The assumption of homogeneity of variance and homogeneity of covariance were not met, Box' M=11311.980, F(234, 4698031.627)=57.789, p<.001, Mauchly's W=.016.

The interaction of model, factor, and ICC, the interaction of factor loading and model, the interaction of sample and model, and model all had significant effect on the relative bias of the within-level factor loading. Table 27 reported the partial effect size of the relative bias of the within-level factor loading in three models.

Table 27. Summary of η_p^2 for the relative bias of the within-level factor loading from mixed ANOVA

Source	RB_meanW
Model*sample*corr*FactorICC	0
Model*corr*FactorICC	0
Model*sample*FactorICC	.004
Model*sample*corr	0
Model*FactorICC	.929*
Model*corr	0
Sample*FactorICC	.001
Model*sample	.015*
model	.508*
sample	.007
corr	0
FactorICC	.922*
Factor*icc*model	.028*
Factor*model	.462*

The pattern of difference on relative bias of mean of within-level factor loading among models was significantly different among sample sizes averaged across factor correlation and factor loading, F(3.055,3973.264)=20.371, p<.001, η_p^2=.015. The pattern of difference on relative bias of mean of within-level factor loading among models was significantly different among FactorICCs averaged across sample sizes and factor correlation, F(4.033,3973.264)= 12729.334, p<.001, η_p^2=.929 . There was a significant difference on the relative bias of the mean

of within-level factor loading among models averaged across FactorICCs, sample sizes and correlations, F(1.018, 3540.172)= 4023.571 , p<.001, η_p^2=.508. There was a significant difference on relative bias of the mean of within-level factor loading among FactorICCs averaged across models , sample sizes, and factor correlations, F(4,3902)= 11517.462 , p<.001, η_p^2=.922. However, there was no significant difference on relative bias of the mean of within-level factor loading among sample sizes averaged across models, FactorICCs, and factor correlations, F(3, 3902)= 9.192 , p<.001, η_p^2=.007.

The pattern of difference among three models on the relative bias of the mean of the within-level factor loading among factor loadings and between ICCs was significantly different averaged across sample size, F(2, 3901)=72.369, p<.001, η_p^2=.028. The pattern of difference among models among factor loadings was significantly different averaged across ICC, factor correlation, and sample size, F(2, 3901)=9465.315, p<.001, η_p^2=.829 .

There was a significant difference among models averaged across ICC and sample size for 0.5 0.5 , F(2, 3901)=1567.827, p<.001, η_p^2=.446. The relative bias of the two-level model was significantly smaller than that of the complex model and normal model, F(1, 3902) = 2957.418, p<.001, η_p^2=.431; F(1, 3902)=3049.333, p<.001, η_p^2=.439, respectively. Also, the relative bias estimated from the complex model was significantly smaller than that estimated by the normal model, F(1, 3902)=56.90, p<.001, η_p^2=.014. There was a significant difference among models averaged across ICC and sample size for 0.8 0.5 , F(2, 3901)= 856.915, p<.001, η_p^2=. 305. The relative bias of the two-level model was significantly smaller than that of the complex model and normal model, F(1, 3902)=1622.242, p<.001, η_p^2=.294; F(1, 3902)

=1671.029, p<.001, η_p^2=.300, respectively. However, the relative bias estimated from the complex model was not significantly smaller than that estimated by the normal model, F(1, 3902)=27.565, p<.001, η_p^2=.007(Figure 19). The mean and standard error of the relative bias of the within-level factor loading among three models at different factor loading and ICC was reported in Table 28.

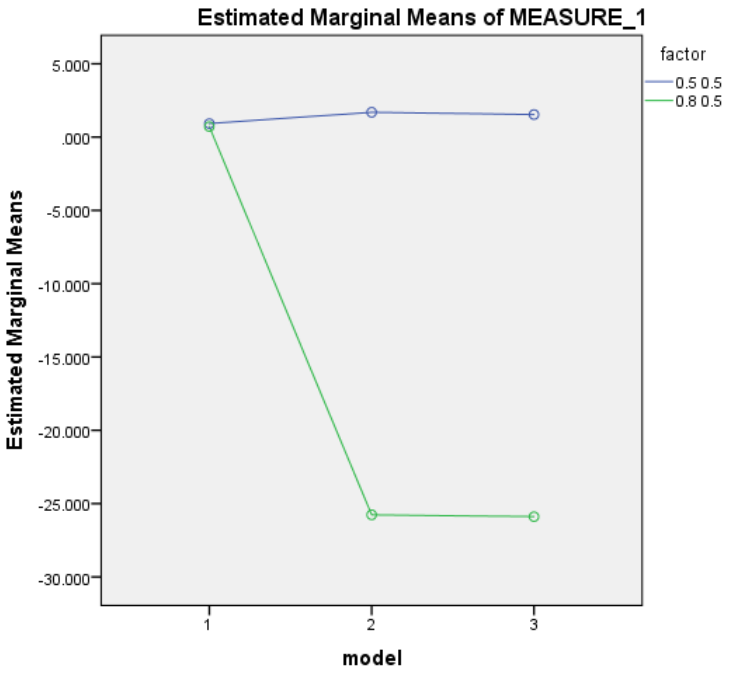

Figure 19. Relative bias of the factor loading as a function of model and factor loading averaged across sample sizes, ICCs, and factor correlations

Table 28. Mean and the standard error of the mean of the factor loading by model and FactorICCs

FactorICC	model	*M*	*SE*
0.5 0.8 0.45	1	1.145	.160
	2	24.801	.171
	3	24.648	.171
0.5 0.5 0.45	1	.914	.161
	2	-5.982	.172
	3	-6.217	.172
0.5 0.5 0.25	1	.932	.163
	2	9.369	.174
	3	9.291	.174
0.8 0.5 0.45	1	.836	.161
	2	-31.291	.172
	3	-31.468	.172
0.8 0.5 0.25	1	.593	.161
	2	-20.238	.172
	3	-20.306	.172

There was a significant difference in the relative bias among models for each sample size averaged across other factors, p<.001 , respectively; η_p^2 =.082, η_p^2 =.083, η_p^2 =.055, η_p^2 =.055, respectively. The relative bias of mean of within-level factor loading of two-level model was significantly smaller than that of complex model for each sample size, p<.001, respectively. The relative bias of mean of within-level factor loading of two-level model was significantly smaller than that of normal model for each sample size, *p*<.001, respectively. The relative bias of mean of within-level factor loading of complex model was significantly smaller than that of normal model for each sample size, *p*<.001, respectively (Figure 20) (Table 29).

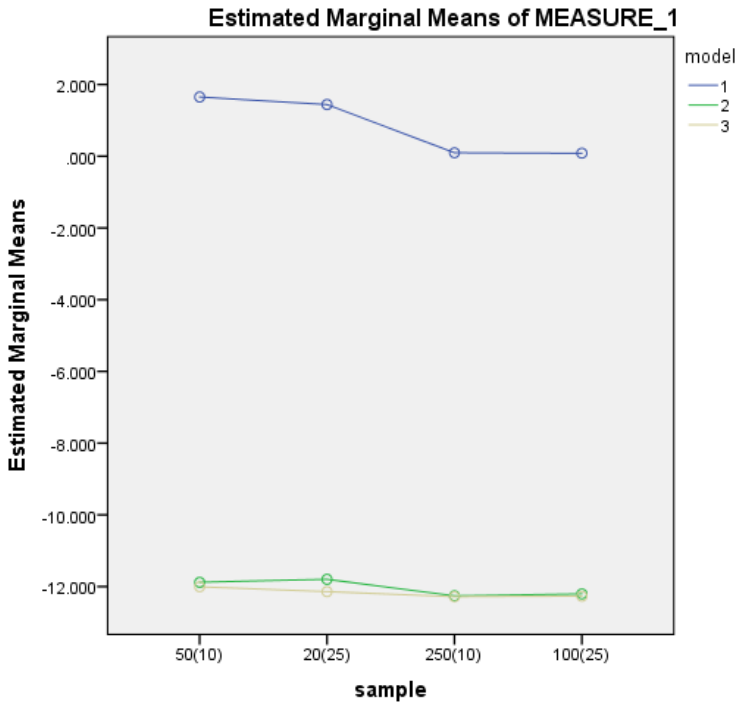

Figure 20. Relative bias of the factor loading as a function of model and factor loading averaged across sample sizes, ICC, and factor correlations

Table 29. Mean and standard error of the mean of the factor loading by model and sample size

sample	model	*M*	*SE*
50(10)	1	1.704	.144
	2	-4.469	.154
	3	-4.602	.154
20(25)	1	1.623	.146
	2	-4.580	.156
	3	-4.923	.156
250(10)	1	.098	.143
	2	-4.826	.153
	3	-4.850	.153
100(25)	1	.110	.143
	2	-4.798	.153
	3	-4.866	.153

In conclusion, similar to the ANOVA result of W2B2 model, it was found that the relative bias of the within-level factor loading estimated by the two-level model was significantly

smaller than the complex model and the normal model for each factor loading averaged across ICC , sample sizes, and factor correlations.

Similar to the result of W2B2 model, when the two-level model was used to estimate the parameters, the large relative bias of the within-level factor loading was affected by the small total sample size. There was no difference whether the within-level or between-level sample size was large or small. As for the single-level complex model and normal model, sample size did not have any effect on the relative bias of the factor loading. When the factor loadings across levels were the same, the factor loadings estimated from the single-level complex model and single-level normal model were all within 0.1, which were still acceptable. However, the factor loadings estimated from the single-level model when factor loadings across levels in the true model were different were much larger than those when factor loadings across levels in the true model were the same.

Similar to the result of W2B2 model, when the single-level model was used to estimate the clustered data, the factor loading was positively biased and above 0.1 when the between-level factor loading was higher than the within-level factor loading in the true model. The factor loading was negatively biased and above -0.1 when the within-level factor loading was higher than the between-level factor loading regardless of the ICC and sample size. Under the condition of the same factor loadings across levels, the higher ICC resulted in the higher relative bias in the single-level model.

Different from the result of W2B2 model, the between-level factor loading was not affected by any of the factors. Different from the result of W2B2 model, the relative bias of the within-level factor loading estimated by the complex model was significantly smaller than that

estimated by the normal model for 0.5 0.5 averaged across ICC, sample sizes, and factor correlations (Figure 19).

4.3.2.7 Compare the standard error among three models in W2B1 Model

A 5×4×2×3 mixed analysis of variance was performed on the standard error of the mean of within-level factor loading as a function of FactorICC, sample size, correlation, and different models. The within-subjects independent variable was model with 3 levels (two-level model, complex model, and normal model). The pattern of difference among samples, models, and FactorICCs was significantly different averaged across other factors, F(15.150, 4926.123)= 206.019, p<.001, η_p^2 =.388. All three-way interactions, two-way interactions, and main effects had significant effect on the standard error of the factor loading. Table 30 reported the partial effect size of standard errors in three models.

Table 30. Summary of η_p^2 for the standard error of the within-level factor loading from mixed ANOVA

Source	η_p^2
model	.934*
Model*sample	.756*
Model*corr	.040*
Model*FactorICC	.522*
Model*sample*corr	.029*
Model*sample*FactorICC	.388*
Model*corr*FactorICC	.013*
Model*sample*corr*facoricc	.009
Model*factor	.268*
Sample	.971*
Corr	.208*
FactorICC	.804*
Sample*corr	.069*
Sample*FactorICC	.580*
Corr*FactorICC	.022*
Sample*corr*FactorICC	.012*

Since the pattern of difference among FactorICC and models averaged across sample size and factor correlation was significant, F(2,3901)=69.498, p<.001, η_p^2=.034, the two-way interaction of factor loading and model was performed. The pattern of difference among models

and factor loadings was significantly different averaged across ICC, factor correlation, and sample size, F(2,3901)=714.885, p<.001, η_p^2=.268 (Figure 21). There was a significant difference among models averaged across ICC and sample size for 0.5 0.5, F(2,3901)= 6499.264, p<.001, η_p^2=.769. The standard error of the two-level model was significantly larger than that of the complex model and the normal model, F(1,3902)= 649.379, p<.001, η_p^2=.143; F(1,3902)=429.580, p<.001, η_p^2=.099, respectively. The standard error of the complex model was significantly larger than that of the normal model, F(1, 3902)=12511.227, p<.001, η_p^2=.762.

There was a significant difference among models averaged across ICC and sample size for 0.8 0.5 , F(2,3901)= 4864.763, p<.001, η_p^2=.714. The standard error of the two-level model was significantly larger than that of the complex model and the normal model, F(1,3902)= 310.939, p<.001, η_p^2=.074; F(1,3902)=499.180, p<.001, η_p^2=.113, respectively. The standard error of the complex model was significantly larger than that of the normal model, F(1,3902)=9515.193, p<.001, η_p^2=.709(Figure 21)(Table 31).

Table 31. Mean and standard error of the standard error of the mean of the factor loading by FactorICC and model

FactorICC	*model*	*M*	*SE*
0.5 0.8 0.45	1	7.734	.030
	2	5.010	.015
	3	2.867	.009
0.5 0.5 0.45	1	7.900	.030
	2	7.127	.015
	3	3.980	.009
0.5 0.5 0.25	1	6.397	.031
	2	4.500	.016
	3	3.460	.009
0.8 0.5 0.45	1	8.408	.030
	2	6.219	.015
	3	3.597	.009
0.8 0.5 0.25	1	6.728	.030
	2	3.765	.015
	3	2.970	.009

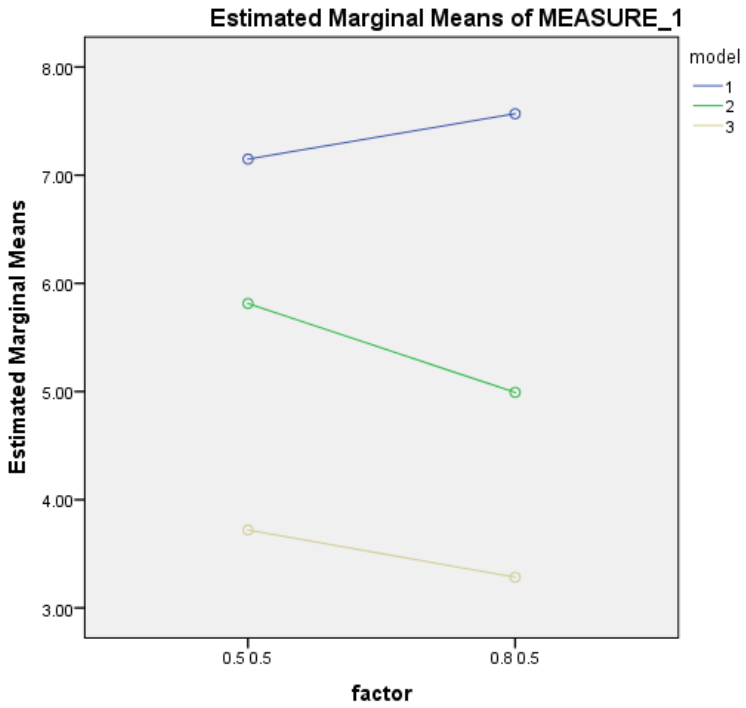

Figure 21. Relative bias of the standard error of factor loading as a function of model and factor loading

Therefore, similar to W2B2 model, standard errors estimated from the two-level model were the largest, followed by the complex model and the normal model. For three models, the standard errors in the condition of factor loadings of 0.5 across levels and ICC of 0.25 were generally the smallest. Also, similar to W2B2 model, when ICC was the same, the lower between-level factor loading resulted in the higher SE and the higher between-level factor loading resulted in the lower SE in the single-level model (Figure 22).

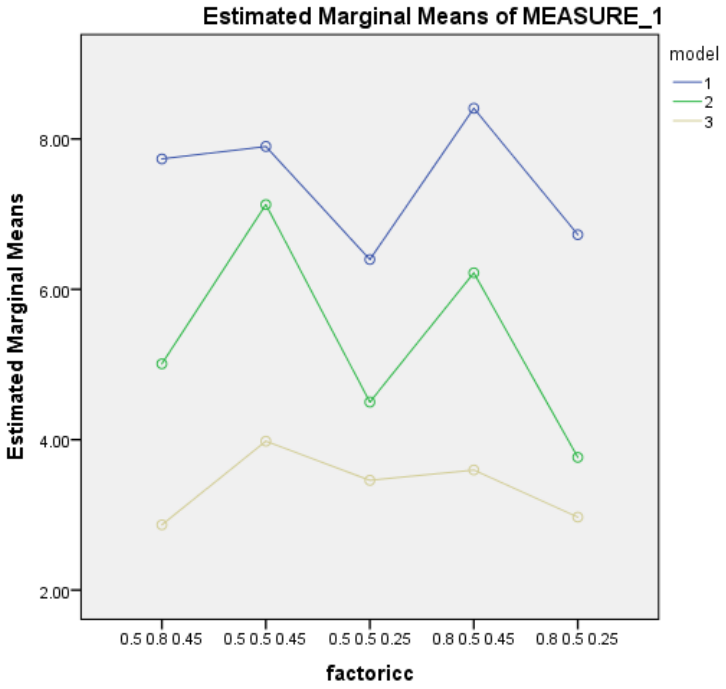

Figure 22. Relative bias of the standard error of factor loading as a function of model and FactorICC

There were statistically significant differences in the standard errors between two-level CFA model and complex single-level CFA model, and between complex single-level CFA model and normal single-level CFA model. To look at the practical differences of the standard errors among these three models, the empirical 95% confidence interval of the standard errors were examined for each model by calculating the 2.5^{th} and 97.5^{th} percentile of the empirical distribution of standard errors computed from simulated datasets in each condition. It was found

that there was no overlap in the confidence intervals among three models in the W2B2 model (Table 32) and W2B1 model (Table 33). Standard errors had obvious biases when the single-level normal CFA model was used to estimate the two-level data for both W2B2 model and W2B1 model.

Table 32. 95% confidence interval for the standard errors of the within-level factor loading of the two-level model, the complex model and the normal model in the W2B2 model (In this table, "m" represents the multilevel model, "c" represents the complex model, and "n" represents the normal model).

Obs	cond	true_m	2.5_m	97.5_m	true_c	2.5_c	97.5_c	true_n	2.5_n	97.5_n
1	111	0.112	0.099	0.101	0.064	0.061	0.063	0.064	0.043	0.044
2	112	0.112	0.100	0.102	0.063	0.062	0.064	0.063	0.042	0.044
3	121	0.117	0.127	0.141	0.087	0.079	0.083	0.088	0.043	0.045
4	122	0.118	0.126	0.139	0.087	0.079	0.083	0.088	0.041	0.043
5	131	0.046	0.046	0.046	0.028	0.029	0.029	0.029	0.020	0.020
6	132	0.046	0.046	0.046	0.028	0.029	0.029	0.028	0.019	0.020
7	141	0.047	0.047	0.048	0.040	0.039	0.039	0.040	0.020	0.020
8	142	0.046	0.047	0.048	0.040	0.040	0.040	0.041	0.019	0.020
9	211	0.112	0.099	0.102	0.098	0.088	0.091	0.098	0.060	0.062
10	212	0.106	0.099	0.102	0.095	0.087	0.090	0.096	0.058	0.060
11	221	0.125	0.134	0.145	0.146	0.112	0.117	0.149	0.058	0.061
12	222	0.123	0.132	0.142	0.112	0.111	0.115	0.138	0.056	0.058
13	231	0.048	0.046	0.047	0.135	0.043	0.043	0.043	0.028	0.028
14	232	0.048	0.046	0.046	0.043	0.042	0.043	0.042	0.027	0.027
15	241	0.046	0.048	0.049	0.042	0.058	0.059	0.059	0.028	0.028
16	242	0.047	0.048	0.049	0.059	0.058	0.059	0.059	0.027	0.027
17	311	0.102	0.091	0.093	0.059	0.058	0.060	0.059	0.051	0.052
18	312	0.104	0.092	0.094	0.059	0.058	0.060	0.060	0.050	0.052
19	321	0.101	0.085	0.089	0.060	0.067	0.070	0.068	0.051	0.053
20	322	0.104	0.086	0.089	0.068	0.067	0.070	0.071	0.050	0.052

Table 32 (continued)

Obs	cond	true_m	2.5_m	97.5_m	true_c	2.5_c	97.5_c	true_n	2.5_n	97.5_n
21	331	0.044	0.044	0.044	0.070	0.027	0.028	0.027	0.023	0.023
22	332	0.044	0.044	0.044	0.027	0.027	0.028	0.027	0.023	0.023
23	341	0.043	0.042	0.042	0.027	0.032	0.033	0.033	0.023	0.023
24	342	0.042	0.042	0.042	0.033	0.033	0.033	0.033	0.023	0.023
25	411	0.116	0.104	0.106	0.033	0.075	0.077	0.084	0.051	0.053
26	412	0.117	0.104	0.106	0.084	0.075	0.078	0.083	0.051	0.052
27	421	0.122	0.138	0.149	0.083	0.094	0.099	0.116	0.050	0.052
28	422	0.124	0.139	0.149	0.116	0.094	0.098	0.114	0.050	0.051
29	431	0.048	0.048	0.048	0.113	0.036	0.037	0.035	0.024	0.024
30	432	0.048	0.048	0.048	0.036	0.036	0.037	0.035	0.024	0.024
31	441	0.048	0.050	0.050	0.036	0.049	0.050	0.049	0.024	0.024
32	442	0.049	0.049	0.050	0.049	0.049	0.050	0.050	0.023	0.024
33	511	0.105	0.095	0.096	0.050	0.047	0.049	0.050	0.042	0.043
34	512	0.104	0.095	0.096	0.049	0.048	0.049	0.050	0.042	0.043
35	521	0.109	0.089	0.093	0.050	0.053	0.056	0.058	0.042	0.043
36	522	0.107	0.089	0.092	0.057	0.054	0.057	0.059	0.041	0.042
37	531	0.045	0.045	0.045	0.059	0.022	0.022	0.022	0.019	0.019
38	532	0.045	0.045	0.045	0.022	0.022	0.023	0.022	0.019	0.019
39	541	0.044	0.042	0.043	0.022	0.026	0.027	0.027	0.019	0.019
40	542	0.044	0.043	0.043	0.027	0.027	0.027	0.027	0.019	0.019

Table 33. 95% confidence interval for the standard errors of the within-level factor loading of the two-level model, the complex model and the normal model in the W2B1 model (In this table, "m" represents the multilevel model, "c" represents the complex model, and "n" represents the normal model).

Obs	cond	true_m	2.5_m	97.5_m	true_c	2.5_c	97.5_c	true_n	2.5_n	97.5_n
1	111	0.116	0.099	0.101	0.062	0.059	0.060	0.062	0.040	0.041
2	112	0.105	0.091	0.093	0.059	0.057	0.058	0.060	0.038	0.039
3	121	0.117	0.127	0.136	0.088	0.077	0.080	0.088	0.039	0.041
4	122	0.109	0.111	0.118	0.084	0.074	0.077	0.085	0.038	0.040
5	131	0.047	0.046	0.046	0.027	0.028	0.028	0.027	0.018	0.018
6	132	0.045	0.043	0.043	0.026	0.027	0.027	0.026	0.017	0.018
7	141	0.046	0.047	0.048	0.039	0.038	0.039	0.039	0.018	0.018
8	142	0.043	0.043	0.044	0.037	0.037	0.037	0.037	0.017	0.018
9	211	0.113	0.100	0.102	0.095	0.085	0.088	0.095	0.056	0.057
10	212	0.103	0.092	0.094	0.088	0.080	0.082	0.089	0.053	0.054
11	221	0.122	0.132	0.141	0.131	0.108	0.112	0.133	0.054	0.056
12	222	0.110	0.116	0.123	0.121	0.101	0.105	0.123	0.051	0.053
13	231	0.048	0.046	0.047	0.040	0.041	0.041	0.040	0.026	0.026
14	232	0.044	0.043	0.043	0.037	0.038	0.039	0.037	0.024	0.025
15	241	0.047	0.048	0.049	0.057	0.056	0.057	0.057	0.026	0.026
16	242	0.043	0.044	0.044	0.054	0.053	0.054	0.054	0.024	0.025
17	311	0.106	0.092	0.094	0.059	0.057	0.058	0.060	0.048	0.049
18	312	0.097	0.085	0.087	0.056	0.054	0.055	0.056	0.046	0.047
19	321	0.108	0.086	0.089	0.073	0.067	0.069	0.073	0.048	0.050
20	322	0.098	0.078	0.081	0.069	0.063	0.065	0.069	0.046	0.047
21	331	0.045	0.044	0.044	0.026	0.027	0.027	0.026	0.022	0.022
22	332	0.042	0.041	0.041	0.024	0.025	0.026	0.024	0.021	0.021
23	341	0.043	0.042	0.042	0.033	0.032	0.033	0.033	0.022	0.022
24	342	0.040	0.039	0.039	0.031	0.031	0.031	0.031	0.021	0.021

Table 33 (continued)

Obs	cond	true_m	2.5_m	97.5_m	true_c	2.5_c	97.5_c	true_n	2.5_n	97.5_n
25	411	0.116	0.104	0.106	0.084	0.075	0.078	0.084	0.051	0.052
26	412	0.110	0.099	0.101	0.078	0.070	0.072	0.078	0.048	0.049
27	421	0.121	0.141	0.150	0.112	0.095	0.098	0.112	0.049	0.051
28	422	0.115	0.127	0.134	0.103	0.086	0.089	0.104	0.047	0.048
29	431	0.048	0.048	0.048	0.035	0.036	0.037	0.035	0.023	0.024
30	432	0.046	0.046	0.047	0.033	0.034	0.034	0.033	0.022	0.022
31	441	0.049	0.049	0.050	0.050	0.050	0.050	0.050	0.023	0.023
32	442	0.048	0.047	0.048	0.046	0.045	0.046	0.046	0.022	0.022
33	511	0.106	0.095	0.096	0.051	0.048	0.049	0.052	0.041	0.042
34	512	0.102	0.091	0.093	0.048	0.046	0.047	0.049	0.040	0.040
35	521	0.108	0.090	0.093	0.060	0.055	0.057	0.060	0.041	0.042
36	522	0.101	0.085	0.088	0.057	0.052	0.054	0.057	0.039	0.040
37	531	0.045	0.045	0.045	0.022	0.022	0.023	0.022	0.019	0.019
38	532	0.044	0.043	0.044	0.021	0.021	0.022	0.020	0.018	0.018
39	541	0.044	0.043	0.043	0.027	0.027	0.028	0.028	0.019	0.019
40	542	0.042	0.041	0.041	0.026	0.026	0.026	0.026	0.018	0.018

5.0 DISCUSSION

5.1 SUMMARY AND CONCLUSIONS

Researchers are looking into different techniques to deal with the data obtained from the complex sampling design. It is questionable whether the multilevel CFA is superior to the single-level CFA with or without the adjustment of the standard errors. This study used the simulation method to compare multilevel and single-level models in CFA of ordinal items with clustered data. Specifically, the purpose of the study is to compare the accuracy of estimating the factor loading and relative standard errors among the two-level model, complex single-level model with adjusted standard error, and the normal single-level model. The study also aims to examine the impact on the model performance from 1) factorial structure, 2) number of cluster members, 3) number of clusters, 4) factor correlation, 5) factor loading, and 6) Item ICC. The result will be discussed in the order of the proposed research questions:

1. What is the difference in terms of model fit indices calculated from the two-level CFA model, single-level CFA model with normal standard error or complex standard error? What model fit indices, if any, are recommended in model selection?
2. How are three models compared in estimating the within-level factor loading and their standard errors? What design factors may impact the performance of these models?

3. What factors influence the performance of the two-level CFA model in recovering between-level factor loading?
4. What factors affect the performance of the two-level CFA model in recovering residual variance?

5.1.1 Model Fit Indices

All fit indices succeeded in identifying the misfit of the normal single-level model. In the complex single-level CFA model, the model fit indices are also adjusted according to Stapleton, Yang, and Hancock (2016). However, their performance is different. Chi-square *p* value and RMSEA suggest a good fit while CFI and TLI indicate worse fit when compared to the two-level model in the conditions with small sample sizes. The complex model fit the multilevel data well in the conditions of the large sample sizes. In the study of Hsu, Kwok, Lin, and Acosta (2014), it stated that CFI and TLI were sensitive to the misspecification of the factor loading at the within-level. In this study, since the two-level model is the true model, it indicates that the chi-square *p* value and RMSEA is not as sensitive as the CFI and TLI in the small sample size. The difference among RMSEA, CFI, and TLI is that RMSEA is absolute fit index while CFI and TLI are comparative fit indices.

Wu (2010) indicated that when the higher-level was neglected, the $\chi 2$ statistic tended to be small and could not identify the bad model fit. This study corroborates prior research in supporting the use of CFI and TLI as goodness of fit index in the complex single-level CFA model.

5.1.2 Parameter Estimates of Three Models

In addition to the model fit, the recovery of the factor loading at both levels, the recovery of the residual variance in the two-level model, and the standard errors of the within-level factor loadings were examined to decide whether the single-level CFA model could be good alternative to the two-level CFA model in the multilevel data. If the differences among the relative bias of the factor loadings of three models are not large, then it is promising to use the single-level CFA model in the multilevel data and the single-level model is easier to implement than the multilevel model.

However, it was found that only when the factor loadings across levels were the same, the relative bias of the within-level factor loadings estimated from the single-level complex model and normal model were within 0.1, which were acceptable. When factor loadings were different across levels, the relative bias of the within-level factor loading were as large as 0.3. Wu and Kwok (2012) and Porprasertmanit et al. (2014) also found that when the factor loadings were the same across levels, the factor loading estimated from the single-level complex model were more accurate than when the factor loadings were different across levels . So it is thought that when factor loadings across levels are the same and the model fit of the complex model is also good, the complex single-level CFA model could be used to estimate the two-level data.

In addition to the influence of the factor loading across levels, the influence of the ICC should be noted. In the study, it was found that under the condition of same factor loading, the higher ICC resulted in a larger relative bias of factor loading in both complex model and normal model. The effect of ICC on the parameter estimate in the complex model and normal model had also been demonstrated by Pornprasertmanit et al. (2014) and Stochl et al. (2015). The smallest ICC of this study is 0.25. According to the result of the current study, the complex model is not

recommended to replace the two-level model when the ICC was 0.25 or 0.45 because both ICCs resulted with biased estimates of factor loadings especially when factor loadings are different at the two levels.

The relative bias of the between-level factor loading in W2B2 model and W2B1 model looked small in most conditions, indicating the good recovery of the between-level factor loading in the true model. The relative bias of the residual variance looked small in most conditions in the W2B2 model and W2B1 model, indicating the good recovery of the residual variance in the true model. Relative bias of the residual variance was a little larger when the sample size was small than when the sample size was large.

There are some differences in the findings between W2B2 model and W2B1 model comparing the complex model and normal model. The relative bias of the within-level factor loading estimated by the complex model was significantly smaller than that estimated by the normal model for factor loading of 0.5 0.5 while averaged across ICC, sample sizes, and factor correlations in W2B1 model. However, the relative bias of the factor loading estimated by the complex model in the W2B2 model was not significantly different than that estimated by the normal model for various factor loading conditions.

The effect of simulation design factors on the relative bias of the factor loading and standard errors between W2B2 model and W2B1 model were similar in most conditions. In both W2B2 model and W2B1 model, only sample size significantly affected the relative bias of the within-level factor loading in the two-level CFA model; only the interaction of factor loading and ICC significantly affected the relative bias of the factor loading in the complex model and normal model. The most obvious difference was that the between-level factor loading in the

W2B2 model was affected by the interaction of the sample and FactorICC while the between-level factor loading in the W2B1 model was not affected by any of the factors.

As for the standard errors, the standard errors of the within-level factor loading estimated by the normal model was significantly smaller than that estimated by the two-level model and the complex model. This finding agreed with the result from Stapleton (2006) and Julian (2001) that the standard error was biased if the standard error was not adjusted in the complex sampling design. The deflation of the standard errors is concerning when the single-level model is used to estimate the two-level data. In this study, the standard errors estimated from the complex model were just a bit smaller than those estimated from the two-level model, but the difference was still significant and there was no overlap between the 95% confidence interval between the standard errors estimated from the two-level model and complex model.

In conclusion, the estimate of the relative bias of the factor loading from the two-level model was more accurate than that estimated from the complex model. Model fit indices (CFI and TLI) and ANOVA result reached to the same conclusion. It is recommended that when the researcher considers which model should be used for the multilevel data, the researcher should consider not only the model fit, but also the accuracy of the factor loading and the related standard errors.

There is certain circumstance that the research should use the complex model adjusting for the standard error instead of the two-level model in the multilevel data structure. Stapleton, Yang, and Hancock (2016) found that a construct may appear to be meaningful at both levels while in fact the construct was theoretically meaningful at the individual level. In this context, the complex model resulted with better fit than the two-level model looking at the model fit

indices. It is important to examine whether the construct is really a cluster-level construct or it just reflects a spurious clustering effect.

In this study, when the true model is the two-level CFA model, the relative bias of the within-level factor loading of the two-level CFA model was significantly smaller than that of the complex model. In the study of Porprasertmanit et al. (2014), the factor loading could be overestimated or underestimated using the single-level CFA model when factor loadings across levels were different. Julian (2001) also found that the factor loading could be overestimated using the single-level CFA model. Based on the results of the previous studies and the current study, consistent estimates of within-level factor loadings between the two-level CFA model and the complex CFA model might indicate that either the factor loadings are the same at the two levels or there is no true multilevel structure of the data and the clustering effect is actually spurious.

5.2 IMPLICATIONS FOR APPLIED RESEARCH

One of the most relevant implications from the present study for the applied researchers is to decide whether the two-level CFA model or the single-level CFA model should be used in the hierarchical data. In educational and psychological research, it is common to have the data with the multilevel structure, but the information about the higher level might not always be available. Under this circumstance, the design-based approach using adjusted standard error is a good choice. In addition, the single-level CFA model adjusting for the standard error is much easier to implement than the two-level CFA model.

In practice, it is suggested that the analysist should first consider whether the interest of the measurement was at the individual level or at the cluster level. If the interest of the research is at the individual level, the complex single-level model is a good choice. The researcher still need to examine the standard error of the parameter estimate of the complex model. One reason that the single-level model was criticized was the deflation of the standard error. Although the standard error was adjusted in the complex model, there was still significant difference in the standard error between the complex model and the two-level model in the current study. If the standard error looked unreasonably small, the researcher could perform the two-level CFA model and compare the standard errors between two models.

Even when the interest of the research is at the individual level, the researchers could perform the two-level model. It is possible that the target data is not suitable for the single-level analysis. It is important to perform the model fit of two models. If the two-level model fit the data but the single-level model does not fit the data, the data is not suitable for the single-level analysis.

On the other hand, if the researcher's interest is at the cluster level and the data does look like having the multilevel structure, there still exists the possibility that the clustering effect is spurious like it is described in the study of Stapleton, Yang, and Hancock (2016). Under this circumstance, it is important for the researcher to compare the multilevel model with the complex model. A better model fit of a multilevel model suggests the data truly has the multilevel structure, while a similar model fit suggests possibility of spurious between-level factor.

5.3 LIMITAIONS AND FUTURE DIRECTIONS

Limitations of the current study were recognized. First, the combination of the within-level and between-level factor loadings was limited to 0.5 0.8, 0.5 0.5, and 0.8 0.5 and the ICC was limited to 0.45 and 0.25. Porprasertmanit et al. (2014) used 5 levels of ICC and Stochl et al. (2015) used 11 levels of ICC. In the future, more combinations of the factor loading and ICC should be conducted especially the low ICCs such as 0.10.

Second, this study did not investigate the condition while the between-level factor structure is more complex than the within-level one. Even though such condition is not common in applied educational and psychological studies, Wu and Kwok (2012) examined this condition and found that the single-level model demonstrated the poor fit and the factor loading was seriously biased. Julian (2001) also examined this condition and found moderate bias of the factor loading. In the future, the condition with the between-level structure more complex than the within-level structure can be examined.

APPENDIX A

MPLUS CODE FOR DATA GENERATION OF W2B2 MODEL WITH FACTOR LOAING OF 0.5 AT WITHIN-LEVEL, 0.8 AT BETWEEN-LEVEL, ICC OF 0.45, AND FACTOR CORRELATION OF 0.3 AT BOTH LEVELS

```
MONTECARLO:
      NAMES ARE u1-u10;
      NOBSERVATIONS = 500;
      SEED = 4526;
      GENERATE=u1-u10(4);
      CATEGORICAL=u1-u10;
      NREPS=50;
      NCSIZES=1;
      CSIZES=50(10);
      REPSAVE=ALL;
   save=model1CFA*.dat;
MODEL POPULATION:
      %WITHIN%
     fw1 by u1*0.5 u2*0.5 u3*0.5 u4*0.5 u5*0.5 ;
      fw2 by  u6*0.5 u7*0.5 u8*0.5 u9*0.5 u10*0.5 ;
       fw1 with fw2*0.30;
    fw1*1;
  fw2*1;


      %BETWEEN%
    fb1 by u1*0.8 u2*0.8 u3*0.8 u4*0.8 u5*0.8 ;
     fb2 by  u6*0.8 u7*0.8 u8*0.8 u9*0.8 u10*0.8 ;
      fb1 with fb2*0.30;
    fb1*1;
  fb2*1;

  u1*0.383;
  u2*0.383;
u3*0.383;
```

```
u4*0.383;
  u5*0.383;u6*0.383;
  u7*0.383;u8*0.383;
       u9*0.383;
u10*0.383;


  [u1$1*-1.464 u2$1*-1.410 u3$1*-1.829 u4$1*-1.591
  u5$1*-1.150 u6$1*-2.081 u7$1*-1.274 u8$1*-1.384
  u9$1*-0.509 u10$1*-0.584];
  [u1$2*-1.298 u2$2*-1.273 u3$2*-1.647 u4$2*-1.422
   u5$2*-0.739 u6$2*-1.884 u7$2*-0.980 u8$2*-1.145
  u9$2*0.049 u10$2*0.561];
  [u1$3*-0.258 u2$3*-0.252 u3$3*-0.632 u4$3*-0.332
    u5$3*-0.135 u6$3*-1.063 u7$3*0.069 u8$3*-0.082
   u9$3*1.051 u10$3*1.319 ];
   [u1$4*0.682 u2$4*0.719 u3$4*0.261 u4$4*0.605
   u5$4*0.470 u6$4*-0.262 u7$4*0.929 u8$4*0.886
  u9$4*1.653 u10$4*1.793 ];

        ANALYSIS:
             TYPE=TWOLEVEL;
             ESTIMATOR=WLSMV;
      !       parameterization=theta;
```

BIBLIOGRAPHY

Arnold-Berkovits, I(2002). Structural modeling with order polytomous and continuous variables: a simulation study comparing full-information Bayesian estimation to correlation/covariance methods. *Unpublished doctoral dissertation, University of Maryland.*

Asparouhov, T. & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. *Proceedings of the 2007 JSM meeting in Salt Lake City, Utah, Section on Statistics in Epidemiology.*

Asparouhov, T. & Muthen, B. (2006). Multilevel modeling of complex survey data. *Proceedings of the Joint Statistical Meeting in Seattle, August 2006. ASA section on Survey Research Methods, 2718-2726.*

Asparouhov, T. & Muthen, B. (2006). Comparison of estimation methods for complex survey data analysis. *Download paper.*

Beauducel, A., & Herzberg, P. Y. (2006*).* On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13, 186–203.*

Bollen, K. A.(1989). Structural equations with latent variables. *New York, NY: Wiley*

Bollen, K. A., Tueller, S. J. and Oberski, D.(2013). Issues in the Structural Equation Modeling of Complex Survey Data. *59th ISI World Statistics Congress, Hong Kong (Session STS010).*

Breevaart, K. et al.(2012). The Measurement of State Work Engagement : A Multilevel Factor analytic Study. *The Journal of Psychological Assessment, vol. 28, No. 4.*

Brondino, M. & Pasini, M. & da Silva, S. C. A.(2013). Development and Validation of an Integrated Organizational Safety Climate Questionnaire with Multilevel Confirmatory Factor Analysis. *Qual Quant , 47.*

Cassidy, D. J., Hestenes, L. L., Hegde, A., Hestenes, S., & Mims, S. (2005). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the early childhood environment rating scale-revised. *Early Childhood Research Quarterly, 20, 345–360.*

Cheah B. Clustering standard errors for modeling multilevel data(2009). *Working paper, Columbia University (https://sites.google.com/site/bancheah/).*

Dedrick, R. F. and Greenbaum, P. E.(2010). Multilevel Confirmatory Factor Analysis of a Scale Measuring Interagency Collaboration of Children's Mental Health Agencies. *Journal of Emotional and Behavioral Disorders,1-14.*

Distefano C.(2002). The Impact of Categorization with Confirmatory Factor Analysis. *Structural Equation Modeling, 9(3).*

Dolan, C.V.(1994). Factor analysis of variables with 2,3,5,and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematics and Statistics, vol 47.*

Dyer, N. G., Hanges, P. J. and Hall, R.J.(2005). Applying Multilevel Confirmatory Factor Analysis Techniques to the Study of Leadership. *The Leadership Quarterly, Vol 16.*

Edwards, M. C. (2010). A Markov Chain Monte Carlo Approach to Confirmatory Item Factor Analysis. *Psychometrika, Vol. 75, No. 3.*

Flora. D. B. and Curran, P. J.(2004). An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis with Ordinal Data. *Psychological Methods, Vol. 9. No. 4.*

Finch, J. F.,West, S. G.,&MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling, 4, 87–107.*

Greenbaum et al.(2011). Multilevel Confirmatory Factor analysis of the systems of Care Implementation Survey. *Journal of Behavioral Health Services and Research. Vol. 2.*

Grilli, L. and Rampichini, C.(2007). Multilevel Factor Models for Ordinal Variables. *Journal of Structural Equation Modeling, 14:1, 1-25.*

Gajewski, B. J. et al.(2010). A multilevel Confirmatory Factor Analysis of the Practice Environment Scale : A Case Study. *Nursing Research, Vol 59, No 2.*

Gay et al.(2006). Educational Research: Competencies for Analysis and Applications. *E. Mills, Peter Airasian, 8th Edition.*

Green, S. B., Akey, T. M., Fleming, K. K., Hershberger, S. L., & Marquis, J. G. (1997). Effect of the number of scale points on chi-square fit indices in confirmatory factor analysis. *Structural Equation Modeling, 4, 108–120.*

Grilli L. & Rampichini C. (2007) Multilevel factor models for ordinal variables. *Structural Equation Modeling, 14(1).*

Hofmann, D. A.(1997). An Overview of the Logic and Rationale of Hierarchical Linear Models. *Journal of Management, vol. 23, No. 6.*

Haenens, E., Van Damme, J. a, and Onghena, P.(2012). Constructing Measures for School Process Variables: The Potential of Multilevel Confirmatory Factor Analysis. *Quality and Quantity, Vol. 46.*

Huang, F. L.(2014). Using a Bifactor Model to Assess the Factor Structure of the Phonological Awareness Literacy Screening for Grade I Through 3. *Journal of Psychoeducational Assessment, Vol. 32(7).*

Hsu , H-Y, et al.(2014). Detecting Misspecified Multilevel Structural Equation Models with Common Fit Indices: A Monte Carlo Study. *Multivariate Behavioral Research: Vol. 50, No. 2.*

Hox, J. J. and C.J. M.Maas(2001). The Accuracy of Multilevel Structural Equation Modeling with Pseudobalanced Groups and Small Samples. *Structural Equation Modeling 8, 157-174.*

Hox, J. J.(1998). An introduction to structural equation modelling. *Family Science Review, 11.*

Hedges, L. V. and Hedberg, E. C.(2007). Intraclass Correlations for Planning Group Randomized Experiments in Rural Education. *Journal of Research in Rural Education, 22(10).*

Jackson, D. L., Gillapsy, J. A. and Pruc-Stephenson, R.(2009). Reporting Practices in Confirmatory Factor Analysis: An Overview and Some Recommendation. *Psychological Methods, Vol.14, No.1.*

Julian, M. W.(2001) The Consequences of Ignoring Multilevel Data Structures in Nonhierarchical Covariance Modeling. *Structural Equation Modeling, 8(3).*

Kalton, G.(1983a). Models in the Practice of Survey Sampling. *International Statistical Review, Vol. 51.*

Kaplan, D.(2009). Structural Equation Modeling: Foundations and Extensions(2nd ed). *Thousand Oaks, CA: Sage.*

Klangphahol, K. , Traiwichitkhun, D., and Kanchanawasi, S.(2010). Applying multilevel confirmatory factor analysis techniques to perceived homework quality, *Scholar, Vol 2, No.2.*

Kim, E. S. and Yoon, M.(2011). Testing Measurement Invariance: A Comparison of Multiple-Group Categorical CFA and IRT. *Structural Equation Modeling, 18.*

Little, J.(2013). Multilevel Confirmatory Ordinal Factor Analysis of the Life Skills Profile-16. *Journal of Psychological Assessment, Vol. 25, No.3.*

Lei, P.W.(2009). Evaluating Estimation Methods for Ordinal Data in Structural Equation Modeling, *Quality and Quantity, No. 43.*

Li, Cheng-Hsien(2015). Confirmatory Factor Analysis with Ordinal Data: Comparing Robust Maximum Likelihood and Diagonally Weighted Least Square. *Journal of Behavioral Research Methods, Jul 15.*

MAAS, C. J. M. and Hox, J. J.(2004). The Influence of Violation of Assumptions on Multilevel Parameter Estimates and Their Standard Errors. *Computational Statistics and Data Analysis, Vol. 46.*

Moulton, Brent (1990), An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units. *The Review of Economics and Statistics, Vol. 72, No. 2, pp. 334-338.*

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39, 129–149.*

Muthén, B. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28, 338–354.*

Muthen, L. K. & Muthen, B. O.(1998-2012). Mplus user's guide(7th edition). *Los Angeles, CA: Muthen & Muthen.*

Muthe´n, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Unpublished manuscript.*

Muthen, B.O(1994). Multilevel covariance structure analysis. *Sociol. Methods Res. 22, 376–398.*

Muthén, B., & Satorra, A. (1991). Complex sample data in structural equation modeling. *Unpublished manuscript.*

Oranje, A. (2003). Comparison of estimation methods in factor analysis with categorized variables: Applications to NAEP data. *Paper presented at the annual meeting of the American Educational Research Association, Chicago.*

Opdenakker, M.-C. & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement, 11, pp. 103-130.*

Pornprasertmanit, S., Lee, J. and Preacher, K.J.(2014). Ignoring Clustering in Confirmatory Factor Analysis: Some Consequences for Model Fit and Standardized Parameter Estimates. *Multivariate Behavioral Research, Vol. 49.*

Raykov, T. (2012). Scale construction and development using structural equation modeling. *In R. H. Hoyle (Ed.), Handbook of structural equation modeling (pp. 472–492). New York, NY: Guildford Press.*

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd Ed.). *Thousand Oaks, CA: Sage.*

Ryu,E.(2014). Factorial Invariance in Multilevel Confirmatory Factor Analysis. *British Journal of Mathematical and Statistical Psychology, Vol. 67.*

Ryu, E., and West, S. G.(2009). Level-Specific Evaluation of Model Fit in Multilevel Structural Equation Modeling. *Structural Equation Modeling, 16(4).*

Rhemtiulla, M., Brosseau-Liard, P. E., and Savalei, V.(2012). When Can Categorical Variables Be Treated as Continuous? A Comparison of Robust Continuous and Categorical SEM Estimation Methods Under Suboptimal Conditions? *Psychological Methods, Vol. 27, No.3.*

Schmidt, W. (1969). Covariance structure analysis of the multivariate random effects model. *Unpublished doctoral dissertation, University of Chicago.*

Stochl, J., Jones, et al.(2015). Effects of Ignoring Clustered Data Structure in Confirmatory Factor Analysis of Ordered Polytomous Items: A Simulation Study Based on PANSS. *International Journal of Methods in Psychiatric Research. Epub 2015 Jun 20 Wiley Online Library.*

Schmitt, T. A.(2011) Current Methodological Considerations in Exploratory and Confirmatory factor Analysis. *Journal of Psychoeducational Assessment, Vol. 29, No.,4.*

Satorra, A., & Muthén, B. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25, pp. 267-316.*

Satorra, A.(1992). Asymptotic Robust Inferences in the Analysis of Mean and Covariance Structures. *Sociological Methodology, Vol. 22.*

Satorra, C., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. *In A. von Eye & C. C. Clogg (Eds.), Latent variable analysis: Applications for developmental research (pp. 399–419). Thousand Oaks, CA: Sage.*

Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling, 13(1), 28-58.*

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9, 475–502.*

Stapleton, L. M., Yang, J. S. & Hancock, G. R.(2016). Construct Meaning in Multilevel Settings, *Journal of Educational and Behavioral Statistics(2016). Vol. 41, No. 5.*

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sample designs. *Research in Higher Education, 42, 517–540.*

Walker, D. A., & Young, D. Y. (2003). Example of the impact of weights and design effects on contingency tables and chi-square analyses. *Journal of Modern Applied Statistical Methods, 2, 425–432.*

Whitton, S. M. and Fletcher, R. B.(2014). The Group Environment Questionnaire: A Multilevel Confirmatory Factor Analysis. *Small Group Research, Vol. 45(1).*

Wirth.,R.J. and Edward, M. C.(2007), Item Factor Analysis: Current Approaches and Future Directions. *Psychological Methods(2007), Vol 12, No. 1, pp.58-79.*

Wu, J. Y(2010). Comparing Model-Based and Design-Based Strucutral Equation Modeling Approach in Analyzing Complex Survey Data, *Unpublished Dissertation.*

Wu, J. Y., & Kwok, O. M. (2012). Using SEM to analyze complex survey data: a comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling: A Multidisciplinary Journal, 19(1),pp.16-35.*

Wu, C-H.(2009). Factor Analysis of the General Self-Efficacy Scale and Its Relationship with Individualism/Collectivism among Twenty-Five Countries: Application of Multilevel Confirmatory Factor Analysis. *Personality and Individual Differences, Vol. 46.*

Yang-Wallentin, F, Joreskog, K.G., and Luo, H.(2010). Confirmatory Factor Analysis of Ordinal Variables with Misspecified Models. *Structural Equation Modeling, 17.*

Yu, C.-Y. (2002). Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes. *Unpublished doctoral dissertation, University of California, Los Angeles.*

Yuan, K. H. and Bentler P. M.(2007). Multilevel Covariance Structure Analysis by Fitting Multiple Single-Level Models. *Sociological Methodology, Vol. 37 , pp. 53-82.*

Zhang , N. J. and Wan, T. T. H.(2005). The Measurement of Nursing Home Quality: Multilevel Confirmatory Factor Analysis of Panel Data. *Journal of Medical Systems, vol. 29, No.4.*

Zimprich, D.,& Perren, S. & Hornung, R.(2005). A Two-Level Confirmatory Factor Analysis of a Modified Rosenberg Self-Esteem Scale. *Educational and Psychological Measurement, Vo. 65, No. 3.*