Title
Limited Information Estimation and Model Fit Evaluation: Towards Quantifying Complexity in Item Response Models

Author
Suh, Yon Soo

Publication Date
2022

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Limited Information Estimation and Model Fit Evaluation: Towards Quantifying Complexity in Item Response Models

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Education

by

Yon Soo Suh

2022

ABSTRACT OF THE DISSERTATION

Limited Information Estimation and Model Fit Evaluation:
Towards Quantifying Complexity in Item Response Models

by

Yon Soo Suh
Doctor of Philosophy in Education
University of California, Los Angeles, 2022
Professor Li Cai, Chair

In model selection, we seek a balance between goodness-of-fit and generalizability, for which model complexity is key. Fitting Propensity (FP) has been suggested as an ideal measure of complexity that refers to a model's inherent flexibility to fit diverse patterns of data, all else being equal. Assessing the FP of item response theory (IRT) models requires random and uniform sampling of all item response patterns for a set of items and fitting the sampled data to one or more models repeatedly many times. The model fit information across the replications is summarized for each model and examined. In the case of multiple models, comparisons between models are also made. Computational issues due to the high-dimensional discrete space involved in the generation of random datasets have rendered it infeasible to investigate FP for more than a handful of dichotomously scored items under the conventional full information (FI) approach of the multinomial framework.

This study turns to limited information (LI) methods as an alternative, capitalizing on the fact that IRT models can be realized as contingency tables using marginal probabilities. LI methods use information from only the lower-order, usually univariate and bivariate, margins of IRT models as opposed to full response patterns. Thus, they not only significantly reduce the number of response probabilities to be generated in the first place but can also make model estimation computationally simpler. The computational gain afforded by the proposed LI approach opens doors for investigating the FP of more complex IRT modeling schemes which traditionally require many more response patterns.

A data-generating algorithm founded on classical literature on sampling contingency tables with fixed margins along with sequential importance sampling (SIS) of contingency tables is introduced for random and uniformly distributed sampling across all univariate and bivariate margins of items. To estimate the data consisting of solely the lower-order margins, a pairwise marginal maximum likelihood (PMML) estimator tailored to fit a wide variety of IRT models is introduced. Lastly, the feasibility of the proposed LI data generation algorithm and estimation approach to assess the FP of IRT models is tested under various combinations of data sampling and estimation methods.

The dissertation of Yon Soo Suh is approved.

Minjeong Jeon

Mark P. Hansen

Wesley E. Bonifay

Li Cai, Committee Chair

University of California, Los Angeles

2022

To my mother, Dr. Hey Jung Jun

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am indebted to many people because of whom my graduate experience has been one that I will cherish forever. My deepest gratitude is to my advisor, Dr. Li Cai. He granted me the freedom to explore and grow on my own whilst making sure I recovered whenever my steps faltered. I will always be in awe of his brilliance, thought-provoking perceptions and commitment to research as well as care for his students, which I aspire to one day resemble. I would also like to express my gratitude to my other committee members. Dr. Minjeong Jeon is truly a wonder woman and an inspiration to all with her capability and passion for research. Dr. Mark Hansen is always so patient and willing to offer keen insights, solutions, and encouragement. Dr. Wes Bonifay provided the inspiration for this dissertation and thus, was instrumental to my graduation with his intuitive comments and constructive criticisms.

I wholeheartedly thank each member of my fellow dissertation-writing "comrades": Drs. Sijia Huang, Shujin Zhong, Meredith Langi, and Mariana Barragan. With them, I have had the pleasure of sharing all the ups and downs and tears and laughter of graduate life. Kudos to us all. I want to also acknowledge my friends fondly within the Education department: Jevan Luo, Maria Paz Fernandez, Marlene Saint Martin Guerra, Nadia Sabat Bass, and Fernando Mora and others. Special thanks to Minho Lee, Teanna Feng, and Preston Botter who brainstormed with me, laughed with me, and patiently put up with my incessant whining. Good luck to my new lab members Yun Kim and Ryan Lerch. I am also grateful to Dr. Amy Gershon who always saved me from a pinch.

I want to also express my sincere appreciation for my former and present colleagues. Thank you to Drs. Greg Chung, Elizabeth Redman, Eunhee Keum, Alex Sturm, Markus Iseli, Emily Relkin, and Kilchan Choi from CRESST. Working with such talented and motivated researchers as them opened my eyes as well as many doors. Special

thanks to Dr. Greg Chung for his professional and personal support and general life advice, along with orchestrating many good times. I am also extremely grateful for every member of my wonderful and inspiring collegues at NWEA, who have welcomedly me warmly and cheered me on. Thank you particularly to my manager, Dr. Tyler Matta, for being so accommodating and understanding.

Most importantly, none of this would have been possible without the love and patience of my parents, grandparents, and siblings. I am, beyond doubt, privileged because of them. Thank you to my sister, Yon Jin Suh, for editing and proofreading the dissertation and for keeping me sane. I have also had the opportunity to welcome new family members in the form of my husband and in-laws. I could not have done this without my husband, Jin Ho Kang, by my side.

As I embark on the next chapter of my academic journey, I will continuously look back with fondness and gratitude on all the invaluable experiences and each and every one of my family, friends, advisors, and colleagues.

VITA

EDUCATION

2014	Bachelor of Arts, English Language and Literature, Yonsei University, Republic of Korea
2016	Master of Arts, Education Yonsei University, Republic of Korea

WORK

2022-Present	Research Scientist NWEA
2016-2022	Graduate Student Researcher National Center for Research on Evaluation, Standards, and Student Testing (CRESST),University of California, Los Angeles
2015-2016	Research Assistant Yonsei University, Republic of Korea

PUBLICATIONS

Cai, L. & Suh, Y. S. (In press). Diagnostic classification models. In B. Frey (Ed.), *The SAGE encyclopedia of Research Design*. SAGE Publications, Inc.

Suh, Y. S., Hwang, D., Quan, M., & Lee, G. (2016). Optimizing the Costs and GT based reliabilities of Large-scale Performance Assessments. In *Quantitative Psychology Research* (pp. 173-185). Springer, Cham.

Suh, Y. J., & Suh, Y. S. (2013). The Student-Teacher-Facilitator role of Corporations in Human rights. *Human rights law review, 10,* 110-143.

CHAPTER I

Introduction

1.1 Research Background

Models link theory to observed data and shed light on the processes of data generation that are simplified or approximated representations of reality. Box (1976) stated that "all models are wrong, but some are useful." While models being the "truth" have been a topic of debate, most can agree that all models are not equal, but some are better than others in capturing important aspects of various phenomena. In order to search for these more useful models, model parameters need to be estimated appropriately. Furthermore, after a model is fit, we require some measure with which to evaluate the usefulness or appropriateness of the model, often in comparison with competing models. Model estimation and model evaluation are interrelated to influence each other (Myung et al., 2003).

Prevalent in the social and behavioral sciences are models relating observable phenomena to underlying and unobserved causes, termed latent variables (Cai et al., 2016). Frequently, the observed data are categorical item-level responses such as questionnaire or test items with the intent of measuring and understanding continuous latent variables. These latent variables can represent educational achievement, attitudes, and personality. For example, PISA or NAEP routinely give out multiple-choice tests to gauge students' latent academic profeciences in various school subjects (Cai et al., 2016). Item response theory (IRT) refers to the latent variable modeling of multivariate categorical response data. IRT resulted from expanding common factor analysis (FA) techniques for continuous observed variables to categorical data (Cai & Thissen, 2014;

Cai & Moustaki, 2018; Jöreskog & Moustaki, 2001). As such, the purpose is very much the same: summarize the dependence structure among a set of categorical variables as alternative, low-dimensional representations using a small number of latent factors, often called abilities.

Over the years, a plethora of different IRT models has been conceived and applied. Technically, any kind of latent variable model can be constructed as long as it appropriately specifies the measurement and structural model (Cai, 2012). However, for IRT models, the difficulty is not with constructing the models per se but with model estimation (Wirth & Edward, 2007) and model fit (Maydeu-Olivares & Joe, 2014). The estimation and model fit of IRT models have been impeded by problems arising due to the variables' categorical nature, such as high-dimensional numerical integration and sparseness (Cai & Moustaki, 2018; Cai & Hansen, 2013). Despite considerable strides made regarding both model estimation and evaluation, they are still active areas of research, especially in light of the continuous development of new and often complicated models coupled with the rise of large-scale datasets.

This study aims to contribute to the model estimation as well as model evaluation of IRT models based on limited-information (LI) methods (e.g., Bolt, 2005). LI methods differ from its counterpart, full information (FI) methods, in terms of the amount of information extracted from the item response matrix (i.e., item response patterns for all examinees). To elaborate, FI methods use all the information from the data, while LI methods use information from only the lower-order margins. Most LI methods are classified as bivariate information methods as they consist of only the first-order (univariate) and second-order (bivariate) margins. Naturally, they are computationally simpler than FI methods (Bolt, 2005; Cai et al., 2006). The

computational feasibility afforded by LI methods is advantageous in more ways than one in the context of the motivation of this research.

The motivation for a LI approach to model estimation and evaluation in this dissertation mainly stems from the desire to contribute to the study of fitting propensity (FP; Preacher, 2006) of IRT models. FP was proposed as a method of evaluating the utility of a model with an emphasis on different types of model complexity and not simply goodness-of-fit (GoF). The reasoning is that a model may fit a dataset better, not because it is a better data-generating model reflecting reality, but because of its tendency to fit any data better. This tendency is likely to increase as models become more complex. FP or model complexity refers to a model's inherent flexibility to fit diverse data patterns relevant to a particular modeling domain. An intuitive method for studying FP entails a data-generating mechanism for generating randomly and uniformly sampled datasets from the complete data space. Random representative data are repeatedly generated and are fit using candidate theoretical model(s) many times and information on the unadjusted fit is recorded. The summary of this information across replications measures of how well each model fits such random data. If the model fits a large amount of the random data space, it is said to have high FP, and thus, we should be more careful in concluding good model fit of the particular model to a particular dataset. On the other hand, if the model fits only a small amount of the data space, we can be more confident about good model fit results. In addition, when multiple models are involved, the FP of models can also be compared, making it useful for relative model fit evaluation and in model selection (Bonifay & Cai, 2017; Falk & Muthukrishna, 2020; Preacher, 2006).

Bonifay and Cai (2017) evaluated the FP of five dichotomous IRT models by fitting many random datasets of full response patterns generated by a simplex sampling method proposed by Rubin (1981) and summarizing the results using IRT model indices

(e.g., Y2/*N*). They found evidence that the conventional method of quantifying model complexity in terms of the number of freely estimated parameters can provide an inadequate and misleading picture regarding model fit. Thus, they advocated for the need to consider functional form complexity along with the number of free parameters in model evaluation. However, a key limitation of their approach is that the number of all possible response patterns to be randomly sampled grows exponentially with the number of items. The problem is only further exacerbated if the number of response categories increases (Figure 1. 1-(A)). More specifically, the total number of response patterns is equal to $\prod_1^J K_j$ where K_j refers to the number of categories for an item j ($j = 1 \cdots J$), which is in line with traditional FI methods grounded in the multinomial framework. Bonifay and Cai's (2017) sampling method quickly becomes computationally infeasible to substantially limit the number and type of items examinable.

A potential solution may lie in LI methods, capitalizing on the fact that response patterns over all individuals can be collapsed into lower-order margins with roots in the multivariate Bernoulli (MVB) framework. Instead of simulating datasets as full multinomial contingency tables where each cell denotes the frequency of a specific response pattern, we would simulate data for only the lower-order margins. Setting J to be the total number of items, only $J + \frac{J(J-1)}{2}$ first and second-order margins need to be fitted when focusing on simply the lower-order margins. This equates to the total number of probabilities involved being $\sum_1^J K_j + \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} K_j K_{j'}$, which is a significant reduction of data elements or response probabilities to be generated, as clearly shown in Figure 1. 1-(B), when compared to sampling full multinomial probabilities. Furthermore, there is the added benefit to model estimation as using lower-order margins is computationally simpler than that using full multinomial cell probabilities.

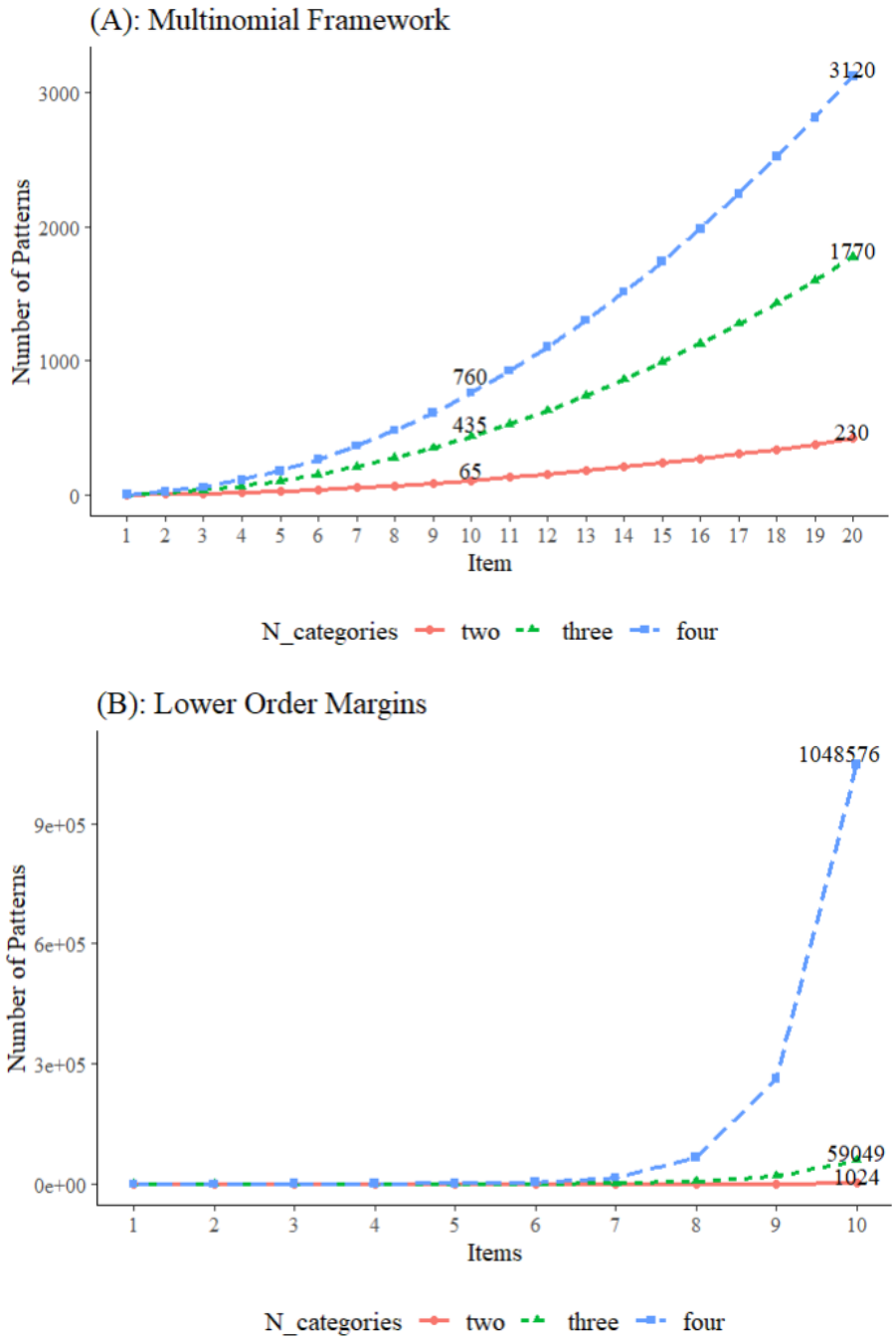


Figure 1. 1: Number of Generated Data Patterns for Full Information versus Limited Information Approach

The challenge is developing an algorithm capable of random and uniform data generation across such lower-order margins, for which the literature on contingency tables with fixed margins proves instrumental. Moreover, as the sampled datasets only

include information about the univariate and bivariate margins, they require LI estimation methods that only use information from these low-order margins to derive parameter estimates. While LI estimation methods are far from new in IRT, they have been restricted to a handful of IRT models that can be estimated with a normal distribution or a probit link function (e.g., 1- and 2-parameter IRT models) under the underlying variable (UV) approach to factor analysis (FA) with categorical variables. The investigation of FP for a wider range of IRT models necessitates a more generalizable LI estimation method based on the IRT approach. Theoretically, this is possible through composite maximum likelihood (CML) estimation (Lindsay, 1988; Varin et al., 2011), based on pseudo-likelihood functions made up of the marginal or conditional distributions of univariate and bivariate responses with a logit link function.

1.2 Research Objectives

Complications due to the categorical nature of item response data sparked much research on the estimation and evaluation of IRT models. FP is an alternative measure of model complexity or parsimony that has recently joined the conversation of model evaluation. Its unique contribution to the model evaluation of IRT models is that it considers the functional form of a model, termed structural complexity, as well as the number of parameters, called parametric complexity (Preacher, 2006). Motivated by the computational issues in generating and estimating many response patterns required to broaden the investigation of FP to a wider range of IRT models, the study proposes a different flavor of model estimation and evaluation using FP based on LI methods. In order to achieve this purpose, the specific research objectives of this study were:

1) to develop a novel data generation method that can produce simulated data randomly and uniformly distributed over the complete categorical data space defined by the lower-order margins.

2) to derive a CML estimator tailored to the IRT approach capable of fitting IRT models using only data from the margins.

3) to validate the proposed data generation algorithm and complementing estimation method to examine the FP of IRT models.

1.3 Research Contributions

This research focused on LI methods has clear advantages for both the model estimation and model evaluation of IRT models in terms of computational feasibility. The use of first- and second-order margins involve only probabilities up to pairs of items, as opposed to the full multinomial probabilities of FI methods. As seen in Figure 1. 1, this leads to large differences in the data elements involved, which become increasingly larger as the number of items, response categories, and factors increase (Bolt, 2005). The significant reduction in computational burden to model estimation and model fitting by employing LI methods paves the way to easily generalize the quantification of FP to IRT models consisting of many items or factors as well as to those with more than two response categories. Considering the continuous rise in large-scale IRT problems where tests consist of many questions along with the likelihood of including polytomously scored items, the computational trackability of LI methods will likely become increasingly favorable.

Furthermore, this study can provide insights into the trade-off relationship between statistical and computational efficiency. Although the loss of information on the higher-order margins hinders the statistical efficiency of LI methods, they can be

appropriate and even preferable if the gains in terms of computational efficiency significantly outweigh this loss. Like this, LI methods can provide insights into whether additional information obtained from higher-order margins such as those from FI leads to meaningful differences relative to LI methods and if so, how much, either for estimated parameters or the assessment of model fit. Through this process, it makes it also possible to study the relative contribution of information of each margin to parameter identification or model misfit because LI methods can easily be decomposed into simple additive pieces (Cai et al., 2006). Furthermore, because the data is generated to be random (i.e., data with no *a priori* underlying structure), there will be issues of model misfit or model misspecification by definition. Results on the effects of model misspecification have yet to be adequately answered for both FI and LI methods. Comparisons on the effects of model misfit between FI and LI methods are also insufficient. The results of this study may shed some light on the behavior of IRT models under both approaches when their assumptions may not hold.

Lastly, this study adds to the literature on data generation mechanisms for binary and categorical variables with a focus on contingency tables, which can be useful for various simulation studies. It can easily generate many different data patterns ranging from plausible to possible (Preacher, 2006; Roberts & Pashler, 2000) based on factors such as model parametrization, estimators, and inferences of interest. Moreover, there is the added benefit of being able to investigate how different sampling schemes targeting different data spaces can influence model estimation and model fit results. In fact, comparison between methods plays an integral role in this study when validating the proposed data generation algorithm and estimation method.

CHAPTER II

Model Evaluation and Fitting Propensity

Model evaluation is essential as the parameters of any statistical model are interpretable only to the extent that the model fits the data and theory. Statisticians are usually interested in quantifiable measures of model evaluation. Bonifay (2015) identified three seemingly dissimilar schools of thought for model evaluation: frequentist statistics, Bayesian inference, and information theory. He sought to develop a theoretical framework that integrated all three perspectives. For this purpose, he focused on the information-theoretic approach of minimum description length (MDL) via the notion of FP (Preacher, 2006), which had been a void in considerations for IRT model fit. This chapter summarizes different aspects of model evaluation to describe the concept and motivation behind FP.

2.1 Bias-Variance Trade-off

Universal to all statistical models is the bias-variance trade-off (Rashidi et al., 2019). Bias refers to the inability of a model to capture enough about the relationship between the variables that is being implied by the dataset. A model with high bias oversimplifies the model to miss the relevant systematic relations or signals. This is called underfitting. Conversely, variance measures a model's tendency to learn too much about the relationship between variables by a certain dataset. A model with high variance may fit a particular dataset well but will not generalize to a different dataset because it is be modeling random noise in the data along with the signal. This is overfitting (Yu, Wang, & Lai, 2005). In an ideal world, one would be able to find a model that simultaneously captures all the regularities of a dataset and generalizes well to

other datasets. However, the reality is that we must seek an optimal balance by trading off between simplifying the modeled relationship (i.e., reducing variance but likely introducing bias) and trying to fit a model more closely to observed values (i.e., reducing bias but likely introducing variance). The goal is to minimize the model's total error and find the most generalizable model (Rashidi et al., 2019; Yu et al., 2005). Total error consists of reducible and irreducible error. The difference between the two errors is whether they can be reduced by choosing a better model. Irreducible error arises from randomness or natural variability and thus is noise that can't be reduced by modeling efforts. Reproducible error can be further decomposed as error due to bias and error due to variance. Reducing reproducible error is the goal. The trade-off relationship is depicted in Figure 2. 1. As seen by the figure, central to this is model complexity (Yu et al., 2005).

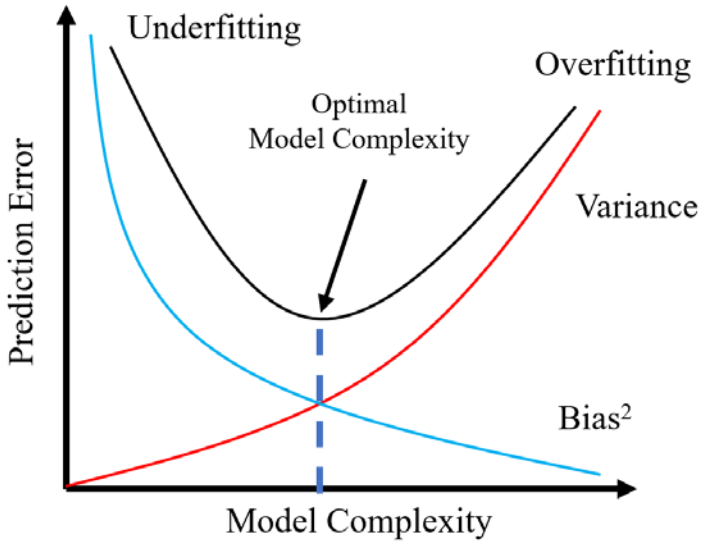


Figure 2. 1:Bias-varianceTrade-Off

Note. Adapted from "An integrated data preparation scheme for neural network data analysis", by L. Yu, S. Wang, and K.K. Lai, 2005, *IEEE Transactions on Knowledge and Data Engineering, 18*(2), p 226.

2.2 Model Evaluation Criteria

Myung et al. (2005) list four quantitative criteria of model evaluation—falsifiability, goodness-of-fit (GoF), simplicity/complexity, and generalizability. The three criteria excluding falsifiability, which is that the model must be falsifiable, are closely related to the concepts introduced in Section 2.1.

GoF represents a model's ability to explain a particular dataset by quantifying the discrepancy between the fitted model and the observed data values (Maydeu-Olivares, 2013). GoF indices are a common method of model evaluation in the social sciences. In fact, it is often the sole measure employed to gauge the fit of a model (Bonifay, 2015). This is based on the logic that the model providing the closest fit to the data is the one that best reflects the underlying regularity (Myung, 2000; Myung & Pitt, 2004). However, this overlooks the role of model complexity (Falk & Muthukrishna, 2020). Myung et al. (2005) define complexity as "a model's inherent flexibility that enables it to fit a wide range of data patterns" (p. 12). Preacher (2006) describes it as the complement of parsimony. Highly complex models can provide good fit, especially in comparison to simpler models, not because it is a better representation of the data-generating process, but simply because of its inherent tendency to fit any data or random noise better (Bonifay & Cai, 2017; Falk & Muthukrishna, 2020; Myung, 2000; Preacher, 2006). GoF statistics include contributions of a model's ability to absorb random error as well as its ability to approximate the underlying process or signal (Myung & Pitt, 2004). In sum, GoF is decomposed as follows:

$$\textit{Goodness-of-fit} = \textit{Fit to regularity (generalizability)} + \textit{Fit to noise (overfitting)} \tag{1}$$

This implies that model evaluation and selection based on only GoF statistics would be justifiable if the data were free of noise, which is highly unlikely. As such, GoF serves a necessary but not sufficient condition for model evaluation and selection (Preacher, 2006). Generalizability and the issue of overfitting should also be considered.

As seen in equation (1), generalizability is a model's ability to fit the underlying data-generating process. It is also a measure of a model's predictive accuracy on future and unseen replication samples arising from the same data-generating process. As model complexity increases, so does generalizability at first. However, after a tipping point, the model begins to fit noise in addition to regularity so that generalizability decreases with higher complexity. This is the concept of overfitting in Section 2.1. The balance point is the same as the optimum in Figure 2. 1. From this, it is also possible to see that bias is related to GoF and variance to generalizability. In short, model selection requires a balance between GoF and generalizability to help ensure that the chosen model captures regularity without overfitting (Myung & Pitt, 2004; Preacher, 2006). Accordingly, the selected model should not only explain the data well enough but also obey Occam's razor or principle of parsimony so that the model as minimally complex as possible. The optimum trade-off is realized when GoF statistics are adjusted for the contribution of model complexity (Myung, 2000; Myung et al., 2005).

Many factors can influence complexity. A facet of complexity that has been considered far more than others is the number of parameters. Models with a greater number of free parameters tend to be more complex. Usually, this is what comes to mind when thinking about model complexity in psychological and educational research (Bonifay & Cai, 2017). For example, Akaike information criteria (AIC) and the Bayesian information criteria (BIC) are two well-known GoF indices for relative model fit that penalize more complex models in terms of the number of parameters. However, research

shows that this is only part of the picture on complexity. Another dimension of complexity increasingly being incorporated is a model's functional form (Myung et al., 2005). Defined as the way in which the parameters are combined in the equations, it shows that models with the same number of parameters can have different model fit due to their functional form (Myung, 2000). The number of free parameters falls into parametrization complexity, and the functional form of a model into structural complexity (Markon & Krueger, 2004). A minimum of four other factors, for which research is scant, have been known to also contribute to model complexity: parameter range, sample size, the shape of the probability distribution in the likelihood function and estimation method, and experimental design (Pitt et al., 2002; Preacher, 2003).

2.3 Fitting Propensity (FP)

One promising alternative to traditional fit indices skewed toward GoF, so that model complexity is included is FP (Preacher, 2006). FP is defined as a model's inherent flexibility to fit diverse data patterns, all else being equal. The premise of FP is that some models will simply have the potential to fit a wider range of data patterns. Thus, FP can be described as the complement of parsimony as models. Thus, higher FP means a model is less parsimonious. Although FP can be examined for a single model, FP is especially beneficial as a relative fit index for comparing competing models in terms of how well each fit to representative data, as implied by the comparative adjectives. Figure 2. 2 shows similar information to Figure 2. 1 but with a focus on the relationships among FP, GoF, parsimony, generalizability, and overfitting. Models with higher FP, meaning lower parsimony (i.e., Figure 2. 2-(c)), tend to exhibit better GoF unadjusted for FP relative to models with lower FP (i.e., Figure 2. 2- (a)). As FP increases, generalizability reaches a

maximum and then decreases, with overfitting occurring beyond the point of maximum generalizability (Myung & Pitt, 2004; Preacher, 2006).

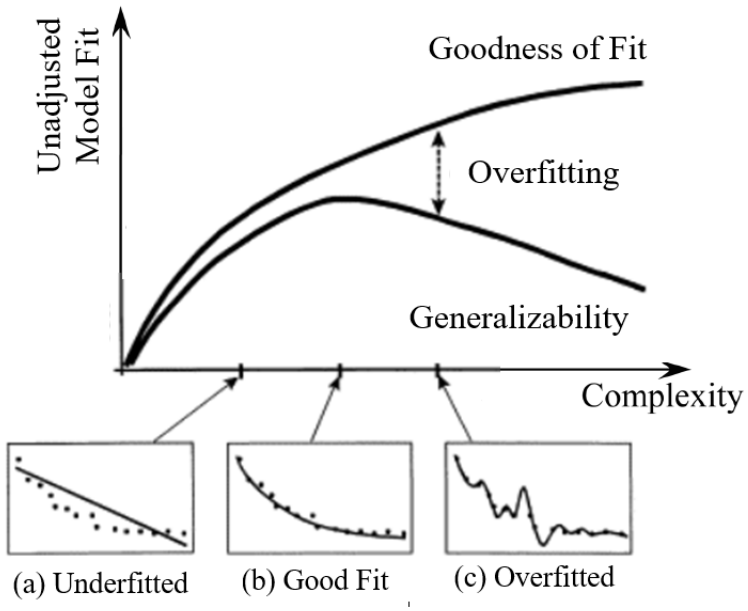


Figure 2. 2: Goodness-of-Fit, Generalizability, and Model Complexity

Note. Adapted from "Model comparison methods", by J. I. Myung and M. A. Pitt, 2004, *Methods in enzymology,* 383, p 355.

FP is based on the MDL principle introduced to the information-theoretic literature for model evaluation (Bonifay & Cai, 2017; Myung & Pitt, 2004; Preacher, 2006), which offers a coherent, intuitive view of model selection. The MDL principle arises from viewing data as well as statistical models as codes or algorithms for compressing data into a sequence of bits (Grünwald, 2000; Myung, 2000; Stine, 2004). A model compresses the regularities in the data by extracting redundancy from it. The more data a model can compress, the more we can learn from and predict using that model. The resulting description length from data compression consists of two parts:

$$DL = L(\text{data}|\text{model}) + L(\text{model}) \quad (2)$$

The description length is the length of the code for the data given the model (i.e., *L(*data|model*)*) plus the length of a description of the model itself (i.e., *L(*model*)*). *L(*data|model*)* represents the GoF of the model to the data while *L(*model*)* quantifies the model complexity (Myung, 2000). Therefore, the MDL principle straightforwardly addresses the trade-off between GoF and model complexity (Bonifay & Cai, 2017; Grünwald et al., 2005). The MDL principle selects the model with the smallest description length, balancing fit versus complexity. In this way, the MDL principle encapsulates Occam's Razor (Grünwald, 2000; Stine, 2004).

In MDL, this shortest total description length of equation (2) is considered over all possible data sequences. Thus, in an information-theoretic approach based on the MDL principle, a model is appraised relative to the complete data space in terms of its description. This reveals the model's innate tendency to fit well with any possible data from the data space. Model complexity based on the MDL includes the multiple aspects of complexity, such as the number of parameters, fictional form, parameter range, sample size, and estimation method. By restricting one or the other, it becomes possible to see the contributions of others (Grünwald et al., 2005; Myung, 2000; Preacher, 2003).

While analytic formulations for variants of MDLs exist, it is intractable for structural equation modeling (SEM) as well as IRT (Preacher, 2006; Bonifay & Cai, 2017). An alternative way to examine the MDL principle or FP is to follow the procedure outlined in Table 2. 1 (Falk & Muthukrishna, 2020). For investigating the FP of dichotomous IRT models, which is of interest in this study, Bonifay and Cai (2017) selected five IRT models that differed in functional form: exploratory factor analytic (EFA) model, bifactor model, two diagnostic classification models (DCMs) of the deterministic input noisy and-gate (DINA) model and the deterministic input noisy or-gate (DINO) model, and the unidimensional 3-parameter logistic (3PL) model. The four models excluding the

unidimensional 3PL model were specified to be equal in terms of the number of estimated parameters. The unidimensional 3PL model served as the model with more parameters but a less complex functional form. For dichotomous items, the data space of interest is the $(2^J\text{-}1)$-dimensional probability simplex where J is the number of items. Random datasets were generated from this space. All five models were fit the simulated datasets and then compared using unadjusted fit statistics designed for categorical data analysis such as the $Y2/N$ (Bartholomew & Leung, 2002; Cai et al., 2006) and LD X^2 (Chen & Thissen, 1997). The cumulative results of these statistics across replications were used to determine each model's inherent propensity to fit any possible data.

The results demonstrated the lowest FP for the unidimensional 3PL model despite an additional parameter. Furthermore, the EFA and bifactor models, which were models capable of handling more complex structures, showed higher FP. Such results suggested the importance of considering functional form or structural complexity, perhaps more so than the number of parameters or parametric complexity. They recommended that researchers de-emphasize good fit for models that are not parsimonious in form regardless of the number of parameters because it is impossible to discern whether regularities or noise is driving the seemingly good fit.

Table 2. 1: Procedure for Testing Fitting Propensity (FP)

1. Define of the model(s) of interest
2. Generate *n* random datasets representing the data space of interest
3. Fit model(s) of interest to the *n* random datasets
4. Record information regarding model fit for each model and dataset
5. Summize model fit using text, graphical displays, and measures of effect size

CHAPTER III

Contingency Tables and Data Generation

The first objective of this study is to develop a novel data generation method that can produce synthetic data randomly and uniformly sampled from the complete categorical data space defined by the lower-order margins. The proposed algorithm exploits the fact that item response data are contingency tables that are collapsible to consecutive marginal moments and takes advantage of long-standing, classical literature on contingency tables with fixed margins. This chapter explains how a contingency table of item response patterns over all individuals can be collapsed into a series of two-way contingency tables (i.e., second-order margins) with fixed row and column margins (i.e., first-order margins). Then, focusing on a general two-way contingency table with fixed margins, of which all two-way contingency tables of item responses fall into, the chapter illustrates what it means to sample randomly and uniformly from a two-way table when univariate margins are fixed, culminating in how it can be built upon to handle all the univariate margins as well as large numbers of two-way contingency tables of item responses simultaneously.

For explanation purposes, let us suppose that J items are measured for N individuals i. Let $\boldsymbol{y}' = (y_1, y_2, \cdots, y_J)$ be the vector of j variables where each j variable has K_j response alternatives that are ordered, $j = 1, \cdots J$. Responses to the items belong to a J -way contingency table with a total of $R = \prod_{j=1}^{J} K_j$ cells that denote the possible response vectors $\boldsymbol{y}'_r = (c_1, c_2, \cdots, c_J)$ where $r = 1, \cdots R$ and $c_j = 1, \cdots K_j$.

3.1 Two Representations of Contingency Tables

Item response probabilities can be realized as contingency tables. These contingency tables have two equivalent representations that extend to tables of any dimension: one based on cell probabilities and the other using moments (Maydeu-Olivares & Joe, 2014). The former is based on the multinomial framework, while the latter has roots in the MVB framework (Cai et al., 2006).

The characterization of IRT models based on the marginal moments of the MVB distribution (Teugels, 1990) is useful when describing LI methods. Let us further simplify things by considering only dichotomously scored items with 0 for incorrect responses and 1 indicating correct responses. Then $R = \prod_{j=1}^{J} K_j$ is equal to 2^J with each cell representing one of the 2^J item response patterns $\boldsymbol{\pi}$. Each of these item response patterns R can be considered as a random J-vector $\boldsymbol{y}' = (y_1, \dots, y_J)$ of Bernoulli random variables for which $\boldsymbol{y} = (y_1, \dots, y_J)', y_j \in \{0,1\}$ is a realization. It is important to note that small letters indicate both the variables and the values that these variables take in this study. The joint distribution of the MVB random vector $(y_1, \dots, y_J)$ is then

$$\pi_{\boldsymbol{y}} = P(y_1 = y_1, y_2 = y_2, \dots y_J = y_J). \tag{3}$$

In the characterization based on marginal moments, the $(2^J - 1)$-vector $\dot{\boldsymbol{\pi}}$ of joint moments of the MVB distribution can be written in the partitioned form $\dot{\boldsymbol{\pi}} = (\dot{\boldsymbol{\pi}}'_{\boldsymbol{1}}, \dot{\boldsymbol{\pi}}'_{\boldsymbol{2}}, \dots \dot{\boldsymbol{\pi}}'_{\boldsymbol{k}}, \dots, \dot{\boldsymbol{\pi}}'_{\boldsymbol{J}})'$, where the dimension of the vector $\dot{\boldsymbol{\pi}}_{\boldsymbol{k}}$ is $\binom{J}{k}$. Accordingly, $\dot{\boldsymbol{\pi}}_{\boldsymbol{1}}$ indicates the set of all J univariate or first-order marginal moments, where $\dot{\pi}_j = E(y_j) = P(y_j = 1) = \pi_j$. $\dot{\boldsymbol{\pi}}_{\boldsymbol{2}}$ denotes the set of $\frac{J(J-1)}{2}$ bivariate or second-order marginal moments, $\dot{\pi}_{jj'} = E(y_j y_{j'}) = P(y_j = 1, y_{j'} = 1) = \pi_{jj'}$ for all distinct j and j' satisfying $1 \le j \le j' \le J$.

The joint moments are defined in the manner up to the last one, $\dot{\boldsymbol{\pi}}_{\boldsymbol{J}} = E(y_j \cdots y_J) = P(y_j = \cdots = y_J = 1)$ with a dimension of $\binom{J}{J} = 1$ (Cai et al., 2006).

Take for example the smallest multivariate categorical data problem, which is a 2 × 2 table constructed using two dichotomously scored items (Table 3. 1). The cell representation uses four cell probabilities which must sum to one as is necessary for both the multinomial and MVB distributions. The joint-moments representation uses three moments consisting of the two means, $\pi_1^{(1)} = P(y_1 = 1)$ and $\pi_2^{(1)} = P(y_2 = 1)$ and the cross product $\pi_{12}^{(1)(1)} = P(y_1 = 1, y_2 = 1)$ (Table 3. 1) to convey the same information. Like this, there is a one-to-one relationship between the two representations, which is invertible regardless of the number of categorical variables involved (Cai et al., 2006; Maydeu-Olivares & Joe, 2014).

Table 3. 1: Two Representations for a 2 × 2 Contingency Table

Cells Representation		
	$y_2 = 0$	$y_2 = 1$
$y_1 = 0$	π_{00}	π_{01}
$y_1 = 1$	π_{10}	π_{11}

Margins Representation			
	$y_2 = 0$	$y_2 = 1$	
$y_1 = 0$			
$y_1 = 1$		$\pi_{12}^{(1)(1)}$	$\pi_1^{(1)}$
		$\pi_2^{(1)}$	

Note. Adapted from "Assessing approximate fit in categorical data analysis", by A. Maydeu-Olivares and H. Joe, 2014, *Multivariate Behavioral Research,* 49(4), p 307.

The example gives a sense of how the response patterns over individuals can be collapsed into consecutive lower-order margins. This study stipulates that it might suffice to generate random datasets for FP based on simply the lower-order moments where most of the information tends to lie and to disregard the higher-order moments.

The sampling of the first- and second-order margins involves probabilities up to pairs of items, as opposed to the full multinomial probabilities, which reduces to having to consider only J univariate and $J(J-1)/2$ bivariate margins rather than 2^J response patterns for dichotomous items.

The logic above easily generalizes to non-dichotomous cases as well. Returning to the ordinal step up with $R = \prod_{i=1}^{J} K_j$ item response patterns, the number of univariate and bivariate margins do not change. That is, they are still equal to J and $J(J-1)/2$, respectively. However, the number of first- and second-order marginal probabilities involved do change. For any item j, there are K_j cells in the corresponding first-order marginal table and for each unique item pair j and j', there are $K_j K_{j'}$ cells in the respective second-order marginal table. Due to the constraint that all cell probabilities of a contingency table should sum to one, there are only $K_j - 1$ linearly independent univariate probabilities per item and similarly, only $(K_j - 1)(K_{j'} - 1)$ linearly independent bivariate probabilities per unique item pair (Cai & Hensen, 2013). This explains why a 2 $\times$ 2 table can be characterized with two univariate margins and one bivariate margin as in Table 3. 1.

Applying this fact to the problem at hand of wanting to randomly sample from the categorical data space defined by the univariate and bivariate margins, this translates to having to generate or sample only $\sum_1^J (K_j - 1) + \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} (K_j - 1)(K_{j'} - 1)$ cell probabilities (i.e., $K_j - 1$ univariate probabilities per item and $(K_j - 1)(K_{j'} - 1)$ bivariate probabilities per unique item pair) in order to get the total number of data elements involved of $\sum_1^J K_j + \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} K_j K_{j'}$. This is also given in Section 1.1.

3.2 Geometry of Two-Way Contingency Tables with Fixed Margins

The issue of categorical data generation based on the lower-order margins is in essence not much different from the random sampling of two-way contingency tables with fixed margins. Geometric interpretations of contingency tables, with emphasis on tables with fixed margins, lie at the core of understanding the many random sampling methods used for such kinds of tables (Diaconis & Efron, 1985; Fienberg, 1970; Fienberg & Gilbert, 1970; Nguyen & Sampson, 1985; Slavković & Fienberg, 2009). This section provides an overview of related core concepts that are applicable to a generic two-way table with any number of rows or columns. Explanations focus on 2×2 tables, as explicit graphical representations are possible (Nguyen & Sampson, 1985).

The generic table can be thought of as one subset of the univariate and bivariate margins made by any item pair. For any pair of items of binary responses, the joint probability mass function (PMF) for items y_j and $y_{j'}$ can be realized as a 2×2 table of cell probabilities p_{ij} where $i = \{0,1\}$ and $j = \{0,1\}$ are from a bivariate Bernoulli distribution. Geometrically, one can identify the set $\mathcal{P}$ of all 2×2 PMF matrices $P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix}$ in a 3-dimensional probability simplex (Δ_3), which is a regular tetrahedron like in Figure 3. 1, with vertices $A_1 = (1,0,0,0), A_2 = (0,1,0,0), A_3 = (0,0,1,0), A_4 = (0,0,0,1)$ when using barycentric coordinates (Slavković & Fienberg, 2009). There is then a one-to-one correspondence between points A of the simplex with coordinates $A = (p_{00}, p_{01}, p_{10}, p_{11})$ and the 2×2 PMF matrices P.

Let $\mathcal{P}(\boldsymbol{R}, \boldsymbol{C})$ be the set of all 2×2 PMF matrices with fixed row marginal probability vector $\boldsymbol{R} = (t, 1 - t)$ and column marginal probability vector $\boldsymbol{C} = (s, 1 - s)$. By fixing one of the cell probabilities, for example, p_{00}, a PMF matrix P of $\mathcal{P}(\boldsymbol{R}, \boldsymbol{C})$ is completely defined

as $P = \begin{bmatrix} p_{00} & t - p_{00} \\ s - p_{00} & 1 - t - s - p_{00} \end{bmatrix}$, which is represented by point $A = (p_{00}, t - p_{00}, s-p_{00}, 1 - t - s - p_{00})$ in subspaces of the simplex Δ_3, and vice versa. Consider two planes $(p_{00}+p_{01}) = t$ and $(p_{00}+p_{10}) = s$ that intersect Δ_3 such that $s_1 = (s, 0,0,1 - s), s_2 = (s, 1 - s, 0,0), s_3 = (0, s, 1 - s, 0), s_4 = (0,0, s, 1 - s)$. t is defined similarly, as illustrated in Figure 3. 1. Each plane geometrically describes the set of points defined by a single fixed marginal. The set $\mathcal{P}(\boldsymbol{R}, \boldsymbol{C})$ is then the line segment given by the intersection of these planes. The two end or extreme points of the line segment are the upper Fréchet bound A^+ and lower Fréchet bound A^-. The independence model for a 2×2 table is also a matrix of $\mathcal{P}(\boldsymbol{R}, \boldsymbol{C})$ denoted by $P_I = \begin{bmatrix} ts & t(1-s) \\ s(1-t) & (t-1)(s-1) \end{bmatrix}$. This is equivalent to the point $A_I = (ts, t(1-s), s(1-t), (t-1)(s-1))$ shown in Figure 3. 1 (Fienberg & Gilbert, 1970; Nguyen & Sampson, 1985).

As t and s take on different possible values between 0 and 1, the set $\mathcal{P}(\boldsymbol{R}, \boldsymbol{C})$ varies accordingly along with points such as A_I, A^+, and A^-. This allows us to move from simply sampling from the line segment produced by the upper Fréchet bound $A^+$and lower Fréchet bound A^- or a point A_I given a particular set of t and s and find the data points that result in various sets of 2×2 PMF matrices conforming to certain models and/or set constraints. Thus, we can explore all parts of the tetrahedron. Simply varying t and s and not applying additional constraints allows us to pick data points from any part of the Δ_3. If additional constraints are added, such as that of the independence model, all relevant points A_I generate a surface of independence, which is a hyperbolic paraboloid (Figure 3. 2). This surface divides the simplex into two subsets where the subset to the left is the set of positively quadrant dependent matrices, and the subset to the right is that of negatively quadrant dependent matrices. Related to this, if association is defined by the coefficient $\alpha = \frac{p_{00}p_{11}}{p_{01}p_{10}}$, $0 \leq \alpha \leq \infty$ (Fienberg & Gilbert, 1970), then to

satisfy $p_{00}p_{11} = p_{01}p_{10}$, the probability distributions of the independence model have $\alpha = 1$. Just as the surface of independence exists for $\alpha = 1$, surfaces of constant degree of association α can also be determined and will always form a hyperboloid of one sheet. The loci of points corresponding to different classes of 2×2 tables constitute discrepant subspaces of Δ_3 depending on the constraints.

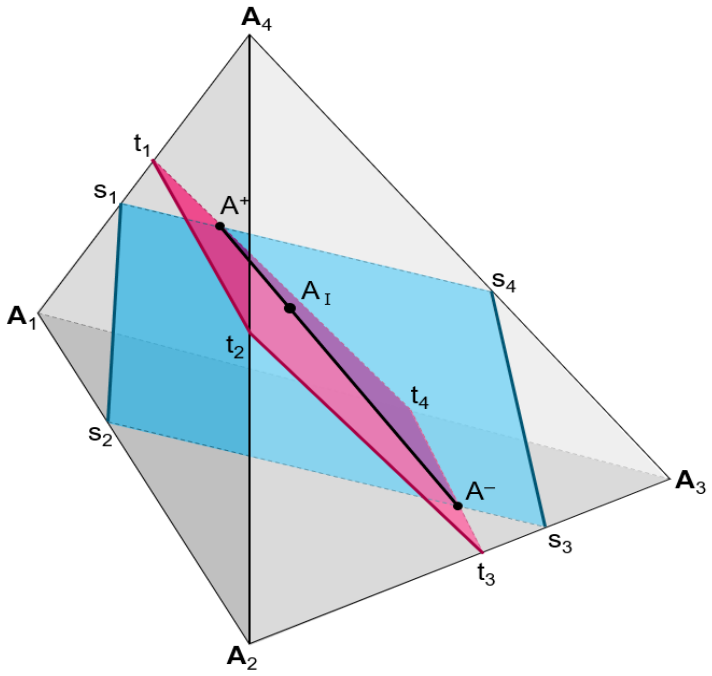


Figure 3. 1: Tetrahedron depicting a 2×2 Contingency Table with Fixed Margins

Note. Adapted from "The Geometry of Certain Fixed Marginal Probability Distributions", by T. T. Nguyen and A. R. Sampson, 1985, *Linear algebra and its applications*, *70*, p 77.

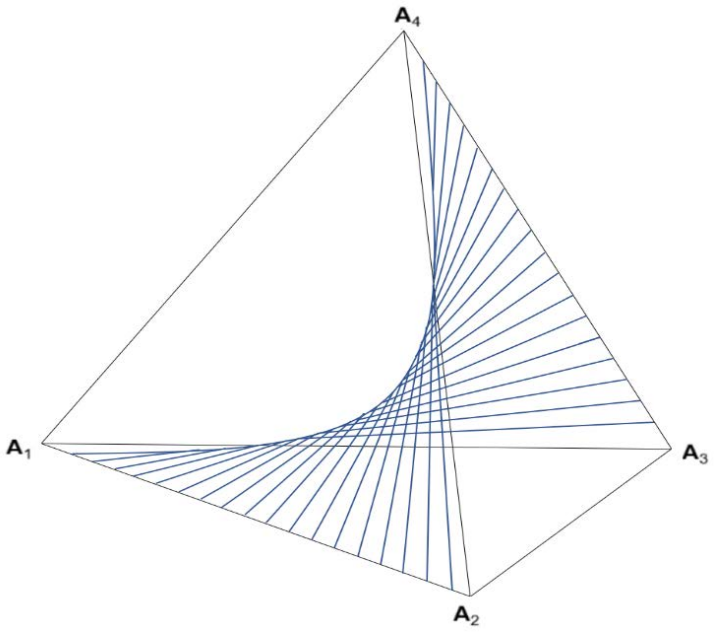


Figure 3. 2: Surface of Independence

Note. Adapted from "The geometry of a two by two contingency table" by S.E. Fienberg and J.P. Gilbert, 1970, *Journal of the American Statistical Association*, *65*(330), p 697.

Although direct visualization is not possible as dimensions get higher, the concepts above generalize to $m \times n$ contingency tables with a fixed row marginal probability vector $\mathbf{R} = (r_1, r_2, \ldots, r_m)$ as well as a fixed column marginal probability vector $\mathbf{C} = (c_1, c_2, \ldots, c_n)$ (Fienberg, 1968; Nguyen & Sampson, 1985; Slavković & Fienberg, 2009). To summarize key points with reference to the scenario listed at the beginning of the chapter, geometrically, the set $\mathcal{P}(\boldsymbol{R}, \boldsymbol{C})$ of all $m \times n$ PMF matrices P consisting of cell probabilities for an item pair reside in a $(mn-1)$-dimensional simplex $(\Delta_{(mn-1)})$. m in the study's context refers to the response frequencies (K_j) of an item j while n refers to those $(K_{j'})$ for an item j'. Every matrix P can be geometrically represented by a point A in the $(mn-1)$ -dimensional simplex with coordinates $A = (p_{00}, p_{01}, p_{10}, p_{11}, \cdots, p_{(m-1)(n-1)})$. The dimension of the simplex is $(mn-1)$ because the probability simplex is constrained by $\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} p_{ij} = 1$ so that we lose one degree of freedom.

Let $\mathcal{P}(\boldsymbol{R},\boldsymbol{C})$ denote the set of all $m \times n$ PMF matrices conforming to the specific marginal constraints of a set of probability vectors $\mathbf{R}$ and $\mathbf{C}$. $\mathcal{P}(\boldsymbol{R},\boldsymbol{C})$ can be found as a subset of the Δ_{mn-1} satisfying a set of conditions laid out by the Fréchet bounds for each cell probability p_{ij} where $i = \{0,(m-1)\}$ and$j = \{0,(n-1)\}$. The bounds are

$$max\big(0, r_i + c_j - 1\big) \leq\ p_{ij} \leq min\,(r_i, c_j). \tag{4}$$

This results in hyperplanes that are bounded by the extreme matrixes P created by the Fréchet bounds and thus, define the subspace of the $(mn-1)$-dimensional probability simplex where valid data points made be found. When $(m-1)(n-1)$ cell probabilities fall within the Fréchet bounds, determined by the given marginal constraints, a unique PMF matrix P of $\mathcal{P}(\boldsymbol{R},\boldsymbol{C})$ can be completely defined. As $\boldsymbol{R}$ and $\boldsymbol{C}$ change, we can expect the $\mathcal{P}(\boldsymbol{R},\boldsymbol{C})$ to vary over the entirety of the simplex in question. If other constraints are added, valid points reside in even more constrained subspaces of the Δ_{mn-1}. The manifold of independence is one example if we consider only points pertaining to the independence model, which is a generalization of the surface of independence for 2×2 tables (Figure 3. 2) to $m \times n$ tables.

3.3 Sequential importance Sampling (SIS) of Contingency Tables with Fixed Margins

While the geometric representation of contingency tables with fixed margins in Section 3.2 provides the theory supporting an LI-based data-generating process, the issue remains how one can randomly sample from the data space with constraints outlined by the theory. Among many possible methods, this study utilizes the sequential importance sampling (SIS) approach. The SIS approach is gaining favor due to its efficiency in sampling multi-way tables of many rows and columns with given marginal

constraints (Chen, Diaconis, et al., 2005). Furthermore, the SIS procedure samples contingency tables independently and uniformly, which fits well with the need of this study to sample many tables simultaneously.

As the name suggests, SIS randomly samples from a target contingency table in a sequential manner, with one cell probability being populated at a time. As the probability of each cell is a random variable, the resulting contingency table is also a random variable. Suppose $\Sigma_{\boldsymbol{rc}}$ denotes the set of all $m \times n$ contingency tables with row marginal probability vector $\boldsymbol{R} = (r_1, r_2, \ldots, r_m)$ and column marginal probability vector $\boldsymbol{C} = (c_1, c_2, \ldots, c_n)$. Let p_{ij} be the element at the ith row and the jth column of a contingency table. Following the logic of SIS, cells are sampled one-by-one, from column to column, beginning with cell p_{11}.

Recall the necessary and sufficient condition for the existence of a contingency table of probabilities with $\boldsymbol{R}$ and $\boldsymbol{C}$ is

$$r_1 + r_2 + \cdots r_m = c_1 + c_2 + \cdots + c_n \equiv 1 \tag{5}$$

Thus, p_{11} needs to satisfy the following conditions:

$$\begin{gathered} 0 \leq p_{11} \leq r_1, \\ c_1 - \sum_{i=2}^{m} r_i = c_1 + (r_1 - 1) \leq p_{11} \leq c_1 \end{gathered} \tag{6}$$

which can be combined as

$$max(0, c_1 + r_1 - 1) \leq p_{11} \leq min\,(r_1, c_1) \tag{7}$$

Notice how this matches the Fréchet bounds for any cell probability p_{ij} defined in equation (4) (Chen, Dinwoodie, et al., 2005; Fienberg, 1999). Fréchet bounds determine the lower and upper limits of a bivariate probability based on the surrounding univariate margins and p_{11} is randomly sampled from the uniform distribution between the lower and upper Fréchet bounds (other distributions can be used to sample cells as well).

After sampling and thus fixing p_{11}, a second cell probability p_{21} is sampled in the same manner but conditional on p_{11}. This equates to the Fréchet bounds being updated to incorporate the information from the previously sampled cell probability p_{11}. Based on the same logic, we can recursively sample the other cells in column 1 that are $1 \leq i \leq m-1$ by uniformly sampling from the range set by the repeatedly updated restriction

$$max\left(0, c_1 - \sum_{k=1}^{i-1} p_{k1} - \sum_{k=i+1}^{m} r_k\right) \leq p_{i1} \leq min\left(r_i, c_1 - \sum_{k=1}^{i-1} p_{k1}\right). \quad (8)$$

This results in all the cells that are free in column 1 being sampled. Then we proceed with a similar process of uniformly sampling the remaining free cell probabilities in the remaining columns ($1 \leq i \leq m-1$ and $1 < j \leq n-1$) within the bounds of

$$max\left(0, c_j - \sum_{k=1}^{i-1} p_{kj} - \sum_{k=i+1}^{m} r_k + \sum_{k=i+1}^{m} \sum_{k'=1}^{j-1} p_{kk'}\right) \leq p_{ij} \leq min\left(r_i - \sum_{k=1}^{j-1} p_{ik}, c_j - \sum_{k=1}^{i-1} p_{kj}\right). \quad (9)$$

The number of free cells in a two-way contingency table with marginal constraints is equal to $(m-1)(n-1)$. All other cell probabilities follow naturally. This is what is called the degrees of freedom. In short, the entire sampled contingency table is the result of sequentially fixing the free cell probabilities in the table (Fienberg, 1999). For example, for a 2×2 table of given row and column sums (i.e., $r_1 + r_2 = c_1 + c_2 \equiv 1$), the degrees of freedom is equal to 1. A cell probability (e.g., p_{11}) is the only variable that needs to be sampled from a uniform or hypergeometric distribution within the range of $[min(r_1, c_1), max(0, c_1 + r_1 - 1)]$. Then all the other cells can be filled as $p_{12} = r_1 - a_{11}, p_{21} = c_1 - p_{11}$ and $p_{22} = 1 - p_{12} - p_{21} - p_{11}$.

3.4 Proposed Data Generation Algorithm

For IRT models specified using the first- and second-order marginal moments, the complete data space of possible data patterns consists of all the bivariate margins simultaneously satisfying the bounds set by all the univariate margins (i.e., Fréchet bounds) for a certain number of items, J in this case. We can propose an algorithm that should be able to randomly and uniformly sample data points from the target data space by answering the following questions:

1) how to appropriately set the distribution to draw the univariate probabilities of an item from,
2) how to randomly sample a bivariate probability arising from an item pair under the pre-generated univariate margin constraints for each item
3) how to do 1) and 2) when the lower order margins of all possible unique item pairs must be considered together. The contingency tables for all item pairs are not entirely independent as they can share some univariate margins with other contingency tables depending on the item pair in question.

For convenience, let's constrain the previous scenario so that J items each have the same m categories with each unique item pair y_j and $y_{j'}$ making up a $m \times m$ contingency table where cells refer to bivariate probabilities. Let p_{ij} be the bivariate probability corresponding to the cell of the ith row and the jth column of a $m \times m$ contingency table. Each item of an item pair also has its respective marginal univariate probabilities that is set as a vector of row probabilities ($\mathbf{R} = (r_1, r_2, \ldots, r_m)$) for y_j and as a vector of column probabilities ($\mathbf{C} = (c_1, c_2, \ldots, c_m)$) for y_j. Again, $r_1 + r_2 + \cdots r_m = c_1 + c_2 + \cdots + c_m = 1$.

Starting with the univariate probabilities, the suitable univariate distribution can be found when considering the relationship between the simplex and the Dirichlet distribution. The probability density function (PDF) of the Dirichlet distribution for k random variables is a standard $(k-1)$ - dimensional probability simplex existing in a k dimensional space (Lin, 2016). This Dirichlet distribution, often denoted as $Dir(\boldsymbol{\alpha})$, has $\alpha_1, \cdots \alpha_k$ concentration parameters where $\alpha_k > 0$. The marginal distributions of the Dirichlet distribution can be derived directly by applying the aggregation property. Suppose we have a 2×2 table of $(p_{00}, p_{01}, p_{10}, p_{11}) \sim Dir(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ where $p_{11} = 1 - p_{00} - p_{01} - p_{10}$. The resulting distribution of the sums of random variables $p_{00} + p_{01}$ and $p_{10} + p_{11}$ are $(p_{00} + p_{01}, p_{10} + p_{11}) \sim Dir(\alpha_1 + \alpha_2, \alpha_3 + \alpha_4)$, which is the same as $Beta(\alpha_1 + \alpha_2, \alpha_3 + \alpha_4)$, and gives the univariate probabilities of the row variable. The column variable counterparts can be found as $(p_{00} + p_{10}, p_{01} + p_{11}) \sim Dir(\alpha_1 + \alpha_3, \alpha_2 + \alpha_4)$. Considering that all α_ks for the $Dir(\boldsymbol{\alpha})$ are set equal to one when uniformly sampling of a unit simplex, which the probability simplex is a part of, the m univariate (category) probabilities for each item y_j should be sampled from a $Dir(\alpha_1 = m, \dots, \alpha_m = m)$ summing relevant α_ks of the joint distribution $Dir(\alpha_1 = 1, \dots, \alpha_{mn} = 1)$ of the $(mm - 1)$- dimensional simplex. These results are corroborated explicitly by research on the tetrahedron, the Dirichlet distribution, and sampling from contingency tables (Draconis & Efron, 1987; Letac & Scarsini, 1998).

The answer for 2) can be found by combining knowledge about Fréchet bounds that dictate the lower and upper bounds of a bivariate probability based on the surrounding univariate margins sampled from the Dirichlet distribution and adapting the SIS proposed by Chen, Diaconis, et al. (2005) accordingly. The previous sections provide relevant details.

Finally, having multiple sets of lower-order margins with overlapping constraints is not an issue when using the SIS method as it samples from each contingency table independently. As long as the respective univariate margins of each specific item pair are correctly specified, this reduces having to only consider one item pair or one contingency table at a time. We just simply repeat the process for all the contingency tables for all the unique item pairs. Weaving the pieces together, the proposed data generation mechanism follows the steps outlined in Table 3. 2.

Table 3. 2: Data Generation Algorithm

1. Randomly sample m univariate (category) probabilities for each item y_j from $Dirichlet(a_1 = m, \dots, a_m = m)$

2. Uniformly sample a bivariate probability cell $p_{ij}, i = 1, \dots, m, j = 1, \dots, m$ from the following Frechet bounds:

If $1 \leq i \leq m-1$ and $j = 1$

$$max(0, c_1 + r_1 - 1) \leq a_{11} \leq \min(r_1, c_1)$$

$$max\left(0, c_1 - \sum_{k=1}^{i-1} a_{k1} - \sum_{k=i+1}^{m} r_k\right) \leq a_{i1} \leq min\left(r_i, c_1 - \sum_{k=1}^{i-1} a_{k1}\right)$$

If $1 \leq i \leq m-1$ and $1 < j \leq m-1$

$$max\left(0, c_j - \sum_{k=1}^{i-1} a_{kj} - \sum_{k=i+1}^{m} r_k + \sum_{k=i+1}^{m} \sum_{k'=1}^{j-1} a_{kk'}\right) \leq a_{ij} \leq min\left(r_i - \sum_{k=1}^{j-1} a_{ik}, c_j - \sum_{k=1}^{i-1} a_{kj}\right)$$

If $i = m$ and $1 \leq j \leq m-1$

$$a_{ij} = c_j - \sum_{k=1}^{i-1} a_{kj}$$

If $1 \leq i \leq m-1$ and $j = m$

$$a_{ij} = r_i - \sum_{k=1}^{j-1} a_{ik}$$

4. Repeat steps 2 for all $J(J-1)/2$ item pairs y_j and $y_{j'}$.

5. Repeat steps 1-4 a very large number of times.

The proposed data generation algorithm can readily generate dichotomous and polytomous item data of large quantities. The algorithm was tested using 10,000 samples for up to 50 items and four categories, which required generating a total of $50 \times 4 + \frac{50 \times 49}{2} \times 4 \times 4 = 19800$ data elements. This number is within an easily manageable range for most computers. The same cannot be said if attempting to use the simplex sampling method, as it requires generating 4^{50} item response probabilities, which is greater than 10^{30}. Theoretically, the proposed algorithm should be able to sample uniformly from the $(mm-1)$-dimensional simplex, which is the desired categorical data space.

An illustration of what this should look like is given in Figure 5. 2 in Chapter 5 for a 2×2 table, resulting in a tetrahedron. It is to note that while random and uniform sampling across the complete simplex was the main goal, it is possible to limit the categorial data space to more plausible data patterns and sample from within it (Preacher, 2006). The criteria for plausible data can vary due to factors such as model parametrization, estimators, and inferences of interest. Geometric descriptions of classes, such as those of 2×2 tables as above can be used to identify the data space or subspaces from which one would need to sample, depending on the research aim. The proposed data generation method then can be modified to sample randomly and uniformly data ranging from any subspace to the entire space of the simplex Δ_3. The

subspace or full space will then serve as the reference for evaluating the inherent complexity of a given categorical data model.

Chapter IV

Limited Information Estimation Methods

The second aim of this study is to derive an LI-based estimator, more generally known as CML estimators, capable of fitting a variety of IRT models using only data from the lower-order margins. This chapter is a review of existing IRT model estimation techniques with a focus on LI methods. We will again suppose the scenario in Chapter 3 where J items are measured for N individuals $i = 1, \cdots N$ and $j = 1, \cdots J$. y_{ij} is the response from an individual i to item j that can have one of K_j categories that are assumed to be ordered. $\boldsymbol{y} = (y_{ij})_{N \times J}$ is the data matrix as $i = 1, \cdots N$ and $j = 1, \cdots J$. $\boldsymbol{y}' = (y_1, y_2, \cdots, y_J)$ denotes the vector of J variables where each variable j has K_j response alternatives. Responses to the items belong to a J-way contingency table with a total of $R = \prod_{j=1}^{J} K_j$ cells that denote the possible response vectors $\boldsymbol{y}'_{\boldsymbol{r}} = (c_1, c_2, \cdots, c_J)$ where $r = 1, \cdots R$ and $c_j = 1, \cdots K_j$. Let us also denote a latent trait vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_q)$ of 1 through q latent factors, using which we specify the probability of each r as

$$\Pr(y_1 = c_1, y_2 = c_2, \ldots, y_J = c_J | \boldsymbol{\xi}) = f(\boldsymbol{\xi}) \tag{10}$$

where $c_1, c_2, \ldots, c_J$ represent the response categories of $y_1, y_2, \ldots, y_J$, respectively (Cai & Moustaki, 2018).

4.1 Classification of Item Response Theory (IRT) Estimation Methods

IRT models were developed for the latent variable modeling of categorical observed variables and continuous underlying variables where traditional linear factor models are no longer suitable. IRT estimation methods fall into two main categories of the underlying variable (UV) approach and the IRT approach. In the former, IRT models

are often called item factor models from the term item factor analysis (IFA). The fundamental distinction between the two approaches lies in how they perceive the categorical outcome variables (Bolt, 2005; Cai & Moustaki, 2018; Jöreskog & Moustaki, 2001; Wirth & Edwards, 2007). The former approach considers ordinal variables as discrete manifestations of underlying continuous variables to meet the assumptions of classical FA for its implementation. The latter approach makes no such assumptions so that ordinal indicators are left as they are. As opposed to using latent variable methods for continuous outcomes, it is modeled by way of distributional assumptions for the observed items conditional on the latent variables that capitalize on the notion of conditional independence (Cai & Moustaki, 2018; Katsikatsou et al., 2012).

Alternatively, latent trait model estimation methods can also be categorized as either FI or LI methods (Bolt, 2005; Maydue-Olivares & Joe, 2005). As repeatedly mentioned, the whole contingency table of the full multinomial probabilities is the unit of analysis in FI estimation methods. LI methods instead make use of lower-order tables, mainly up to two-way tables, and perform the estimation on those tables (Cai & Moustaki, 2018; Maydeu-Olivares & Joe, 2006). To reiterate, the clearest distinction between FI and LI estimation is the difference in the number of data elements fitted, which exponentially increases as items and factors do. The research overall suggests that the performances of FI and LI methods for categorical outcomes are comparable (Bolt, 2005).

4.2 Limited Information (LI) Estimation Methods

LI is the terminology used in psychometrics referring to estimation and inference procedures based on low-dimensional margins. LI methods are much more prevalent in the UV approach as opposed to the IRT approach. In the UV approach, the observed ordinal variable vector $\boldsymbol{y}' = (y_1, \ldots, y_J)$ is connected to an underlying continuous variable

vector $\boldsymbol{y}^{*\prime} = (y_j^*, \ldots, y_j^*)$ usually assumed to be multivariate normal (MVN) with $K_j - 1$ thresholds for each y_j that discretize the MVN distribution. Elaborating, an observed ordinal variable y_j is connected to an underlying continuous variable y_j^* by

$$y_j = c_j \Leftrightarrow \tau_{c_j-1}^{(y_j)} < y_j^* < \tau_{c_j}^{(y_j)}, c_j = 1, \cdots K_j \tag{11}$$

where

$$-\infty = \tau_0^{(y_j)} < \tau_1^{(y_j)} < \cdots < \tau_{K_j-1}^{(y_j)} < \tau_{K_j}^{(y_j)} = +\infty. \tag{12}$$

The $\tau_{c_j}^{(y_j)}$ is the c_jth threshold of variable y_j and together these thresholds define the K_j categories. Because only ordinal information is available, the scale of the latent response variable is indeterminate and most often we set each y_j^* as a standard normal distribution. The observed ordinal variable vector $\boldsymbol{y}' = (y_j, \ldots, y_J)$ has now been transformed into a *J*-dimensional vector of latent continuous variables $\boldsymbol{y}^{*\prime} = (y_j^*, \ldots, y_J^*)$ to which usual FA methods can be applied.

However, it becomes evident that this is not feasible when considering the (log-) likelihood function to be maximized where the data is assumed to have been generated by a multinomial distribution. As y_{ij} is the response from an individual i to item j with data matrix $\boldsymbol{y}$, the likelihood function can be written as

$$L(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{N} \pi_i(\boldsymbol{\theta}) \tag{13}$$

where $\boldsymbol{\theta}$ is a parameter vector, and $\pi_i(\boldsymbol{\theta})$ is the probability under the model for the response vector from person i, $\Pr(y_{i1} = y_{i1}, \cdots y_{iJ})$. This is the likelihood contribution of a single observation i. The log-likelihood function, which is what is actually maximized is then

$$l(\boldsymbol{\theta};\boldsymbol{y}) = lnL(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^{N} ln\pi_i(\boldsymbol{\theta}) \tag{14}$$

Since every response pattern, denoted r, given the latent ability factors, have equal contributions to the log-likelihood, equation (14) can be rewritten as

$$l(\boldsymbol{\theta};\boldsymbol{y}) = ln\, L(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{r=1}^{R} n_r ln\pi_r(\boldsymbol{\theta}) = N\sum_{r=1}^{R} p_r ln\pi_r(\boldsymbol{\theta}) \tag{15}$$

where n_r and $\pi_r(\boldsymbol{\theta})$ are each the observed frequency and the probability under the model for the response pattern r with $\pi_r(\boldsymbol{\theta}) > 0$ and $\sum_{r=1}^{R}\pi_r(\boldsymbol{\theta}) = 1$. $N = \sum_{r=1}^{R} n_r$ and $p_r = n_r/N$ is the sample proportion of r. In short, n_r serve as frequency weights for each pattern instead of multiplying over all individuals one-by-one.

In the UV approach, the maximization of this log-likelihood function over the parameter vector $\boldsymbol{\theta}$ requires the evaluation of the J-dimensional integral with no closed form of

$$\pi_r(\boldsymbol{\theta}) = \pi(y_1 = c_1, y_2 = c_2, \ldots, y_J = c_J;\ \boldsymbol{\theta}) = \int_{\tau_{c_1-1}^{(y_1)}}^{\tau_{c_1}^{(y_1)}} \cdots \int_{\tau_{c_J-1}^{(y_J)}}^{\tau_{c_J}^{(y_J)}} \phi_J(\boldsymbol{y}^*|\boldsymbol{\Sigma}_{\mathbf{y}^*})\, d\boldsymbol{y}^* \tag{16}$$

where $\phi_J(\boldsymbol{y}^*|\boldsymbol{\Sigma}_{\mathbf{y}^*})$ is a J-dimensional normal density with zero mean and correlation matrix $\boldsymbol{\Sigma}_{\mathbf{y}^*}$. Due to this, LI methods are often used under the UV approach which falls into either the class of multiple-stage (Jöreskog, 1990; Muthén 1984) or CML (Lindsay, 1988) estimators.

At the heart of multiple-stage estimators are the correlations between each pair of variables $(y_j^*, y_{j'}^*)$, which are tetrachoric or polychoric correlations depending on the variables being binary or ordinal (Jöreskog, 1990; Muthen, 1984). They are called multiple-stage estimators because estimation is carried out in multiple stages. In the case of three-stage estimation methods, first-order statistics (i.e., $K_j - 1$ thresholds per

item) are estimated using the univariate margins of each item. Then, given such first-stage estimates, second-order statistics (i.e., $J(J-1)/2$ tetrachoric/polychoric correlations) are estimated by maximizing the bivariate marginal likelihoods for each observed data of item pairs $(y_j, y_{j'})$ separately, given the first-stage estimates. In the third and last stage, the structural model is estimated using conventional FA given the estimated correlation matrix from stage two based on $\boldsymbol{y}^{*\prime} = (y_{ij}^*, \ldots, y_{iJ}^*)$ (Cai & Moustaki, 2018). Three-stage methods differ in this step depending on the version of generalized least squares (GLS) used such as unweighted least squares (ULS), diagonally weighted least squares (DWLS), and weighted least squares (WLS). Their main advantage is that they are computationally less demanding than FIML. The limitations are that they require multiple stages of estimation and tend to have issues estimating the weight matrix, especially in small sample sizes with zero or small frequencies in the bivariate margins (Cai & Moustaki, 2012; Katsikatsou et al., 2012, Xi, 2011).

CML estimation is the other branch of LI methods in psychometrics. Among various CML estimation methods, the use of composite marginal maximum likelihoods composed by low-dimensional marginal distributions has received at lot of attention. The advantage of the CML approach when compared with the multi-stage estimation methods is that all the model parameters are estimated simultaneously so that there is less room for error. Moreover, the standard errors of the estimates are straightforward to calculate without the need for any weight matrix.

Let us again consider the log-likelihood in equation (14). CML methods developed under the UV approach include pseudo-likelihood functions where the sum of both the univariate and bivariate marginal distributions are maximized, coined the underlying bivariate normal (UBN) method, as well as functions where simply the bivariate marginal

distribution or pairwise marginal likelihoods are maximized (Jöreskog & Moustaki, 2001, 2018; Katsikatsou et al., 2011; Nuo & Browne, 2014). The latter is called the pairwise maximum likelihood (PML). The UBN fit function can be written as

$$\sum_{j=1}^{J} lnL(\boldsymbol{\theta}; y_j) + \sum_{j=2}^{J-1} \sum_{j'=j+1}^{J} lnL(\boldsymbol{\theta}; y_j, y_{j'}) = \sum_{j=1}^{J} \sum_{c_j=1}^{K_j} n_{c_j}^{(y_j)} ln\left[\pi_{c_j}^{(y_i)}(\boldsymbol{\theta})\right] + \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \sum_{c_j=1}^{K_j} \sum_{c_{j'}=1}^{K_{j'}} n_{c_j c_{j'}}^{(y_j y_{j'})} ln\left[\pi_{c_j c_{j'}}^{(y_j y_{j'})}(\boldsymbol{\theta})\right]. \quad (17)$$

The PML method involves only the latter part of equation (17) with the bivariate likelihoods. $n_{c_j}^{(y_j)}$ and $n_{c_j c_{j'}}^{(y_j y_{j'})}$ are the univariate and bivariate frequency of a response in category c_j and $c_{j'}$ for variables y_j and $y_{j'}$, respectively, and $\pi_{c_j}^{(y_i)}(\boldsymbol{\theta})$ and and $\pi_{c_j c_{j'}}^{(y_j y_{j'})}(\boldsymbol{\theta})$ are the corresponding univariate and bivariate marginal probabilities under the model. Elaborating, univariate marginal probabilities are

$$\pi_{c_j}^{(y_j)}(\boldsymbol{\theta}) = \Pr(y_j = c_j) = \int_{\tau_{c_j-1}^{(y_j)}}^{\tau_{c_j}^{(y_j)}} \phi_1(y_j^*) dy_j^* \quad (18)$$

where $\phi_1(\cdot)$ is the standard normal density function while bivariate marginal probabilities are

$$\pi_{c_j c_{j'}}^{(y_j y_{j'})}(\boldsymbol{\theta}) = \Pr(y_j = c_j, y_{j'} = c_{j'}) = \int_{\tau_{c_j-1}^{(y_j)}}^{\tau_{c_j}^{(y_j)}} \int_{\tau_{c_{j'}-1}^{(y_{j'})}}^{\tau_{c_{j'}}^{(y_{j'})}} \phi_2(y_j^*, y_{j'}^* | \rho_{jj'})\, dy_j^* dy_{j'}^* \quad (19)$$

where $\phi_2(\cdot,\cdot,\rho_{ij})$ is the standardized bivariate normal distribution density function with correlation $\rho_{jj'}$.

Like this, the above CML methods only require one- and two-dimensional integrations, instead of the general *J*-dimensional integration, one for each item j,

needed for a conventional UV approach as in equation (16). Thus, it is always computationally feasible (Katsikatsou et al., 2012) and a larger number of item variables as well as a larger number of factors are estimable (Joreskong & Moustaki, 2001; Xi, 2011). From this, it is possible to see why the UV approach frequently favors LI estimation. As dimensionality requiring integration is the same as the number of observed variables, the integration becomes computationally infeasible even for a handful of items. This is less so in an IRT context, where the dimensionality increases as a function of the number of latent factors (Cai, 2010).

4.3 Composite Maximum Likelihood (CML) Estimation

CML methods expand upon Fisherian likelihood theory driven by the need to reduce computational burden in likelihood estimation of high-dimensional data. For this purpose, they replace high-dimensional complicated likelihood functions with any product of conditional or marginal lower-dimensional densities that are more computationally feasible (Joe et al., 2012; Lindsay, 1988; Varin, 2008; Varin et al., 2011). Modeling of lower-order dimensional distributions is frequently easier and more straightforward as modeling uncertainty tends to increase with dimensionality. As such, possible model misspecification in higher-order dimensional distributions can be avoided so that CML is a robust modeling alternative. To add, a model of lower order distributions is likely to be compatible with multiple modeling options designed for higher dimensional distributions.

Composite likelihood methods achieve this by piecing together individual component likelihoods, each of which corresponds to a marginal or conditional event (Lindsay, 1988). Let Y be a j-dimensional vector with probability density function $f(y;\theta)$ for some unknown q-dimensional parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ Suppose $(A_1, \ldots, A_D)$ are a set of

marginal or conditional events with associated likelihoods $L_d = f(\boldsymbol{y} \in A_d;\ \boldsymbol{\theta}) = Pr(\boldsymbol{y} \in \boldsymbol{A}_d), d = 1, \ldots, D$. A composite likelihood is the weighted product of each of the individual likelihoods,

$$CL(\boldsymbol{y}; \boldsymbol{\theta}) = \prod_{d=1}^{D} f(\boldsymbol{y} \in A_d;\ \boldsymbol{\theta})^{w_d} \tag{20}$$

where $\{w_d\}$ are a set of a non-negative weights associated with event A_d. The log-likelihood is

$$cl(\boldsymbol{y}; \boldsymbol{\theta}) = lnCL(\boldsymbol{y}; \boldsymbol{\theta}) = \sum_{d=1}^{D} w_d lnf(\boldsymbol{y} \in A_d;\ \boldsymbol{\theta}) \tag{21}$$

If its maximizer $\widehat{\boldsymbol{\theta}}_{CL}$ if unique, it is the maximum composite likelihood estimator (MCLE; Xu & Reid, 2011; Varin, 2008). That is,

$$\widehat{\boldsymbol{\theta}}_{CL} = \text{argmax}_{\boldsymbol{\theta}} cl(\boldsymbol{y}; \boldsymbol{\theta}) \tag{22}$$

CML estimation can also to divided into two categories (Varin et al., 2011). One category consists of "subsetting methods," (Cox & Reid, 2004) where the joint likelihood is replaced with any product of conditional or marginal densities that is easier to evaluate, and hence to maximize (Varin, 2008; Varin et al., 2011). The other category is "omission methods," which remove elements of the likelihood that are likely to complicate the full likelihood but provide little information about model parameters (Varin, 2005). From this, it is evident that information will be lost when using CML. A CML estimator needs to provide significant computational savings while at the same time keeping the loss of efficiency tolerable.

As mentioned above, composite marginal likelihood methods for up to bivariate likelihoods are common in psychometrics (and research in general). Thus, CML methods

in psychometric fall under "subsetting methods." The simplest composite marginal likelihood is constructed under working independence assumptions resulting in

$$L_{ind}(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{j=1}^{J} f(y_j;\boldsymbol{\theta})^{w_j} \tag{23}$$

$f(y_i;\boldsymbol{\theta})$ refers to the probability of observing each individual variable y_i. Equation (23) is coined the independence marginal likelihood function because it only considers univariate marginal events (Xi, 2011; Xu & Reid, 2011). $L_{ind}(\boldsymbol{y};\boldsymbol{\theta})$ is equal to the true likelihood if independence among all variables holds. If this is violated, it is no longer the true likelihood because the dependency among variables has not been factored in. Inferences on marginal parameters, mainly thresholds, are possible with $L_{ind}(\boldsymbol{y};\boldsymbol{\theta})$. However, often parameters regarding the dependence between variables are also of interest as real data are usually correlated. Thus, composite likelihood modeling pairs of observations, such as the pairwise likelihood (PL) or those constituted of larger subsets like triplets are used (Cox & Reid, 2004; Varin, 2008). PLs can be written as

$$L_{pair}(\boldsymbol{y};\boldsymbol{\theta}) = \prod_{j=1}^{J-1} \prod_{j'=j+1}^{J} f(y_j; y_{j'};\boldsymbol{\theta})^{w_{j,j'}} \tag{24}$$

Like this, the PL $L_{pair}(\boldsymbol{y};\boldsymbol{\theta})$ calculates the probability of every possible variable pair. The most widespread form of composite marginal or conditional likelihood in CML applications is the pairwise likelihood (PL).

Despite being pseudo-likelihoods, CML estimators inherit many of the desirable properties of inference based on the full likelihood function. For one, they have the properties of being consistent, and asymptotically normally distributed (Lindsay, 1988; Varin, 2008; Varin et al., 2011) under regularity conditions. That is,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{CL} - \boldsymbol{\theta}) \rightarrow \boldsymbol{N_q}(0,\ G^{-1}(\boldsymbol{\theta})) \tag{25}$$

where q is the dimension of $\boldsymbol{\theta}$, and $G(\boldsymbol{\theta})$ is the Godambe (Godambe, 1960) or "sandwich" information matrix of a single observation (Varin, 2008; Varin et al., 2011). The Godambe information matrix is used in the CML method for the calculation of standard errors. When employing CML methods, the asymptotic covariance matrices of the CML estimators are different from the Fisher information matrix $I(\boldsymbol{\theta})$ as they are not from full likelihoods and thus not fully efficient. The Godambe information matrix is defined as

$$G(\boldsymbol{\theta}) = H(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}H(\boldsymbol{\theta}) \tag{26}$$

where $H(\boldsymbol{\theta})$is the sensitivity matrix and equal to the negative Hessian matrix of composite log-likelihood or

$$H(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial \boldsymbol{\theta}} s(\boldsymbol{\theta}; \boldsymbol{y})\right] = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} c(\boldsymbol{\theta}; \boldsymbol{y})\right], \tag{27}$$

$s(\boldsymbol{\theta}, \boldsymbol{y})$ is the composite score function and $c(\boldsymbol{\theta}, \boldsymbol{y})$ is a composite log-likelihood function. As the composite score is a linear combination of valid likelihood score functions, its unbiasedness follows under the regularity conditions. $J(\boldsymbol{\theta})$ is the variability matrix calculated as

$$J(\boldsymbol{\theta}) = Var\big(s(\boldsymbol{\theta}; \boldsymbol{y})\big) = E_{\boldsymbol{\theta}}[s(\boldsymbol{\theta}; \boldsymbol{y}) s(\boldsymbol{\theta}; \boldsymbol{y})'] = E_{\boldsymbol{\theta}}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} c(\boldsymbol{\theta}; \boldsymbol{y})\right)\left(\frac{\partial}{\partial \boldsymbol{\theta}} c(\boldsymbol{\theta}; \boldsymbol{y})\right)'\right], \tag{28}$$

which is the expected outer product of the composite score function. For true likelihoods, the $H(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$ so that $G(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$. In general, $H(\boldsymbol{\theta}) \neq J(\boldsymbol{\theta})$ because the likelihood terms forming the composite score function are likely to be correlated. The fact that $H(\boldsymbol{\theta}) \neq J(\boldsymbol{\theta})$ in composite likelihood is indicative of the loss of efficiency of CML estimators compared to MLE (Martin et al., 2019; Varin, 2008).

Analytical forms of particularly $J(\boldsymbol{\theta})$ are difficult to derive so that sample or empirical estimates need to be used instead for which a consistent estimate is available

by the Law of Large Numbers (Xi & Browne, 2014). The sample estimates of $H(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ are

$$\widehat{H}\left(\widehat{\boldsymbol{\theta}}_{PML}\right) = -\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\boldsymbol{\theta}}s(\boldsymbol{\theta};\boldsymbol{y}_n)\frac{\partial}{\partial\boldsymbol{\theta}}s(\boldsymbol{y}_n,\boldsymbol{\theta}) = -\frac{1}{N}\sum_{n=1}^{N}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}c(\boldsymbol{\theta};\boldsymbol{y}_n), \tag{29}$$

and

$$\hat{J}\left(\widehat{\boldsymbol{\theta}}_{PML}\right) = \frac{1}{N}\sum\nolimits_{n=1}^{N}s(\boldsymbol{\theta};\boldsymbol{y}_n)s(\boldsymbol{\theta};\boldsymbol{y}_n)' = \frac{1}{N}\sum\nolimits_{n=1}^{N}\left(\frac{\partial}{\partial\boldsymbol{\theta}}c(\boldsymbol{\theta};\boldsymbol{y}_n)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}c(\boldsymbol{\theta};\boldsymbol{y}_n)\right)' \tag{30}$$

Standard errors (SE) of the CML estimates are then obtained as

$$SE\left(\widehat{\boldsymbol{\theta}}_{PML}\right) = \frac{1}{\sqrt{\hat{G}(\widehat{\boldsymbol{\theta}}_{PML})}} \tag{31}$$

which is similar to how standard errors are derived using the Fisher information matrix.

Test statistics for inference and model selection criteria are also available for CML estimation, which requires the calculation of $J(\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$. Possible test statistics include LI goodness-of-fit statistics (Maydeu-Olivares & Joe, 2006) and composite likelihood information criterion introduced by Varin and Vidoni (2005) through which indices similar to AIC and BIC (Gao & Song, 2010) can be found.

4.3 Proposed CML Estimation and the IRT Approach

4.3.1 Pairwise Estimation and the IRT Approach

Recall the likelihood to be maximized for estimating categorical item responses given in equation (13), $L(\boldsymbol{\theta};\boldsymbol{y}) = \sum_{i=1}^{N}\pi_i(\boldsymbol{\theta})$. In the IRT approach with the assumption of local or conditional independence, the probability of response pattern $\boldsymbol{y_i} = (y_{i1},\ldots,y_{iJ})$ (i.e, $\pi_i(\boldsymbol{\theta})$) factors into a product over the individual item response category probabilities of

$$P(y_{ij} = y_{ij} |\, \boldsymbol{\xi}_i; \boldsymbol{\theta}_j) = f_j(y_{ij} | \boldsymbol{\xi}_i; \boldsymbol{\theta}_j), \tag{32}$$

where $\boldsymbol{\theta_j}$ is a vector of item-specific parameters and f_j is an item response function, which gives us

$$P(\boldsymbol{y}_i |\, \boldsymbol{\xi}_i; \boldsymbol{\theta}) = \prod_{j=1}^{J} P(y_{ij} = y_{ij} |\, \boldsymbol{\xi}_i; \boldsymbol{\theta}) = \prod_{j=1}^{J} f_j(y_{ij} | \boldsymbol{\xi}_i; \boldsymbol{\theta}). \tag{33}$$

From this, one can determine the joint distribution of $\boldsymbol{y}$, which leads to the joint likelihood function

$$L(\boldsymbol{\xi}; \boldsymbol{\theta} | \boldsymbol{y}) = \prod_{i=1}^{N} \prod_{j=1}^{J} f_j(y_{ij} | \boldsymbol{\xi}_i; \boldsymbol{\theta}_j). \tag{34}$$

Further assuming that the entries in $\boldsymbol{\xi_i}$ are independent and identically distributed and follow a cumulative distribution function F, the full marginal likelihood is

$$L(\boldsymbol{\theta} | \boldsymbol{y}) = \prod_{i=1}^{N} \int \left[\prod_{j=1}^{J} f_j(y_{ij} | \boldsymbol{\xi}_i; \boldsymbol{\theta}_j) \right] \phi(\boldsymbol{\xi}_i) d\boldsymbol{\xi}_i, \tag{35}$$

where $\phi(\boldsymbol{\xi}_i)$ refers to the density function of F that is usually set to standard normal.

In place of maximizing this full marginal likelihood, the proposed estimator will maximize the following composite likelihood function:

$$\begin{aligned} L_c(\boldsymbol{\theta} | (\boldsymbol{y_j}, \boldsymbol{y_j'})) = \prod_{i=1}^{N} \Bigg\{ & \left[\prod_{j=1}^{J} \int f_j(y_{ij} | \boldsymbol{\xi}_i; \boldsymbol{\theta}_j) \phi(\boldsymbol{\xi}_i) d\boldsymbol{\xi}_i \right] \\ & \times \left[\prod_{j=1}^{J-1} \prod_{j'=j+1}^{J} \int f_j(y_{ij} | \boldsymbol{\xi}_i; \boldsymbol{\theta}_j) f_{j'}(y_{ij'} | \boldsymbol{\xi}_i; \boldsymbol{\theta}_{j'}) \phi(\boldsymbol{\xi}_i) d\boldsymbol{\xi}_i \right] \Bigg\} \end{aligned} \tag{36}$$

This function is the product of all of the univariate and bivariate or pairwise marginal likelihoods that each correspond to $P(y_{ij} = y_{ij} |\, \boldsymbol{\theta_j})$ and $P(y_{ij} = y_{ij}, y_{ij'} = y_{ij'} |\, \boldsymbol{\theta_j}; \boldsymbol{\theta_{j'}})$, respectively. The log-likelihood then becomes

$$l_c(\boldsymbol{\theta}|(\boldsymbol{y_j}, \boldsymbol{y_j'})) = \sum_{i=1}^{N}\sum_{j=1}^{J} ln\left[\int P(y_{ij} = y_{ij}|\boldsymbol{\theta}_j, \boldsymbol{\xi}_i)\phi(\boldsymbol{\xi}_i)d\boldsymbol{\xi}_i\right]$$

$$+\sum_{i=1}^{N}\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J} ln\left[\int P(y_{ij} = y_{ij}, y_{ij'} = y_{ij'}|\boldsymbol{\theta}_j; \boldsymbol{\theta}_{j'}, \boldsymbol{\xi}_i)\phi(\boldsymbol{\xi}_i)d\boldsymbol{\xi}_i\right] \quad (37)$$

Again, as $\boldsymbol{y}_r$, $r = 1, \ldots, R = \prod_{j=1}^{J} K_j$, denotes the rth possible item response vector for J, equation (35) is equal to

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{r=1}^{R}\int_{\boldsymbol{\xi}} \left[\prod_{j=1}^{J} f_j(y_j|\boldsymbol{\xi}; \boldsymbol{\theta_j})\right]\phi(\boldsymbol{\xi})d\boldsymbol{\xi}, \quad (38)$$

And equation (37) becomes

$$l_c(\boldsymbol{\theta}|(\boldsymbol{y_j}, \boldsymbol{y_j'})) = \sum_{j=1}^{J}\sum_{c_j=1}^{K_j} n_{c_j}^{(y_j)} ln\left[\int_{\boldsymbol{\xi}} P(y_j = c_j|\boldsymbol{\theta_j}, \boldsymbol{\xi})\phi(\boldsymbol{\xi})d\boldsymbol{\xi}\right]$$

$$+\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\sum_{c_j=1}^{K_j}\sum_{c_{j'}=1}^{K_{j'}} n_{c_j c_{j'}}^{(y_j y_{j'})} ln\left[\int_{\boldsymbol{\xi}} P(Y_j = c_j, Y_{j'} = c_{j'}|\boldsymbol{\theta_j}; \boldsymbol{\theta_{j'}}, \boldsymbol{\xi})\phi(\boldsymbol{\xi})d\boldsymbol{\xi}\right]. \quad (39)$$

Equations (37) and (39) are obviously similar to equation (17) as they both intend to maximize the univariate and bivariate marginal likelihoods of item responses. However, it is important to remember that equation (17) is based on the UV approach and thus, integrates over the items. On the other hand, the pairwise log-likelihoods of Equations (37) and (39) are built upon the IRT approach where integration is over the latent ability vector. Notice that this difference results in differences regarding the integrals. In the UV approach, the high-dimensional integrals are replaced to require evaluation of up to only two-dimensional integrals, one-dimensional for univariate likelihoods and two-dimensional for bivariate or pairwise likelihoods, regardless of the number of observed or latent variables. However, in the IRT approach, the integrals are

not affected in any way by using the CML estimator. This may explain why LI methods are not as favored in the IRT approach as the UV approach. We want to avoid dealing with high-dimensional integrals, especially in cases where no closed-form solutions exist. Using CML estimation based on the UV approach enables this. While this particular benefit does not apply to CML estimation under the IRT approach as defined above, it can help to significantly reduce the number of response patterns to account for in parameter estimation. Furthermore, it may be possible to extend CML estimation to reduce the dimensionality under the IRT as well.

One thing that CML estimation using both approaches as above have in common is that they are both composite marginal maximum likelihood estimation methods. Marginal maximum likelihood (MML) methods make assumptions about the latent distributions (i.e., items for the UV approach and latent abilities for the IRT approach) and integrate them out to maximize marginal likelihoods.

4.3.2 Classification of IRT Models and Applications of Pairwise Estimation

For this dissertation, the data generation method only produces the univariate and bivariate probabilities for a set of items. Such data is fit to IRT models using only the bivariate (log-)likelihoods in equations (37) and (39). The univariate likelihoods were dropped following general consensus that they add little to estimation (e.g., Katsikatsou et al., 2012). In short, the proposed CML method is the pairwise marginal maximum likelihood (PMML) estimation. Equations (37) or (39) is a generalized version of the (log-) likelihood function that serves as the basis for many IRT models that may be of interest.

IRT models are classified according to the type of items they model. Dichotomous IRT models are used for binary scored items or responses. Polytomous models are used for nominal or ordered categorical items. IRT models can also be divided as

unidimensional or multidimensional. Unidimensional IRT models assume a single latent trait while multidimensional models assume multiple latent traits underlying the item response data. Also, models can be categorized depending on whether the latent variable distribution is continuous or categorical. Although the models that have been reviewed so far all assume a continuous latent factor, diagnostic classification models (DCMs; Rupp, Templin, & Henson, 2010) assume multiple categorical latent factors coined attributes. DCMs are special types of IRT models modified to handle latent abilities that are categorical in nature (i.e., mastery or non-mastery of particular sets of attributes).

Specifically, this dissertation focused on the five IRT models found in Bonifay and Cai (2017). The five models were the exploratory factor analytic (EFA) two-parameter logistic (2PL) model; bifactor 2PL model; deterministic input, noisy and-gate (DINA) model; deterministic input, noisy or-gate (DINO) model; and the unidimensional three-parameter logistic (3PL) model. As mentioned in Bonifay (2015), while the models involve different multidimensional factor structures, they are all dichotomous IRT models with four out of five of them being 2PL models. Thus, let us consider a multidimensional version of the 2PL model under the same usual scenario with $i = 1, \cdots, N$ respondents and $j = 1, \cdots, J$ items. Let $y_{ij} \in \{0,1\}$ denote the item score for respondent i to item j. Furthermore, let's change the notation slightly so that $\boldsymbol{\theta}$ now denotes the vector of q latent abilities. The item response function denoting the conditional probability of a correct or positive response by an individual with ability vector $\boldsymbol{\theta}$ to item j is

$$P_j(\boldsymbol{\theta}) = P\left(y_j = 1 | \boldsymbol{\theta}; \boldsymbol{\gamma}\right) = \frac{1}{1 + \exp\left(-\left(\boldsymbol{a}_j'\boldsymbol{\theta} + c_j\right)\right)} \tag{40}$$

$\boldsymbol{a}_j'$ and c_j are the item discrimination and intercept parameters in an item and $\boldsymbol{\gamma}$ is the vector of all freely estimated item parameters. The probability of an individual incorrectly answering an item is then simply $Q_j(\boldsymbol{\theta}) = 1 - P_j(\boldsymbol{\theta})$. The slope-intercept

parameterization is used as it is simpler to differentiate (Baker & Kim, 2004) and more readily interpretable for multidimensional models (Bonifay, 2019). Let $\boldsymbol{u}_k, k = 1, \ldots, 2^J$ denote the kth possible item response vector where $2^J = \prod_{j=1}^{J} K_j$ is the total number of possible response patterns $\boldsymbol{u}_k$. Under the assumption of local independence, the probability of score pattern $\boldsymbol{u}_k$ for an individual of ability $\boldsymbol{\theta}$ is

$$P(\boldsymbol{u}_k \mid \boldsymbol{\theta}) = \prod_{j=1}^{J} P\left(y_j = y_j | \boldsymbol{\theta}; \boldsymbol{\gamma}\right) = \prod_{j=1}^{J} P_j(\boldsymbol{\theta})^{u_{kj}} Q_j(\boldsymbol{\theta})^{1-u_{kj}}. \tag{41}$$

The unconditional probability of observing pattern k is

$$P(\boldsymbol{u}_k) = \int_{\boldsymbol{\theta}} [P(\boldsymbol{u}_k \mid \boldsymbol{\theta})]\phi(\boldsymbol{\theta})d\boldsymbol{\theta} = P_k \tag{42}$$

An individual score pattern assigns him or her to one of the 2^J mutually exclusive item response vectors. Under the assumption that they are independent, the marginal likelihood function is

$$L = \prod_{k=1}^{2^J} P_k^{n_k} = \prod_{k=1}^{2^J} \left[\int_{\boldsymbol{\theta}} \prod_{j=1}^{J} P_j(\boldsymbol{\theta})^{u_{kj}} Q_j(\boldsymbol{\theta})^{1-u_{kj}} \phi(\boldsymbol{\theta}) d\boldsymbol{\theta} \right]^{n_k} \tag{43}$$

Thus, the log-likelihood function becomes

$$l = \sum_{k=1}^{2^J} n_k P_k = \sum_{k=1}^{2^J} n_k \ln \left[\int_{\boldsymbol{\theta}} \prod_{j=1}^{J} P_j(\boldsymbol{\theta})^{u_{kj}} Q_j(\boldsymbol{\theta})^{1-u_{kj}} \phi(\boldsymbol{\theta}) d\boldsymbol{\theta} \right] \tag{44}$$

where n_k and P_k are respectively the observed frequency and probability under the model for the response pattern k with $\sum_{k=1}^{2^J} n_k = N$, and $\sum_{k}^{2^J} P_k = 1$.

In PML estimation, we are not interested in the full 2^J item response vectors but only item response vectors for item pairs $\left(y_j, y_{j'}\right), j \neq j$. In other words, only the bivariate observed frequencies and probabilities under the model for every item pair need to be considered. The pairwise marginal log-likelihood function to be maximized is

$$pl = \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \ln L\left(y_j, y_j'\right) = \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \sum_{c_j=1}^{m_j} \sum_{c_{j'}=1}^{m_{j'}} n_{c_j c_{j'}}^{(y_j y_{j'})} \ln P_{c_j c_{j'}}^{(y_j y_{j'})}$$

$$= \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \sum_{k=1}^{2^2} n_{c_j c_{j'}}^{(y_j y_{j'})} \ln P_{c_j c_{j'}}^{(y_j y_{j'})} \tag{45}$$

$$= \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \sum_{k=1}^{2^2} n_{c_j c_{j'}}^{(y_j y_{j'})} \ln \left[\int_{\boldsymbol{\theta}} P_j(\boldsymbol{\theta})^{u_{kj}} Q_j(\boldsymbol{\theta})^{1-u_{kj}} P_{j'}(\boldsymbol{\theta})^{u_{kj'}} Q_{j'}(\boldsymbol{\theta})^{1-u_{kj'}} \right]$$

which reduces having to multiply over 2^J probabilities to only $\frac{(J-1)\times J}{2} \times 2^2$ for dichotomous models. $\frac{(J-1)\times J}{2}$ refers to the number of possible unique item pairs so that in essence, we only have to account for four marginal pairs of probabilities involved given by

$$\begin{aligned} P\left(Y_j = 1, Y_{j'} = 1\right) &= P_j(\boldsymbol{\theta}) P_{j'}(\boldsymbol{\theta}) \\ P\left(Y_j = 1, Y_{j'} = 0\right) &= P_j(\boldsymbol{\theta}) Q_{j'}(\boldsymbol{\theta}) \\ P\left(Y_j = 0, Y_{j'} = 1\right) &= Q_j(\boldsymbol{\theta}) P_{j'}(\boldsymbol{\theta}) \\ P\left(Y_j = 0, Y_{j'} = 0\right) &= Q_j(\boldsymbol{\theta}) Q_{j'}(\boldsymbol{\theta}). \end{aligned} \tag{46}$$

Item parameters are estimated by finding the sets of item parameters $\boldsymbol{c}$ and $\boldsymbol{a}$ that maximize equation (45). The integral(s) has no closed form but can be approximated by numerical methods such as the Gauss-Hermite quadrature (Baker & Kim, 2004). Also, the usual dimension reduction can be employed for models such as the bifactor model (Cai, Yang, & Hansen, 2011).

The likelihood function of the equation does not change for different dichotomous IRT models such as the unidimensional 3PL. For polytomous IRT models, only the number of marginal pairs of probabilities would increase. For example, if there were 3 categories per item *j*, we would need to consider $3^2 = 9$ for every item pair.

For DCMs, the likelihood function differs in relation to the latent ability vector $\boldsymbol{\theta}$. In DCMs, $\boldsymbol{\theta}$ is not continuous but a vector of categorical latent factors called attributes. Let $\boldsymbol{\theta}$ be a collection of binary attributes so that they become a total of $C = 2^q$ classes consisting of combinations of attribute mastery levels. Let us consider $\boldsymbol{\theta}$ as synonymous with $\boldsymbol{\alpha}_c = (\alpha_1, \dots, \alpha_q)$ where $\alpha_q \in \{0,1\}$ that denotes a specific attribute profile c. Then, the interpretation of $P_j(\boldsymbol{\theta})$ of equation (40) changes to be the probability of correct response to item j by a respondent in latent class c, $c = 1, \cdots C$ or or $P_j(\boldsymbol{\alpha}_c)$. Also, the conditional probability of observing a particular response pattern is given $\boldsymbol{\alpha}_c$ as

$$P(\boldsymbol{u}_k \mid \boldsymbol{\alpha}_c) = \prod_{j=1}^{J} P_j\big(y_j = y_j|\boldsymbol{\alpha}_c\big) = \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_c)^{u_{kj}}\, Q_j(\boldsymbol{\alpha}_c)^{1-u_{kj}} \tag{47}$$

The unconditional probability is

$$P(\boldsymbol{u}_k) = \sum_{c=1}^{C} v_c[P(\boldsymbol{u}_k \mid \boldsymbol{\alpha}_c)] = P_k \tag{48}$$

with v_c being a mixing probability $\big(\sum_{c=1}^{C} v_c = 1.0\big)$ and denotes the probability of membership in latent class or profile c. Then the full marginal likelihood function becomes

$$L = \prod_{k=1}^{2^J} P_k^{n_k} = \prod_{k=1}^{2^J} \left[\sum_{c=1}^{C} v_c \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_c)^{u_{kj}} Q_j(\boldsymbol{\alpha}_c)^{1-u_{kj}}\right]^{n_k} \tag{49}$$

which makes the log-likelihood function to be maximized

$$l = \sum_{k=1}^{2^J} n_k \ln P_k = \sum_{k=1}^{2^J} n_k \ln \left[\sum_{c=1}^{C} v_c \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_c)^{u_{kj}} Q_j(\boldsymbol{\alpha}_c)^{1-u_{kj}}\right] \tag{50}$$

and the pairwise counterpart is

$$pl = \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \sum_{k=1}^{2^2} n_k^{(y_j y_{j'})} \ln \left[\sum_{c=1}^{C} v_c \big(P_j(\boldsymbol{\alpha}_c)^{u_{kj}} Q_j(\boldsymbol{\alpha}_c)^{1-u_{kj}} P_{j'}(\boldsymbol{\alpha}_c)^{u_{kj'}} Q_{j'}(\boldsymbol{\alpha}_c)^{1-u_{kj'}}\big)\right] \tag{51}$$

v_c refers to parameters of the structural model of a DCM (as opposed to the measurement model relating attributes and observed item responses). While various structural models are present in the literature (Rupp et al., 2010; Thompson, 2018), Bonifay and Cai's (2017) study chose to impose a higher-order structure on v_c where we regress the attributes on a higher-order, continuous latent trait θ (multiple traits are also possible). The probability of mastering each attribute is assumed to depend on a respondent's place in this higher-order dimension. Assuming that the mastery of a set of skills for a respondent is related to a unidimensional trait θ, and assuming conditional independence of the latent attributes given θ, the probability model of $\boldsymbol{\alpha}_c$ conditional on θ is

$$P(\boldsymbol{\alpha}_c \mid \theta) = \prod_{k=1}^{q} P(\alpha_k \mid \theta) \tag{52}$$

where q is the number of attributes. Because the attributes are binary variables, they can be treated as if they were items and technically any IRT model may be used for $P(\alpha_k = 1 \mid \theta)$ (Hansen, 2013). Bonifay and Cai (2017) impose a 2PL model for all attributes so that

$$P(\alpha_k = 1 \mid \theta) = \frac{1}{1 + \exp\left(-(c_k + a_k\theta)\right)} \tag{53}$$

where c_k and a_k are the intercept and slope parameters, respectively, that resemble item easiness and discrimination parameters in IRT. However, we should keep in mind that these are higher-order structural parameters and that the higher-order model is being fit to attribute profile probabilities and not the item response patterns (Hansen, 2013). The marginal probability of an observed response pattern of $\boldsymbol{u}_k$ then can be represented as

$$P(\boldsymbol{u}_k) = \int_{\theta} \left\{ \prod_{k=1}^{q} P(\alpha_k \mid \theta) \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_c)^{u_{kj}} Q_j(\boldsymbol{\alpha}_c)^{1-u_{kj}} \right\} g(\theta) d\theta \tag{54}$$

In many applications, θ is assumed to be normally distributed with mean 0 and variance 1. Then the full marginal log-likelihood function becomes

$$l = \sum_{k=1}^{2^J} n_k \left[\int_{\theta} \left\{ \prod_{k=1}^{q} P(\alpha_k \mid \theta) \prod_{j=1}^{J} P_j(\boldsymbol{\alpha}_c)^{u_{kj}} Q_j(\boldsymbol{\alpha}_c)^{1-u_{kj}} \right\} g(\theta) d\theta \right] \tag{55}$$

and the pairwise counterpart is

$$pl = \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J} \sum_{k=1}^{2^2} n_k^{(y_j y_{j'})} \ln \left[\int_{\theta} \left\{ \prod_{k=1}^{q} P(\alpha_k \mid \theta) \left(P_j(\boldsymbol{\alpha}_c)^{u_{kj}} Q_j(\boldsymbol{\alpha}_c)^{1-u_{kj}} (\boldsymbol{\alpha}_c)^{1-u_{kj'}} \right) \right\} g(\theta) d\theta \right] \tag{56}$$

The resulting pairwise likelihoods for each model in question can then be maximized using their logarithms and standard numerical procedures such as the Newton-Raphson (NR) algorithm (Bock & Lieberman, 1970) or the Expectation-Maximization (EM) algorithm (Bock & Akin, 1981; Dempster, Laird, Rubin, 1977).

Although the results are not reported, both the MML method in Bock and Lieberman (1970) and the EM algorithm performed well for the Bifactor and EFA models but only the latter worked well for DCMs. In EM algorithms, each response pattern for item pairs is evaluated over the grid of $\boldsymbol{\theta}$ values using initial parameter estimates. Then, given the marginal evaluated pairwise response patterns, we generate an "expected" table of response patterns across the grid of $\boldsymbol{\theta}$ values. This is the expectation step of the EM algorithm. Then, the item parameters are maximized as if the expectation table was what was really observed, using the $\boldsymbol{\theta}$ grid as the predictor variable. This is the maximization step of the EM algorithm. The item parameters are updated from the maximization step, which in turn are used to calculate new expected tables. Like this,

the EM algorithm is an iterative approach that toggles between the two modes of the E-step and M-step. The algorithm is terminated or said to have converged if change in some convergence criteria between successive iterations is smaller than a designated convergence tolerance value. A relative error criterion $\frac{|l_{n-1}-l_n|}{|l_{n-1}|} < 10^{-6}$ where l_n refers to the log-likelihood value at the nth iteration and an absolute error criterion maximum parameter change $< 10^{-6}$ were mainly used.

After obtaining the parameters using the PML approach, we can compute SE using the Godambe information matrix (Godambe, 1960) constructed from the sample estimates of the variability matrix ($\hat{J}(\hat{\boldsymbol{\theta}}_{PML})$) and the sensitivity matrix ($\hat{H}(\hat{\boldsymbol{\theta}}_{PML})$). The former was calculated as

$$\begin{aligned}\hat{J}(\hat{\boldsymbol{\theta}}_{PML}) &= \frac{1}{n}\sum_{i=1}^{n}\left(\nabla pl(\hat{\boldsymbol{\theta}}_{PML};\ \boldsymbol{y_i})\right)\left(\nabla pl(\hat{\boldsymbol{\theta}}_{PML};\ \boldsymbol{y_i})\right)' \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\nabla pl(\hat{\boldsymbol{\theta}}_{PML};\ y_{ij}, y_{ij'})\right)\left(\sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\nabla pl(\hat{\boldsymbol{\theta}}_{PML};\ y_{ij}, y_{ik})\right)'\end{aligned} \tag{57}$$

The latter can be consistently estimated with the observed pairwise likelihood information as

$$\hat{H}(\hat{\boldsymbol{\theta}}_{PML}) = -\frac{1}{n}\sum_{i=1}^{n}\left(\nabla^2 pl(\hat{\boldsymbol{\theta}}_{PML};\ \boldsymbol{y_i})\right) \tag{58}$$

which is equal to minus of the Hessian matrix. The square root of the inverse of the resulting Godambe information matrix gives the SEs as shown in equation (31).

The calculation of the sensitivity matrix is not an issue as the Hessian matrix can be easily calculated as a byproduct of the item parameter estimation process. However, the variability matrix is tricky as it requires the calculation of the variance of the composite score vector across individuals. Furthermore, by definition, it is possible to calculate the variability matrix only when the full response or data pattern is available.

Therefore, while the proposed PML estimation approach provides the equations to calculate the SEs, if the data fed into the estimator has information up to only bivariate margins (as opposed to being higher-dimensional data being collapsed into the margins), they cannot be calculated. We might consider using other alternative methods such as bootstrapping (Lui et al., 2015) for the calculation of variance in the estimation. Nonetheless, although it is favorable to have the option to compute SEs, it does not necessary deter us from investigating FP using the proposed LI-based data generation and estimation method, which is the main motivation of this study.

Chapter V

Simulation Study

This chapter summarizes a series of simulation studies conducted to achieve the three main objectives of this study in the order of data generation, model estimation, and evaluation of FP using the proposed LI approach. Although the proposed LI method to IRT data generation and model estimation described in Chapters 3 and 4 apply in theory to a myriad of different IRT models, there is yet to validate their appropriateness empirically. Bonifay and Cai's (2017) study was set as the reference point in these first steps to exploring the utility of the proposed LI methods, particularly in terms of FP investigation. The goal was to see if the same results could be derived using the FI approach under the well-defined multinomial framework that Bonifay and Cai (2017) employed. Replications of the results of Bonifay and Cai (2017) using the newly proposed methods will imply the suitability of the approach for appraising the FP of IRT models. Through this, the effects of complexity due to the estimation method and data generation method may also be realized.

5.1 Overview of Common Simulation Conditions

All simulation studies in this chapter are centered around the data and model conditions found in Bonifay and Cai (2017). Thus, the number of items was set to seven dichotomously scored or binary items. Also, four out of the five IRT models of Bonifay and Cai (2017) were chosen as the IRT models of interest. The 3PL unidimensional model was excluded due to two main reasons. First of all, the initial results of fitting the unidimensional 3PL model using the PML estimation failed to provide any explainable

results. This was regardless of placing priors on the pseudo-guessing parameters as commonly done. Furthermore, and more importantly, recent research has highlighted the impact that the choice of priors for item parameters can have on FP. The 3PL model in Bonifay and Cai (2017) also used priors on the pseudo-guessing parameters to promote estimation stability, whose influence was not intended nor investigated. The remaining four IRT models were all essentially 2PL models and equal in the number of parameters. Bonifay and Cai (2017) set the number of item parameters to the arbitrary number of 20 and configured the four models to all match in the number of parameters. Path diagrams depicting the four models and their specifications are in Figure 5. 1.

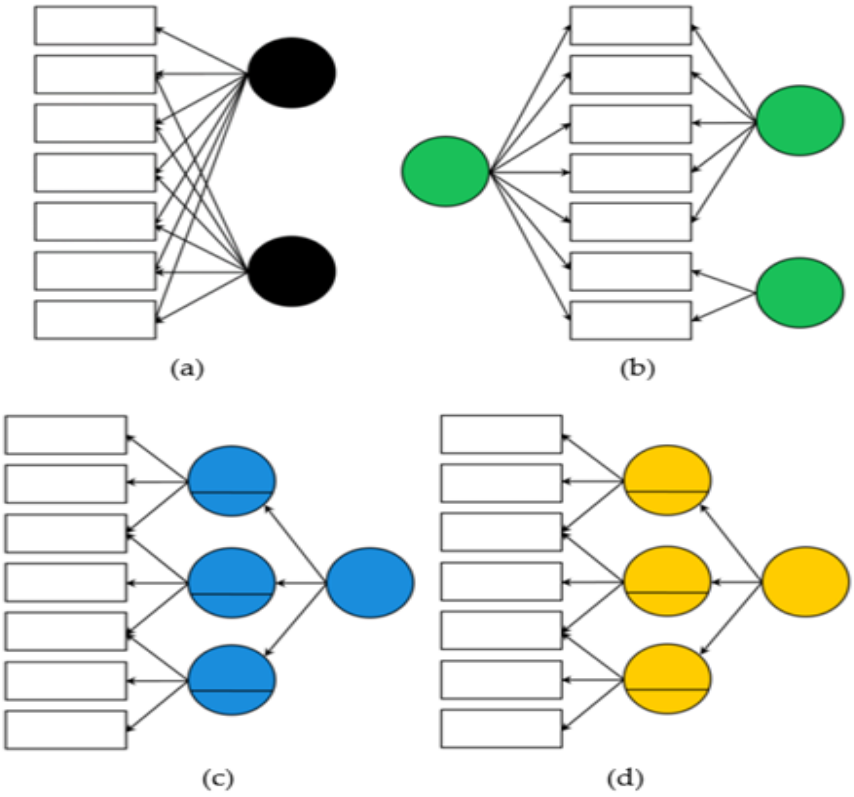


Figure 5. 1: Path diagrams of Models in Bonifay and Cai (2017)

Note. (a) exploratory factor analysis (EFA) 2PL model; (b) bifactor 2PL model; (c) deterministic input, noisy and-gate (DINA) model; and (d) deterministic input, noisy or-gate (DINO) model.
Adapted from "On the Complexity of Item Response Theory Models", by W. Bonifay, and L. Cai, 2017, *Multivariate Behavioral Research*, p 4.

To elaborate on each of the four models, the EFA model was specified so that two factors represented the seven items. This model was considered the baseline model because, being an exploratory model, it is by definition a flexible model that can accommodate a variety of data structures as the particular paths among items and latent factors are not fixed *a priori* (Aytürk Ergin, 2020). In this model, items loaded freely on both factors but the loading for the first item on the second factor being constrained to 0 for identification purposes. The factor loading matrix for the EFA model was

$$\begin{pmatrix} a_{11} & 0 \\ a_{12} & a_{22} \\ a_{13} & a_{23} \\ a_{14} & a_{24} \\ a_{15} & a_{25} \\ a_{16} & a_{26} \\ a_{17} & a_{27} \end{pmatrix}. \tag{59}$$

The bifactor model had two specific factors in addition to one general factor. All seven items loaded on the general factor. The first five items loaded on the first specific factor, and the remaining two items loaded on the second specific factor. The loadings for the second specific factor were set to be equal for model identification purposes. The factor-loading matrix of the bifactor model was

$$\begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & 0 \\ a_{41} & a_{42} & 0 \\ a_{51} & a_{52} & 0 \\ a_{61} & 0 & a_{62} \\ a_{71} & 0 & a_{62} \end{pmatrix}. \tag{60}$$

The first column represents the general factor, the second column represents the first specific factor, and the third column represents the second specific factor.

In the DINA and DINO models, three attributes were chosen. The Q-matrix for these two models is presented in Table 5. 1.

Table 5. 1: Q-matrix for DCMs

Items	Attributes		
	a_1	a_2	a_3
Item 1	1	0	0
Item 2	1	0	0
Item 3	1	1	0
Item 4	0	1	0
Item 5	0	1	1
Item 6	0	0	1
Item 7	0	0	1

As can be seen in Table 5. 1, the Q-matrix consisted of five simple structure items, meaning one attribute was measured per item, and two complex structure items having two out of three attributes load on an item. That is, the probability of responding to items three and five required an interaction of two attributes: Item three required attributes one and two, and Item five required attributes two and three. DINA and DINO models differ in how these interactions are modeled. Because the DINA model is non-compensatory, mastery in both attributes is required for an increase in response probability on these items. On the other hand, the DINO model is a extreme compensatory model that assumes mastery in either of the attributes can result in maximal probability of item response.

Bonifay (2015) provides multiple model indices including test-level indices such as Y2/N and item-level fit indices such as LD-X^2 (Chen & Thissen, 1997) for use in qualifying FP. Aytürk Ergin (2020) added a slew of others. Among the model fit indices, the Y2/*N* statistic (Bartholomew & Leung, 2002; Cai et al., 2006) was selected to measure test-level fit in the simulations. The Y2/*N* was chosen not only because it best represents the best of Bonifay and Cai's (2017) study but also because it is a limited-information index that requires only information about the univariate and bivariate margins

themselves, which is all the combination of the proposed SIS data generation and PML estimation methods can provide. The Y2 statistic is traditionally calculated as

$$Y2 = N\left[\sum_{j=1}^{J}\frac{\left(o_j - e_j\right)^2}{e_j\left(1-e_j\right)} + \sum_{j=1}^{J-1}\sum_{j'=j+1}^{J}\frac{\left(o_{jj'} - e_{jj'}\right)^2}{e_{jj'}\left(1-e_{jj'}\right)}\right]. \quad (61)$$

N is the sample size, J is the number of items, $o_j$and e_j are the observed and expected positive or correct response frequencies for item j, and $o_{jj'}$ and $e_{jj'}$are the observed and expected positive or correct response frequencies for item pair jj'. Bonifay and Cai (2017) actually calculated Y2/N based on equation (61) but using information from all the cells in the univariate and bivariate margins. Their method was employed in this study. As an index of the magnitude of the discrepancy between the data and model, it is a "badness-of-fit" index so that higher values indicate worse fit. Y2 was divided by the sample size N to produce the Y2/N statistic to make it independent of sample size (Bonifay & Cai, 2017).

$Y2$ and $Y2/N$ are distinguished from all other indices described here because they do not depend on the number of parameters in the model. In other words, the Y2 statistic does not penalize for number of free parameters in the model. Thus, it is considered the closest analog to root mean squared residual (RMSR; Jöreskog & Sörbom, 1996) that currently exists for discrete data. RMSR is what Preacher (2006) used as the metric of model fit in his study of FP of SEMs because he considered it a "pure" measure of fit unadjusted for the number of parameters or function form of a model and thus, appropriate to compare FPs across competing models.

5.2 Proposed Data Generation Algorithm and 2×2 Tables

For dichotomously scored items, which is the focus of this study, generating data on the lower-order margins reduces to simply sampling 2×2 tables. Then the proposed data generation algorithm is simplified to that of Table 5. 2.

Table 5. 2: Proposed Data Generation Algorithm for 2×2 Tables

1. Randomly sample *j* univariate probabilities from a $Beta(2,2)$ distribution
2. Uniformly sample $p_{00} = P(y_j = 0, y_{j'} = 0)$ from the range of the lower and upper Fréchet bounds
3. Calculate $p_{01} = p_{0+} - p_{00}, p_{10} = p_{+0} - p_{00}$ and $p_{11} = 1 - p_{01} - p_{10} - p_{00}$

In the case of 2×2 tables, we can plot the results of sampling an arbitrarily large number of points based on one data generation scheme or another and graphically verify whether it fits with theory as well as make comparisons. 10,000 points were sampled following the algorithm in Table 5. 2. Results of the bivariate points, which reside in the tetrahedron, are presented in the left panel of Figure 5. 2 (i.e., Figure 5. 2-(A)). The right panel of Figure 5.2 (Figure 5.2-(B)) shows the sampled bivariate points when using the conventional simplex sampling method with just two binary items used in Bonifay and Cai (2017). In the case of two variables, the two methods should provide identical results. The results of both seemed well-matched with theory in that they gave the impression of more or less uniformly distributed points across the 3-dimensional simplex, which was the categorical data space of interest.

A set of corresponding univariate margins for both data generation methods are given in Figure 5. 3. The plots of univariate margins for the proposed SIS approach (depicted in Figure 5.3 in red) are from applying Step 1 of Table 5. 2. Thus, the univariate

margins were sampled from a $Beta(2,2)$ distribution. On the other hand, the plots for the simplex sampling method (blue in Figure 5. 3) are from collapsing the bivariate margins. The results of the two sets of plots were more or less identical indicating that the distribution from which to sample univariate probabilities was appropriate and functioned as expected. Such results provided evidence of the appropriateness and feasibility of the proposed SIS method.

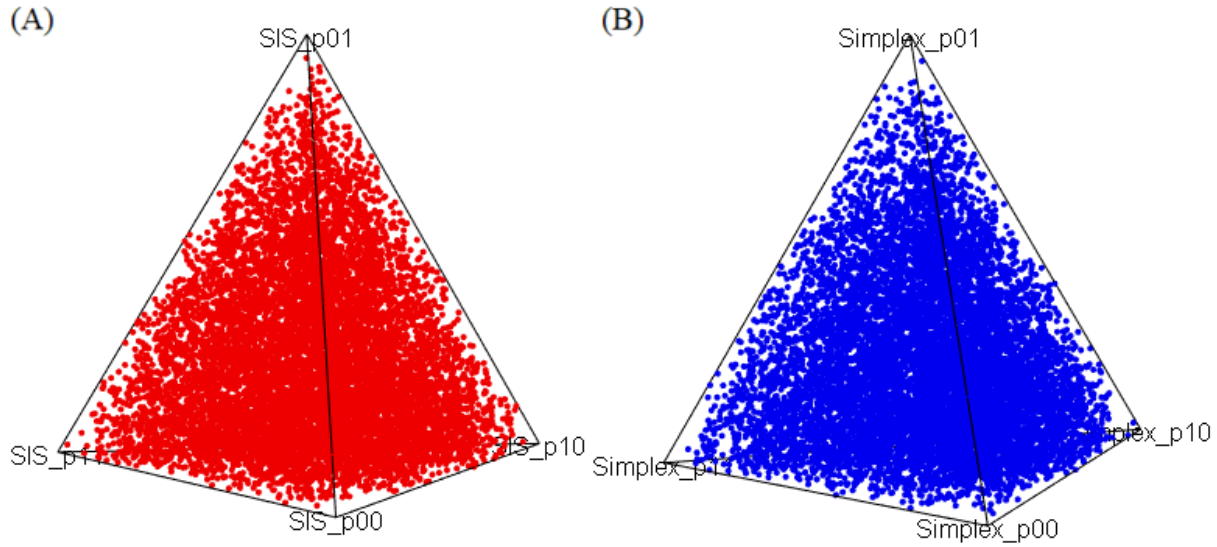


Figure 5. 2: Plot of Bivariate Margins by Sampling Method

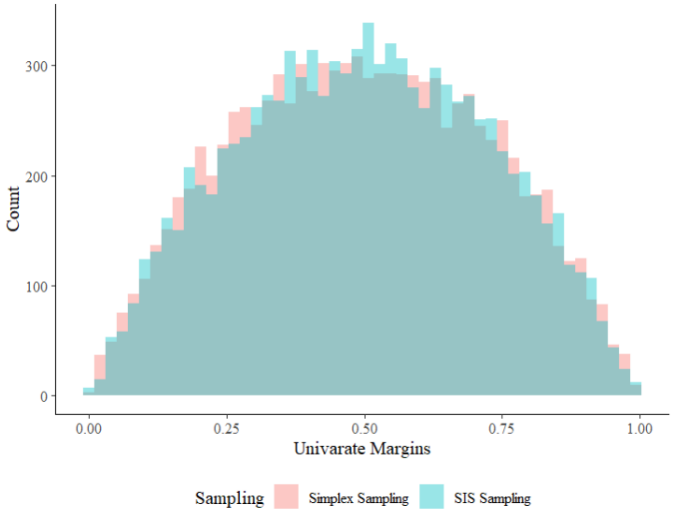


Figure 5. 3: Plot of Univariate Margins by Sampling Method

5.3 Simulation Study 1: Performance of Proposed PML Estimators

The purpose of this simulation study was to examine the performance of proposed PML estimators for the five IRT models in Bonifay and Cai (2017) in terms of item parameter recovery, standard error estimation, and model fit. The focus was not only on whether the proposed PML estimators provided reasonable results when compared to simulated "true" parameters, but also on comparing the performance of the PML method to the full-information maximum likelihood (FIML) method via the widely used expectation-maximization (EM) algorithm (Bock & Akin, 1981).

5.3.1 Simulation Study Setup

Data Generating Models

The data-generating models were set to those of Figure 5. 1. Item parameters needed for data generation were intentionally chosen to be simple but still referencing existing literature and being reflective of realistic settings. For the EFA model, the parameters of interest were standardized factor loadings instead of item parameter estimates. The loadings were drawn from a uniform distribution $U(0.3; 1)$ with a diagonal residual covariance matrix with variances set to 1 (Haslbeck & van Bork, 2022). These values were rounded to the nearest first decimal point. One cross-loading of size 0.1 was included (Li et al., 2020). Furthermore, we allowed the two factors to correlate, which was set to 0.4, which is a moderate correlation (Haslbeck & van Bork, 2022). These served as the parameters for the underlying factor model assuming continuous variables. Thresholds were used to connect these underlying variables with the binary

observed ones. Threshold parameters were set to 0 for all items which split the underlying normal distribution per item roughly in half.

The general consensus on bifactor models is that they are well-suited when the general factor has dominance over the specific factors (Reise et al., 2007; Seo & Weiss, 2015). Data were simulated under such a structure by setting all general factor loadings to 1 and specific factor loadings to 0.6 for the first specific factor and 0.4 for the second one. The item intercept parameters were set at random 0.5 intervals between -1.5 and 1.5. For both the EFA and bifactor, the latent ability vector for respondents was drawn from standard multivariate normal distributions that matched the number of latent factors of each model.

The item parameters for both the DINO and DINA models were generated using response success probabilities where non-masters had relatively low probabilities of correctly responding while masters had relatively high probabilities. The response probabilities for non-masters were drawn from a uniform distribution of *U(0.1,0.3),* while those for masters were drawn from a uniform distribution *U(0.7,0.9)* so that the items could be considered as highly discriminating items. In other words, the guessing (g) and slipping (s) parameters were each drawn from a uniform distribution of *U(0.1,0.3).* These parameters were again rounded. These parameters were then converted to intercept (c in flexMIRT), main and interaction effect (a in flexMIRT) parameters under the log-linear diagnostic classification model (LDCM) parameterization (Rupp et al., 2010) where $g = exp(c)/(1 + exp(c))$ and $s = 1 - (\frac{exp(c+a)}{1+exp(c+a)})$. The base-rate of mastery, also called the marginal attribute difficulty, for each attribute was set equal to 0.5. The tetrachoric correlations reflecting the relationship between factors were also set to an equal value of 0.7 (Kunina-Habenicht et al., 2012).

Data Generation

"True" or "population" parameters drawn according to the settings above were used to generate random response datasets with $N = 10000$ based on a multinomial framework. This resulted in full item response pattern probabilities, which was a conscious choice as we were interested in comparing the proposed PML approach to the conventional FIML approach that requires them for model estimation. There is also the advantage that SEs can be generated for the PML approach as well. For use with the PML approach, the data were collapsed down to their bivariate margins, which is the only data utilized in model estimation.

Analysis Setup

The response datasets were calibrated using the models that matched the respective data generation models for both the LI and FI estimation methods. Bias and RMSE were set as the evaluation criteria for item parameter recovery to gauge the performance of the PML approach of each model (as well as for the FIML approach). Both the performance of the PML approach itself compared with the "true" or population parameters and its relative performance compared to the FIML method were examined. Each model and each estimation method was repeated 50 times. Thus, the total number of conditions was 4 models × 2 estimation methods, which resulted in a total number of iterations of $50 \times 8 = 400$. Data generation, estimation using the PML approach, and calculations used R version 4.1.2. Estimation based on the FIML approach used flexMIRT 3.6.4 (Cai, 2021).

Evaluation Criteria

The accuracy of item parameter recovery was evaluated using bias and root mean square error (RMSE) calculated as follows:

$$\text{Bias}(\hat{\theta}) = \frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta) \tag{62}$$

$$RMSE(\hat{\theta}) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2} \tag{63}$$

R is the number of replications. $\hat{\theta}$ is the estimate of a parameter or its asymptotic SE at the rth replication. θ is the corresponding true value. In the case of SEs, true values (i.e., θs) do not exist, so that the empirical SE was used as a surrogate. The empirical SE for a parameter is the standard deviation of each parameter estimate across replications R or

$$SE_E = \sqrt{\frac{\sum_{r=1}^{R}(\theta_r - \overline{\theta_r})^2}{R-1}} \tag{64}$$

(Katsikatsou et al., 2012; Xi & Browne, 2014).

5.3.2 Results

All results are organized by models following the order in Figure 5. 1. To facilitate comparisons between the "true" values for each model as well as methods, results over replications as well as specific parameter types are provided in a series of tables and graphs. A table for each model is presented with results for each item parameter in the order of the "true" value, the average parameter estimate across replications and corresponding bias and the RMSE for the PML method, and then the average parameter estimate across replications, the bias, and the RMSE for the FIML method. Graphical representations of the bias and RMSE of both methods overlaid are also presented. If

applicable, similar tables and graphs are given for the SE for both methods, starting with PML estimation and following the order of the standard deviation of the parameter estimate across replications, the average SE of a parameter across replications, and the respective bias and RMSE within an estimation method. Lastly, a table with the average bias and RMSE values across replications and all parameters of the same type, including those for SE, and average model fit value (i.e., $Y2/N$) are also provided.

Model recovery results of individuals parameters for the EFA model are provided in Table 5. 3. For EFA models, recovery of the standardized rotated factor loadings (λs) and the correlation parameter between the two factors (i.e., ϕ) were the parameters of interest as opposed to the slope and intercept parameters. For rotated factors, SEs are usually not estimated, and thus results are only given for the loadings and correlation parameters themselves. Figure 5. 4 is a plot of the columns for bias and RMSE for each item parameter in Table 5. 3 for both estimation methods. Averages of bias and RMSE for the same parameter type across replications as well as model fit are summarized in Table 5. 4. The results from the tables and the figures showed that PML estimated loadings and correlation parameters were close to their "true" counterparts. In addition, they were nearly identical to those from the FIML estimation. The parameter that deviated the most from the "true" value for both the PML and FIML methods was the λ_{14}, which was the only parameter with a cross-loading. This is in line with past literature. The average model fit based on the $Y2/N$ was the same for both methods.

Table 5. 3: Recovery Results of Estimates of the EFA Model

Par.	True	PML			FIML		
		Est.	Bias	RMSE	Est.	Bias	RMSE
λ_{11}	0.6	0.595	-0.005	0.014	0.594	-0.006	0.014
λ_{12}	0.5	0.500	0.000	0.019	0.500	0.000	0.019
λ_{13}	0.7	0.700	0.000	0.015	0.701	0.001	0.015
λ_{14}	0.4	0.389	-0.011	0.020	0.389	-0.011	0.020
λ_{15}	0.0	-0.002	-0.002	0.011	-0.003	-0.003	0.011
λ_{16}	0.0	0.004	0.004	0.012	0.004	0.004	0.012
λ_{17}	0.0	0.000	0.000	0.008	0.000	0.000	0.008
λ_{21}	0.0	-0.004	-0.004	0.013	-0.004	-0.004	0.013
λ_{22}	0.0	-0.009	-0.009	0.016	-0.009	-0.009	0.017
λ_{23}	0.0	-0.006	-0.006	0.011	-0.006	-0.006	0.011
λ_{24}	0.1	0.094	-0.006	0.014	0.093	-0.007	0.013
λ_{25}	0.7	0.695	-0.005	0.013	0.695	-0.005	0.013
λ_{26}	0.8	0.802	0.002	0.012	0.802	0.002	0.012
λ_{27}	0.9	0.907	0.007	0.011	0.906	0.006	0.011
ϕ	0.4	0.407	0.007	0.014	0.408	0.008	0.014

Note. λ_1 refers to the first, λ_2 refers to the second factor, and ϕ refers to the correlation parameters. Est.=Estimate.

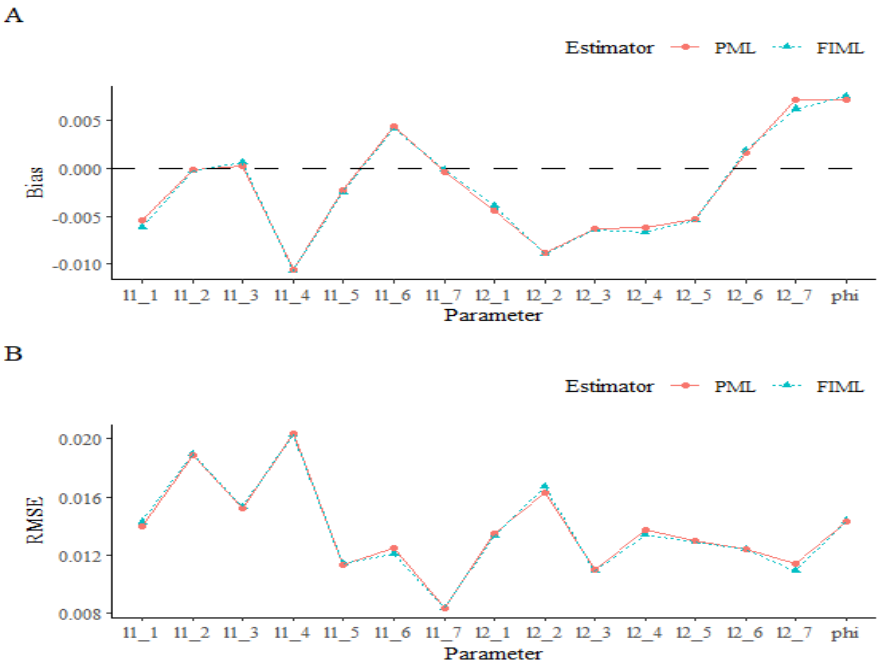


Figure 5. 4: Bias and RMSE of Estimates of the EFA Model
Note. $l = \lambda$ and $phi = \phi$.

Table 5. 4: Average Bias, RMSE of Estimates, and Y2/N values of the EFA Model

	Item Parameters				Model Fit
	λ		ϕ		Y2/N
	Bias	RMSE	Bias	RMSE	
PML	-0.003	0.017	0.007	0.014	0.00062
FIML	-0.003	0.017	-0.003	0.016	0.00065

Note. λ refers to factor loading and ϕ refers to the correlation parameters. Average bias and RMSE was calculated as $Bias(\hat{\theta}) = \frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)}{J}$ and $RMSE(\hat{\theta}) = \sqrt{\frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)^2}{J}}$ averaged over replications. J is the number of parameters of type θ. That is, they were calculated as the average difference between the estimated and "true" parameters across items of parameter types and replications.

In the bifactor model, the parameters of concern were the slopes and intercepts. Model recovery results for average estimates of each of these parameters over replications are provided in Table 5. 5 and Figure 5. 5. The same results for corresponding standard errors are in Table 5. 6 and Figure 5. 6. The overall summary of estimates and standard errors for a parameter type across replications along with model fit is organized in Table 5. 7. The PML method was able to well recover the "true" parameters. Furthermore, its performance was comparable to that of the FIML estimator as well. The results between the FIML and PML estimation were closer to each other than their respective recovery of the population parameters. They also exhibited similar patterns where slope parameters were more difficult to estimate than intercept parameters. This is generally the case and likely there is also the influence of how the factor loading matrix was structured. While results are not presented here, different loading matrices lead to different recovery rates, even when the overall structure and number of parameters were kept the same. For example, when three loadings per specific factor were assumed and the same setting of one general and two specific factors was assumed, the model recovery results for the PML and FIML were hardly distinguishable. In terms of average model fit for the estimator, PML estimation tended to produce smaller $Y2/N$ values.

Table 5. 5: Recovery Results of Estimates of the Bifactor Model

Par.	True	PML			FIML		
		Est.	Bias	RMSE	Est.	Bias	RMSE
a_{G1}	1.0	0.915	-0.085	0.085	0.895	-0.104	0.104
a_{G2}	1.0	0.883	-0.117	0.117	0.877	-0.123	0.132
a_{G3}	1.0	0.919	-0.081	0.081	0.884	-0.116	0.116
a_{G4}	1.0	0.912	-0.088	0.088	0.905	-0.095	0.095
a_{G5}	1.0	0.906	-0.094	0.094	0.861	-0.139	0.139
a_{G6}	1.0	1.096	0.096	0.096	1.034	0.034	0.047
a_{G7}	1.0	1.086	0.086	0.086	1.049	0.049	0.049
a_{S11}	0.6	0.579	-0.021	0.082	0.559	-0.041	0.094
a_{S12}	0.6	0.600	0.000	0.082	0.561	-0.039	0.115
a_{S13}	0.6	0.661	0.061	0.070	0.674	0.074	0.097
a_{S14}	0.6	0.535	-0.065	0.071	0.504	-0.096	0.096
a_{S15}	0.6	0.596	-0.004	0.026	0.601	0.001	0.012
a_{S21}	0.4	0.379	-0.021	0.071	0.439	0.039	0.069
c_1	-1.0	-0.983	0.017	0.017	-0.985	0.015	0.016
c_2	-1.5	-1.496	0.004	0.015	-1.501	-0.001	0.015
c_3	1.5	1.560	0.060	0.060	1.566	0.066	0.066
c_4	-1.5	-1.482	0.018	0.025	-1.487	0.013	0.025
c_5	0.5	0.550	0.050	0.050	0.551	0.051	0.051
c_6	-1.0	-0.982	0.018	0.021	-0.987	0.013	0.019
c_7	-1.5	-1.459	0.041	0.051	-1.473	0.027	0.042

Note. a_G stands for the general factor, a_S to specific factor and c refers to intercept parameters. Est.=Estimate.

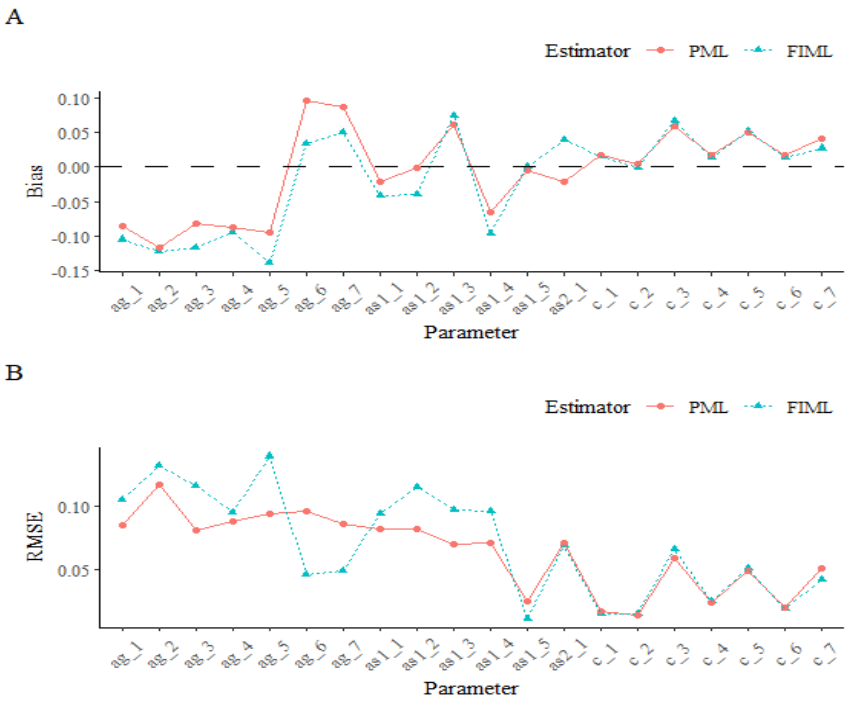


Figure 5. 5: Bias and RMSE of Estimates of the Bifactor Model

Note. a_G stands for the general factor, a_S to specific factor and c refers to intercept parameters.

Table 5. 6: Recovery Results of Standard Errors of the Bifactor Model

Par.	PML				FIML			
	"True"	Est.	Bias	RMSE	"True"	Est.	Bias	RMSE
a_{G1}	0.050	0.073	0.023	0.023	0.031	0.079	0.049	0.049
a_{G2}	0.065	0.077	0.011	0.011	0.101	0.083	-0.018	0.018
a_{G3}	0.037	0.079	0.042	0.042	0.032	0.084	0.052	0.052
a_{G4}	0.017	0.078	0.061	0.061	0.036	0.085	0.049	0.049
a_{G5}	0.053	0.067	0.014	0.014	0.054	0.071	0.016	0.016
a_{G6}	0.030	0.082	0.052	0.052	0.051	0.088	0.038	0.038
a_{G7}	0.031	0.089	0.058	0.058	0.027	0.097	0.070	0.070
a_{S11}	0.106	0.122	0.016	0.016	0.099	0.134	0.035	0.035
a_{S12}	0.100	0.122	0.022	0.022	0.141	0.132	-0.009	0.018

Par.	PML				FIML			
	"True"	Est.	Bias	RMSE	"True"	Est.	Bias	RMSE
a_{S13}	0.076	0.129	0.054	0.054	0.096	0.138	0.042	0.042
a_{S14}	0.058	0.126	0.068	0.068	0.070	0.138	0.068	0.068
a_{S15}	0.035	0.119	0.085	0.085	0.017	0.125	0.108	0.108
a_{S21}	0.094	0.208	0.114	0.114	0.104	0.193	0.089	0.089
c_1	0.010	0.031	0.021	0.021	0.012	0.031	0.019	0.019
c_2	0.021	0.039	0.018	0.018	0.026	0.039	0.013	0.013
c_3	0.026	0.042	0.016	0.016	0.027	0.044	0.017	0.017
c_4	0.028	0.037	0.008	0.008	0.030	0.036	0.006	0.006
c_5	0.023	0.027	0.004	0.004	0.025	0.027	0.002	0.002
c_6	0.020	0.033	0.013	0.013	0.019	0.033	0.013	0.013
c_7	0.038	0.040	0.001	0.001	0.038	0.040	0.002	0.002

Note. a_G stands for the general factor, a_S to specific factor and c refers to intercept parameters. Est.=Estimate.

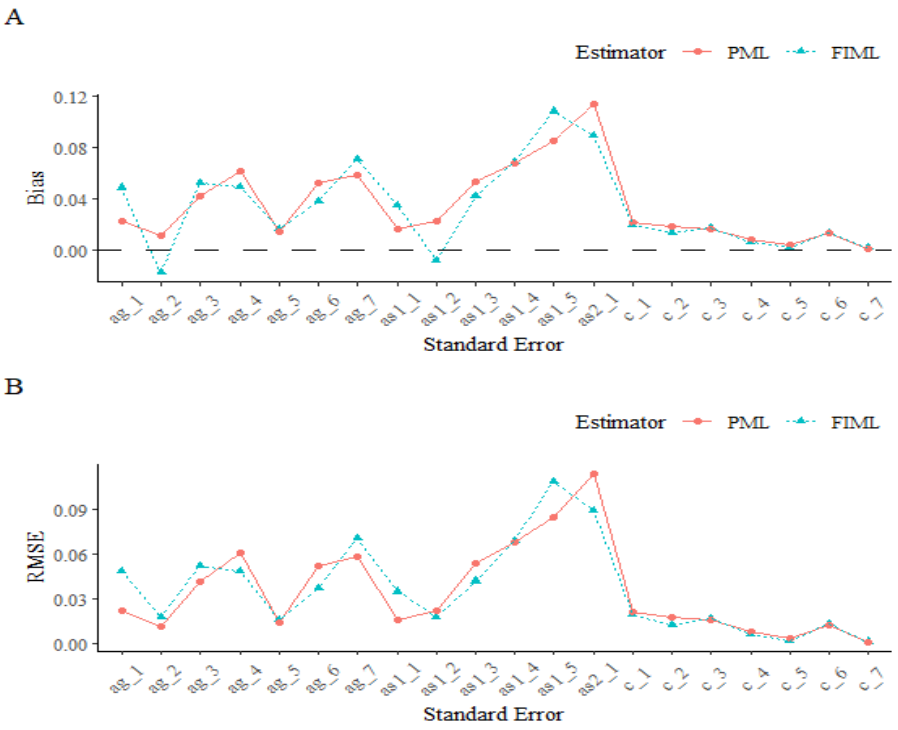


Figure 5. 6: Bias and RMSE of Standard Errors of the Bifactor Model

Note. a_G stands for the general factor, a_S to specific factor and c refers to intercept parameters.

Table 5. 7: Average Bias, RMSE of Estimates and Standard Errors, and Y2/N Values of the Bifactor Model

	Item Parameters						Model Fit
	c		a_G		a_S		Y2/N
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
PML	0.030	0.041	-0.041	0.100	-0.014	0.077	0.00082
FIML	0.026	0.041	-0.070	0.109	-0.010	0.082	0.00112
	Standard Errors						
	c		a_G		a_S		
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
PML	0.012	0.013	0.037	0.042	0.081	0.085	
FIML	0.010	0.013	0.037	0.045	0.069	0.076	

Note. a_G stands for the general factor, a_S to specific factor and c refers to intercept parameters. . Average bias and RMSE was calculated as $Bias(\hat{\theta}) = \frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)}{J}$ and $RMSE(\hat{\theta}) = \sqrt{\frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)^2}{J}}$, which was averaged over replications. J is the number of parameters of type θ.

The model recovery results of the DINA model are organized in Table 5. 8 and Figure 5. 7 for individual parameter estimates and in Table 5. 9 and Figure 5. 8 for the corresponding standard errors. Overall summary results for parameters across replications are in Table 5. 10. The same information for the DINO model is presented in Table 5. 11 and Figure 5. 9 for the estimates, Table 5. 12 and Figure 5. 10 for the standard errors, and Table 5. 13 for overall summaries. As mentioned above, both the DINA and DINO can be formulated as a log-linear cognitive diagnosis model (LCDM). The item parameters were kept in this format of intercept, main, and interaction effects noted using λ s. The item parameter recovery results compared with the "true" parameters showed good fit. These results were consistent when the item parameters of both DCMs estimated using the PML approach were compared to their FIML counterparts. As expected, both DCMs exhibited increased difficulty in estimating the main effect and interaction effect parameters than intercept parameters. In addition, the main effect parameter for Item 4, which according to the Q-matrix in Table 5. 1 was the only simple structure item for Attribute 2, deviated the most from the "true" estimate for both DCMs. The two models also showed the same conclusions regarding model fit where both the DINA and DINO model had smaller $Y2/N$ values across all replications using the PML method when compared to their FIML-based counterparts. In fact, there was evidence from other simulation attempts that the differences in the $Y2/N$ values seemed to grow larger as model misfit grew. The ramifications of this discrepancy will show up in the subsequent simulation studies and will be further discussed then.

Table 5. 8: Recovery Results of Estimates of the DINA Model

Par.	True	PML			FIML		
		Est.	Bias	RMSE	Est.	Bias	RMSE
$\lambda_{1,(1)1}$	2.394	2.404	0.010	0.087	2.402	0.008	0.087
$\lambda_{1,(1)2}$	2.203	2.222	0.019	0.087	2.217	0.014	0.085
$\lambda_{2,(1,2)1}$	2.591	2.599	0.008	0.102	2.584	-0.006	0.088
$\lambda_{1,(2)1}$	1.840	1.823	-0.017	0.131	1.806	-0.033	0.123
$\lambda_{2(2,3)1}$	2.161	2.156	-0.005	0.083	2.149	-0.012	0.081
$\lambda_{1,(3)1}$	2.485	2.518	0.033	0.088	2.508	0.023	0.084
$\lambda_{1,(3)2}$	2.361	2.385	0.024	0.087	2.382	0.021	0.083
λ_{01}	-0.944	-0.953	-0.008	0.083	-0.950	-0.006	0.064
λ_{02}	-1.208	-1.221	-0.013	0.085	-1.217	-0.009	0.067
λ_{03}	-1.266	-1.260	0.005	0.052	-1.263	0.003	0.048
λ_{04}	-0.895	-0.882	0.014	0.112	-0.877	0.018	0.108
λ_{05}	-1.266	-1.244	0.022	0.063	-1.262	0.004	0.052
λ_{06}	-1.099	-1.075	0.023	0.086	-1.109	-0.010	0.060
λ_{07}	-1.153	-1.126	0.026	0.079	-1.161	-0.009	0.058

Note. λ_0 refers to the intercept, λ_1 main effect, and λ_2 interaction parameters. Est.=Estimate.

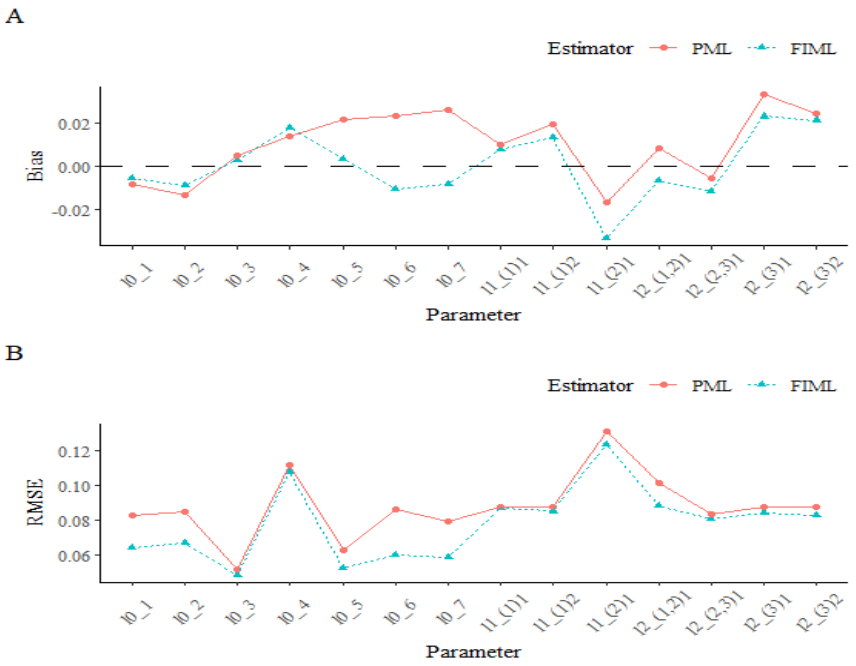


Figure 5. 7: Bias and RMSE of Estimates of the DINA Model

Note. $l = \lambda$. The parameters are indexed in the order of intercept, main effect, and interaction parameters.

Table 5. 9: Recovery Results of Standard Errors of the DINA Model

Par.	PML				FIML			
	"True"	Est.	Bias	RMSE	"True"	Est.	Bias	RMSE
$\lambda_{1,(1)1}$	0.088	0.086	-0.002	0.004	0.087	0.094	0.006	0.026
$\lambda_{1,(1)2}$	0.086	0.080	-0.006	0.006	0.085	0.097	0.012	0.043
$\lambda_{2,(1,2)1}$	0.102	0.089	-0.013	0.014	0.089	0.119	0.031	0.069
$\lambda_{1,(2)1}$	0.131	0.083	-0.048	0.049	0.120	0.155	0.036	0.127
$\lambda_{2(2,3)1}$	0.084	0.076	-0.008	0.008	0.081	0.084	0.003	0.015
$\lambda_{1,(3)1}$	0.082	0.088	0.006	0.007	0.081	0.088	0.006	0.015
$\lambda_{1,(3)2}$	0.085	0.084	-0.001	0.003	0.081	0.085	0.004	0.016
λ_{01}	0.083	0.046	-0.037	0.037	0.065	0.065	0.001	0.031

Par.	PML				FIML			
	"True"	Est.	Bias	RMSE	"True"	Est.	Bias	RMSE
λ_{02}	0.085	0.050	-0.034	0.034	0.067	0.075	0.008	0.042
λ_{03}	0.052	0.041	-0.011	0.011	0.049	0.050	0.001	0.019
λ_{04}	0.112	0.048	-0.064	0.064	0.107	0.110	0.003	0.090
λ_{05}	0.059	0.039	-0.020	0.020	0.052	0.058	0.005	0.024
λ_{06}	0.084	0.048	-0.036	0.036	0.059	0.061	0.002	0.025
λ_{07}	0.076	0.048	-0.027	0.028	0.058	0.061	0.003	0.025

Note. λ_0 refers to the intercept, λ_1 main effect, and λ_2 interaction parameters. Est.=Estimate.

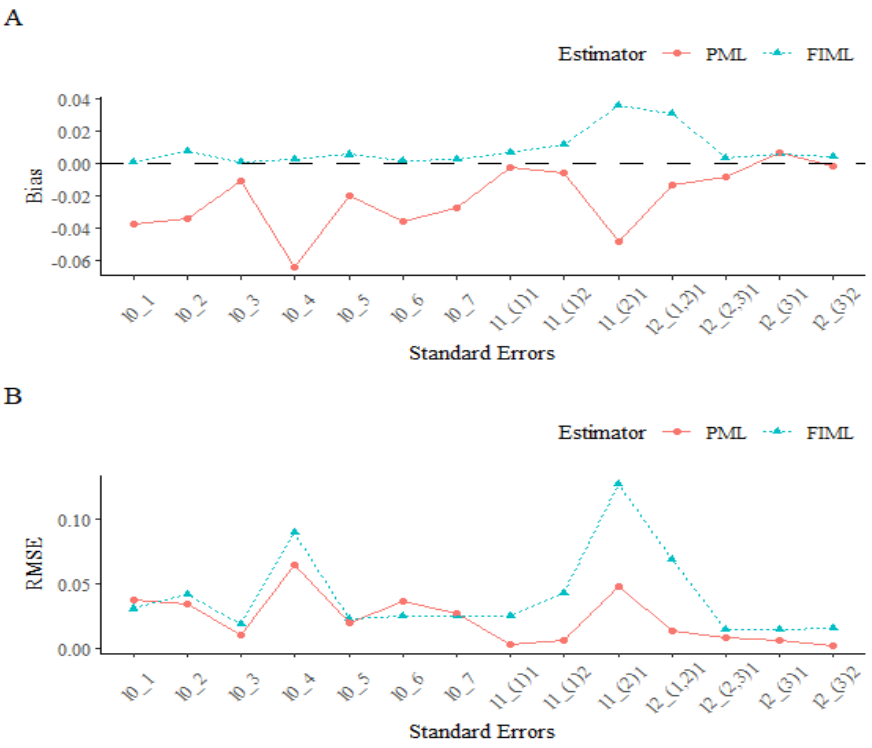


Figure 5. 8: Bias and RMSE of Standard Errors of the DINA Model

Note. $l = \lambda$. The parameters are indexed in the order of intercept, main effect, and interaction parameters.

Table 5. 10: Average Bias, RMSE of Estimates and Standard Errors, and Y2/N Values of the DINA Model

	Item Parameters						Model Fit
	λ_0		λ_1		λ_2		Y2/N
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
PML	0.010	0.077	0.014	0.093	0.002	0.010	0.00078
FIML	-0.001	0.065	0.006	0.088	-0.009	0.072	0.00105
	Standard Errors						
	λ_0		λ_1		λ_2		
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
PML	-0.097	0.051	-0.028	0.041	-0.130	0.135	
FIML	-0.062	0.247	-0.028	0.214	-0.071	0.233	

Note. λ_0 refers to the intercept, λ_1 main effect, and λ_2 interaction parameters. . Average bias and RMSE was calculated as $Bias(\hat{\theta}) = \frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)}{J}$ and $RMSE(\hat{\theta}) = \sqrt{\frac{\sum_{j=1}^{J}(\hat{\theta}_j - \theta_j)^2}{J}}$, which was averaged over replications. J is the number of parameters of type θ.

Table 5. 11: Recovery Results of Estimates of the DINO Model

Par.	True	PML			FIML		
		Est.	Bias	RMSE	Est.	Bias	RMSE
$\lambda_{1,(1)1}$	2.234	2.251	0.017	0.059	2.258	0.025	0.064
$\lambda_{1,(1)2}$	2.773	2.778	0.005	0.088	2.767	-0.006	0.085
$\lambda_{2,(1,2)1}$	2.234	2.227	-0.007	0.091	2.216	-0.018	0.083
$\lambda_{1,(2)1}$	3.584	3.694	0.110	0.298	3.651	0.068	0.236
$\lambda_{2(2,3)1}$	2.234	2.225	-0.009	0.074	2.222	-0.012	0.066
$\lambda_{1,(3)1}$	3.045	3.061	0.016	0.123	3.038	-0.007	0.112
$\lambda_{1,(3)2}$	1.695	1.691	-0.004	0.064	1.692	-0.003	0.064
λ_{01}	-0.847	-0.862	-0.015	0.062	-0.859	-0.011	0.047
λ_{02}	-1.386	-1.399	-0.013	0.078	-1.385	0.001	0.060

Par.	True	PML			FIML		
		Est.	Bias	RMSE	Est.	Bias	RMSE
λ_{03}	-1.386	-1.382	0.004	0.085	-1.368	0.018	0.071
λ_{04}	-2.197	-2.217	-0.020	0.134	-2.191	0.006	0.102
λ_{05}	-0.847	-0.847	0.000	0.060	-0.834	0.013	0.059
λ_{06}	-2.197	-2.235	-0.038	0.146	-2.193	0.004	0.107
λ_{07}	-0.847	-0.854	-0.007	0.055	-0.842	0.005	0.047

Note. λ_0 refers to the intercept, λ_1 main effect, and λ_2 interaction parameters. Est.=Estimate.

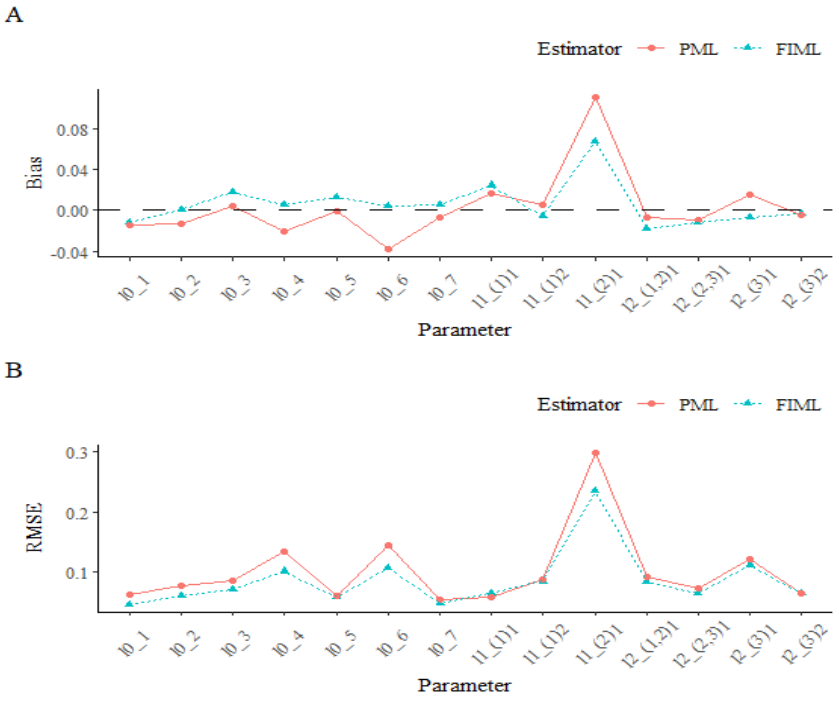


Figure 5. 9: Bias and RMSE of Estimates of the DINO Model

Note. $l = \lambda$. The parameters are indexed in the order of intercept, main effect, and interaction parameters.

Table 5. 12: Recovery Results of Standard Errors of the DINO Model

Par.	PML				FIML			
	"True"	Est.	Bias	RMSE	"True"	Est.	Bias	RMSE
$\lambda_{1,(1)1}$	0.057	0.084	0.027	0.034	0.060	0.075	0.015	0.015
$\lambda_{1,(1)2}$	0.089	0.102	0.013	0.048	0.086	0.081	-0.005	0.014
$\lambda_{2,(1,2)1}$	0.092	0.083	-0.010	0.023	0.082	0.097	0.016	0.030
$\lambda_{1,(2)1}$	0.279	0.168	-0.112	0.142	0.228	0.079	-0.150	0.150
$\lambda_{2(2,3)1}$	0.074	0.075	0.001	0.011	0.066	0.076	0.011	0.011
$\lambda_{1,(3)1}$	0.123	0.162	0.039	0.151	0.113	0.072	-0.041	0.042
$\lambda_{1,(3)2}$	0.065	0.074	0.009	0.021	0.065	0.064	0.000	0.001
λ_{01}	0.061	0.045	-0.016	0.018	0.046	0.042	-0.004	0.004
λ_{02}	0.078	0.060	-0.018	0.026	0.061	0.050	-0.011	0.019
λ_{03}	0.086	0.067	-0.019	0.032	0.069	0.086	0.017	0.031
λ_{04}	0.134	0.118	-0.016	0.074	0.102	0.054	-0.049	0.050
λ_{05}	0.061	0.052	-0.008	0.012	0.059	0.058	0.000	0.002
λ_{06}	0.142	0.126	-0.017	0.115	0.108	0.054	-0.053	0.054
λ_{07}	0.055	0.044	-0.011	0.014	0.047	0.041	-0.006	0.006

Note. λ_0 refers to the intercept, λ_1 main effect, and λ_2 interaction parameters. Est.=Estimate.

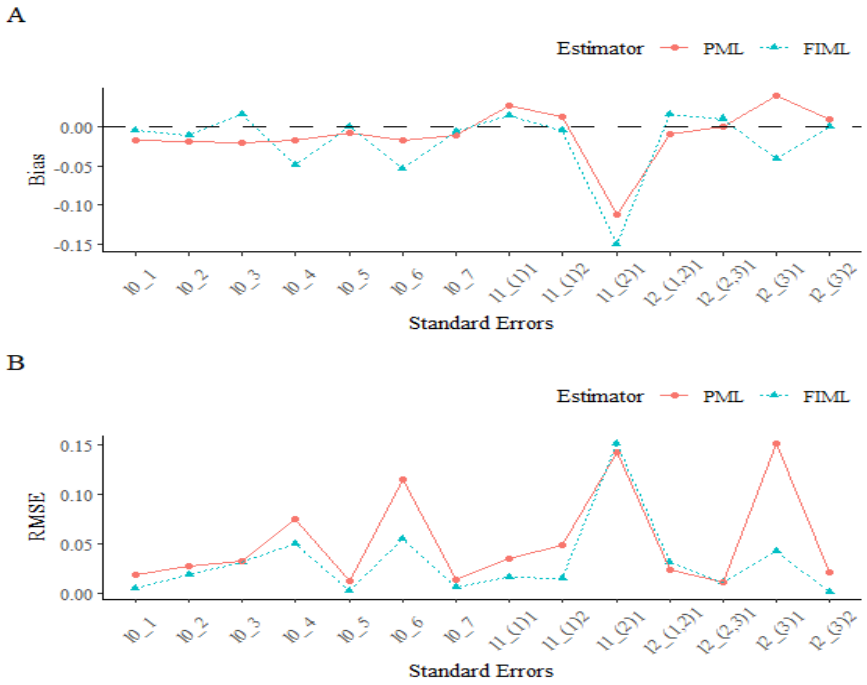


Figure 5. 10: Bias and RMSE of Standard Errors of the DINO Model

Note. $l = \lambda$. The parameters are indexed in the order of intercept, main effect, and interaction parameters.

Table 5. 13: Average Bias, RMSE of Estimates and Standard Errors, and Y2/N Values of the DINO Model

	Item Parameters						Model Fit
	λ_0		λ_1		λ_2		Y2/N
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
PML	-0.012	0.090	0.029	0.137	-0.008	0.071	0.00079
FIML	0.005	0.070	0.015	0.114	-0.015	0.066	0.00109
	Standard Errors						
	λ_0		λ_1		λ_2		
	Bias	RMSE	Bias	RMSE	Bias	RMSE	
PML	-0.015	0.041	-0.005	0.078	-0.004	0.041	
FIML	-0.015	0.030	-0.036	0.070	0.013	0.014	

Note. λ_0 refers to the intercept, λ_1 main effect, and λ_2 interaction parameters. . Average bias and RMSE was calculated as $Bias(\hat{\theta}) = \frac{\Sigma_{j=1}^{J}(\hat{\theta}_j - \theta_j)}{J}$ and $RMSE(\hat{\theta}) = \sqrt{\frac{\Sigma_{j=1}^{J}(\hat{\theta}_j - \theta_j)^2}{J}}$, which was averaged over replications. J is the number of parameters of type θ.

5.3.3. Summary

The results of Simulation Study 1 for item parameter recovery indicated adequate parameter recovery for four out of the five models in Bonifay and Cai (2017) under PML estimation. The exception was the unidimensional 3PL model. Moreover, the parameter recovery results were comparable with those from a conventionally used FIML method; albeit FIML showed slightly superior results in terms of item parameter recovery. This is as expected as FIML methods use all the available information from the data while PML uses only bivariate information. SE estimates were also similar across the two methods of FIML and PML estimation even though PML required the Godambe information matrix for SE calculation. Contrarily, the results showed differences in model fit that was evaluated using Y2/*N* statistic. There was a tendency of smaller Y2/*N* values in PML estimation than FIML estimation for particularly the DCMs.

5.4 Simulation Study 2: Replication of Bonifay and Cai's (2017) Study using the Simplex Sampling Method

Bonifay and Cai (2017) used flexMIRT (Cai, 2021) to fit the IRT models in Figure 5. 1 using a FIML method via the EM algorithm. In this dissertation, PML estimation is the target estimation method. The goal of this simulation study was to determine whether Bonifay and Cai's (2017) major findings are replicable with the PML approach to be used throughout the subsequent studies so that any comparisons of results are meaningful.

5.4.1 Simulation Study Setup

Data Generation

To compare the results of Bonifay and Cai's (2017) study with those based on a PML approach, datasets for seven dichotomous items were sampled using the simplex sampling method presented in Bonifay and Cai (2017). This generated full item response patterns, which flexMIRT expects. Following Bonifay and Cai (2017), a total of 1000 random datasets for seven dichotomously scored items were generated with $N = 10000$. Similar to the previous simulation study on the performance of the proposed PML approach, the data were collapsed to the corresponding bivariate margins for use with PML estimation.

Analysis Setup

The EFA, bifactor, DINA, and DINO models were each fit to the 1,000 random datasets using both flexMIRT and the PML estimation method. For flexMIRT, estimation specifications were identical to that of Bonifay and Cai (2017). They used a more relaxed convergence tolerance of 0.001 for maximum parameter change in consecutive EM cycles as well as an increase in the maximum number of EM cycles (i.e., 20,000) to increase the convergence rate. However, Bonifay and Cai (2017) concluded that Preacher's (2016) argument that even non-converged estimates after many E-step iterations (e.g., 10,000, 20,000) can be considered acceptable estimates was valid following indistinguishability of the converged and non-converged results in their study. Also, Aytürk Ergin (2020) echoed this sentiment in that convergence did not affect their results in FP investigation either. Therefore, for PML estimation, considering that it generally takes more iterations and a lower tolerance to get the same results as the FIML method, we kept to the initial setting of tolerance of 10^{-6} but increased the number of iterations to 25,000.

The Y2/*N* index resulting from fitting each model was recorded and analyzed for all 1,000 replications for both estimation methods, regardless of non-convergence. The results were organized using cumulative percentage curves of Y/*N* for each model. Also, euler diagrams using the *eulerr* package (Larsson, 2021) in R were utilized to examine the approximate degree of overlap among models with respect to the area in the data space they covered based on Y/*N*.

5.4.2 Results: Generated Data and 2×2 Tables

The simplex sampling method for J binary items used in this simulation gives us 2^J multinomial probabilities, which can be collapsed down to their bivariate margins for PML estimation. However, it is both theoretically and empirically evident that the data from these 2×2 tables do not match that of generating 2×2 tables directly using the SIS approach.

The simplex sampling method for J binary items is the same as sampling from a $(2^J - 1)$ probability simplex (Rubin, 1981). Same as before, the minus one comes from the constraint that the sum of all probabilities must be one. We already know that sampling from a $(2^J - 1)$ probability simplex is also the same as sampling from a Dirichlet distribution with 2^J variables where all the alphas are set to 1 or $Dir(\alpha_1 = 1, \cdots, \alpha_{2^J} = 1)$. Applied to the case of Simulation Study 2 with seven items, this means the values of the simulated datasets were essentially drawn from a $Dir(\alpha_1 = 1, \cdots, \alpha_{128} = 1)$. Let's apply the aggregation property repeatedly until we arrive at the bivariate margins (i.e., the cells of a 2×2 table). It becomes that the 2×2 tables obtained from collapsing the multinomial probabilities follow a $Dir(\alpha_{00} = 32, \alpha_{01} = 32, \alpha_{10} = 32, \alpha_{11} = 32)$, which differs significantly from a $Dir(\alpha_{00} = 1, \alpha_{01} = 1, \alpha_{10} = 1, \alpha_{11} = 1)$ that the 2×2 tables using the SIS approach follow.

The α parameters of a Dirichlet distribution determine both the concentration and distribution of the distribution. The higher the value of α, the greater the weight and amount of the total mass assigned to that parameter given to the corresponding outcome. Equal αs give a symmetric or even distribution regardless of their size. α less than one (i.e., $\alpha < 1$) push the outcomes toward the extremes and αs greater than one (i.e, $\alpha > 1$) pull the outcome toward some central value, the force of which increases depending on α. If $\alpha_1 = \cdots = \alpha_k = 1$ for k variables in a Dirichlet distribution, then the points are uniformly distributed (Lin, 2016). A $Dir(\alpha_{00} = 32, \alpha_{01} = 32, \alpha_{10} = 32, \alpha_{11} = 32)$ should generate values that are very concentrated at the center of a $(4-1)$- dimensional simplex, just like Figure 5. 11.

In short, when $J > 2$, the simplex sampling method will not lead to bivariate margins that are uniformly distributed. Instead, they will be concentrated toward the center of the $(2^J - 1)$- dimensional simplex. In turn, this means that the data generated using the SIS method is more random than when using a simplex sampling approach. The values from the latter may be classified to what Preacher (2006) and Roberts and Pashler (2000) call "plausible" values of a target data space rather than "possible" data.

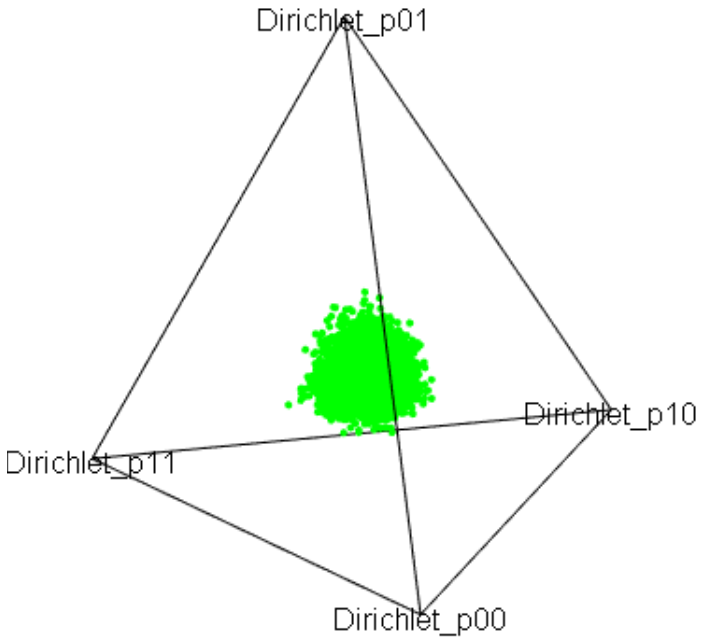


Figure 5. 11: Plot of Bivariate Margins from doing Simplex Sampling with Seven Items

5.4.3 FP Results for FIML Estimation

Table 5. 14 displays the descriptive statistics of the Y2/*N* statistic for the four models. Comparing means shows that on average, the EFA and bifactor models produced Y2/*N* values of 0.05 or lower. The DINA and DINO models tended to have Y2/*N* values of 0.10. The results also show The EFA and bifactor model generally had lower Y2/*N* values compared to the DINA and DINO models.

Table 5. 14: Descriptive Statistics of Y2/*N* for Simplex Method × FIML Estimation

Model	M	SD	Min	Max
EFA	0.037	0.018	0.003	0.127
Bifactor	0.047	0.022	0.005	0.148
DINA	0.103	0.039	0.024	0.261
DINO	0.103	0.039	0.033	0.264

Note. M= mean, SD=standard deviation.

Figure 5. 12 shows the empirical cumulative distribution function (CDF) plot of the Y2/*N* statistic for the four models. The cumulative percentages of datasets across replications that achieve a particular value of Y2/*N* for each model were depicted in the plot. The percentage of data that each model fit according to a reference Y2/*N* value can be found by investigating the vertical distance between models in the plot. For example, using the cutoff in Bonifay and Cai (2017) of $Y2/N \leq 0.05$, 77.9% of the datasets were found to fit under the EFA model, followed by 60.1% for the bifactor model. Less than 5% of datasets showed good fit to the DINA and DINO model when $Y2/N \leq 0.05$. Looking vertically across the values of Y2/*N* shows that the EFA model had the highest percentage of fitting datasets followed by the bifactor model and, lastly, DINA and DINO models.

One can also find the corresponding Y2/*N* value to a benchmark percentage based on the horizontal discrepancy of the CDF plot. Looking horizontally across percentages, we can see that the EFA model had the lowest Y2/*N* value for any benchmark percentage with increasing values in the order of the bifactor and then DINA and DINO models. The large gap in the CDF curves between the EFA and bifactor models and the DINA and DINO models points to significant differences in Y2/*N*s values between of EFA and

bifactor models compared to their DCM counterparts to fit the same amount of data. There is also a distinction between the EFA and bifactor to a lesser extent. On the other hand, the two DCMs had nearly identical curves.

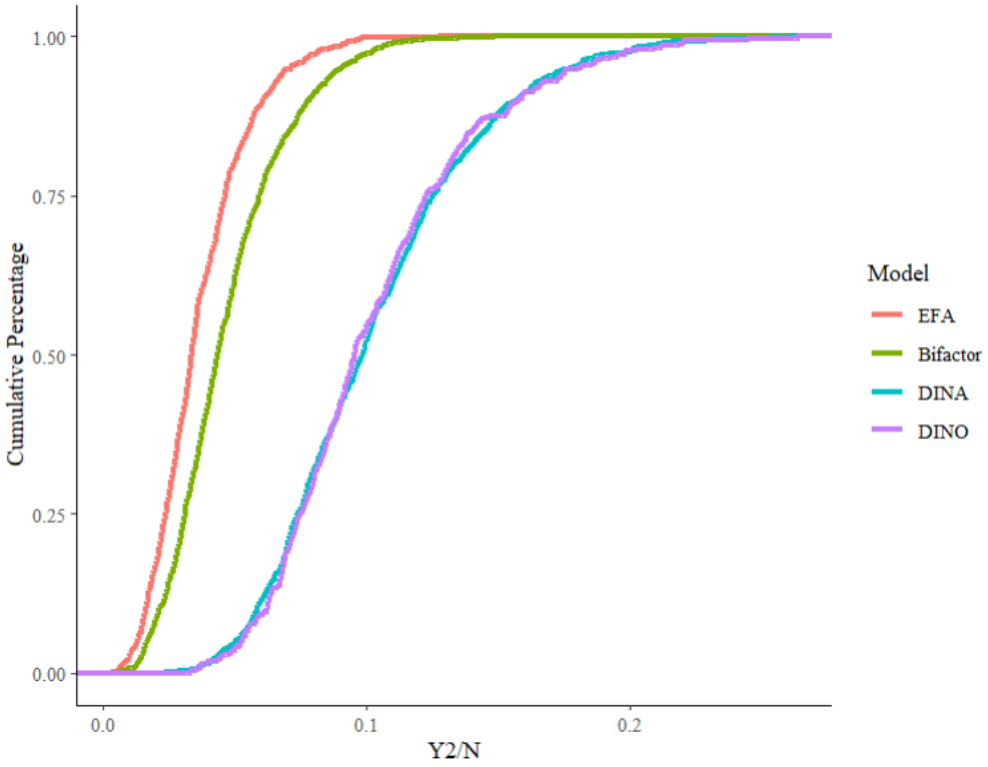
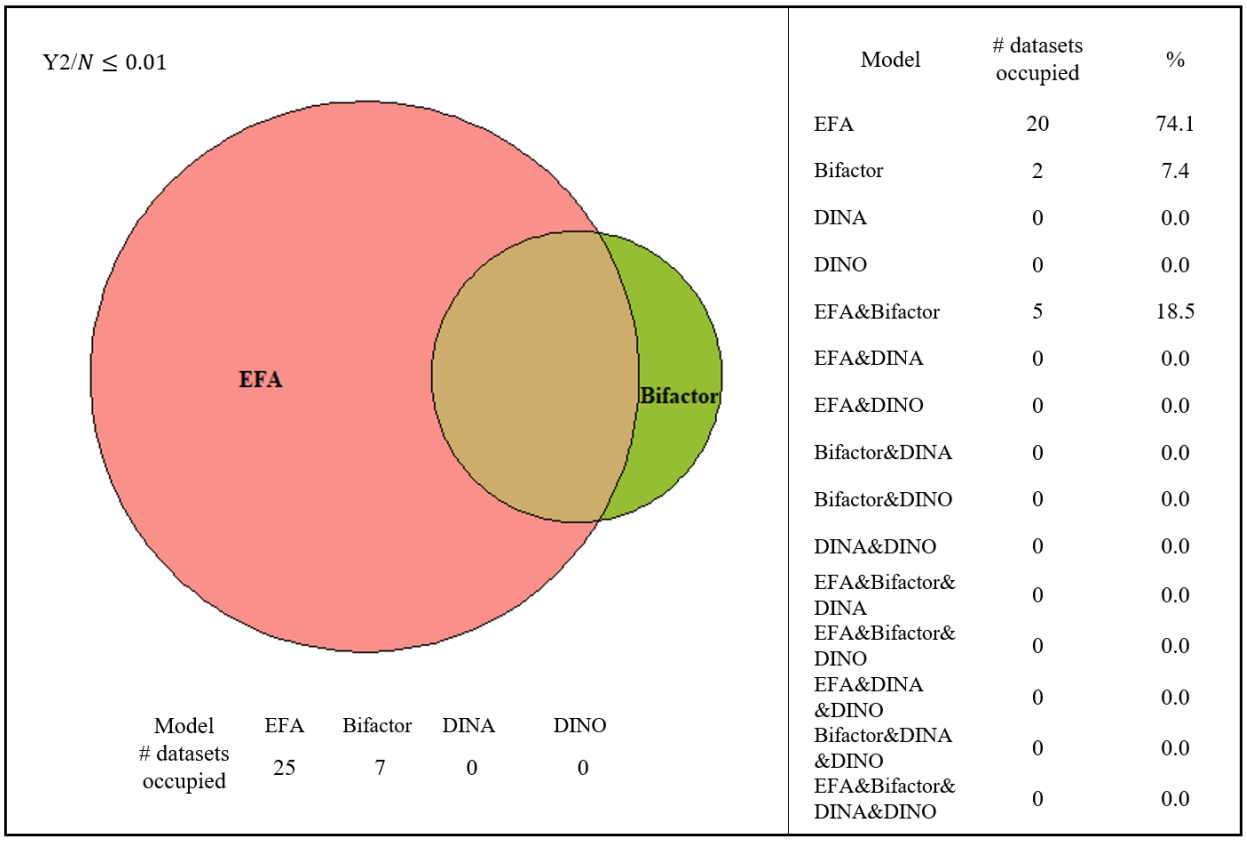


Figure 5. 12: Cumulative Percentage Distributions of the Y2/*N* statistic Simplex Method × FIML Estimation

Euler diagrams can be used to approximately examine the overlap among models in addition to the areas of the data space that models show fit at specific cut-points. The overlap in areas between models indicates datasets for which the overlapping models all satisfy a cut-point. Figure 5. 13, Figure 5. 14, and Figure 5. 15 depict the regions in the data space captured by the models at increasingly large cutoffs for the Y2/*N* statistic of 0.01, 0.03, and 0.05 (Bonifay & Cai, 2017).

Under the most conservative cutoff of $Y2/N \leq 0.01$ (Figure 5. 13), only the EFA and bifactor model were present with EFA occupying a larger area among the two. Also, even though the larger portion of the bifactor model area overlapped with the EFA model (71.4% of bifactor models overlapped with EFA model), there was still a region where only the bifactor model fit (2% of the complete data space and 7.4% of all the fitted region).

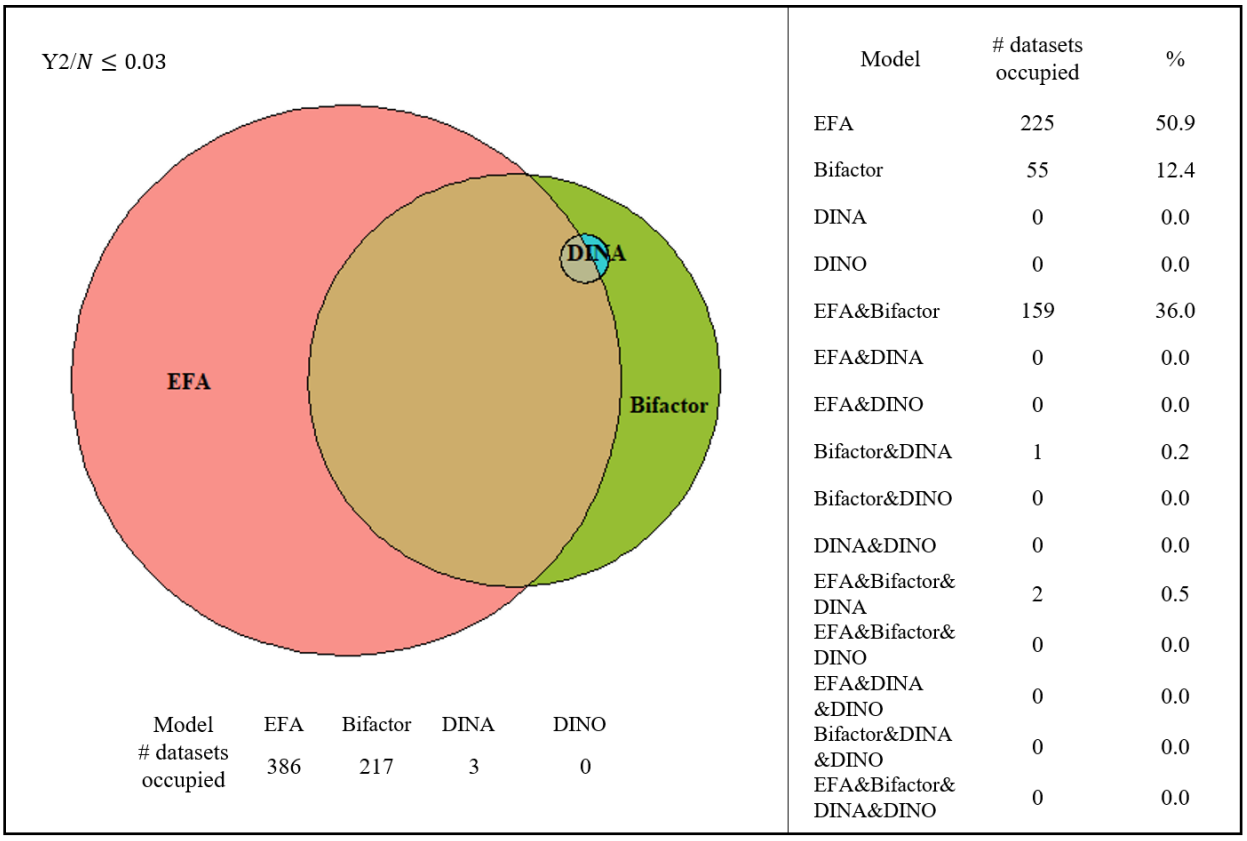


Model	# datasets occupied	%
EFA	20	74.1
Bifactor	2	7.4
DINA	0	0.0
DINO	0	0.0
EFA&Bifactor	5	18.5
EFA&DINA	0	0.0
EFA&DINO	0	0.0
Bifactor&DINA	0	0.0
Bifactor&DINO	0	0.0
DINA&DINO	0	0.0
EFA&Bifactor&DINA	0	0.0
EFA&Bifactor&DINO	0	0.0
EFA&DINA&DINO	0	0.0
Bifactor&DINA&DINO	0	0.0
EFA&Bifactor&DINA&DINO	0	0.0

Figure 5. 13: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 0.01$ Simplex Sampling Method × FIML Estimation

Figure 5. 14 shows that with $Y2/N \leq 0.03$, although the EFA and bifactor still dominated the regions of good fit, the DINA model was also introduced. No DINO models were included. The degree of overlap between the EFA and bifactor models increased as well, from 18.5% to 35.7%, but the region where only the bifactor model showed good fit

also increased (5.5% of total datasets and 12.4% of fitted datasets). The DINA datasets were subsumed completely with the bifactor model, among which one did not overlap with the EFA model.

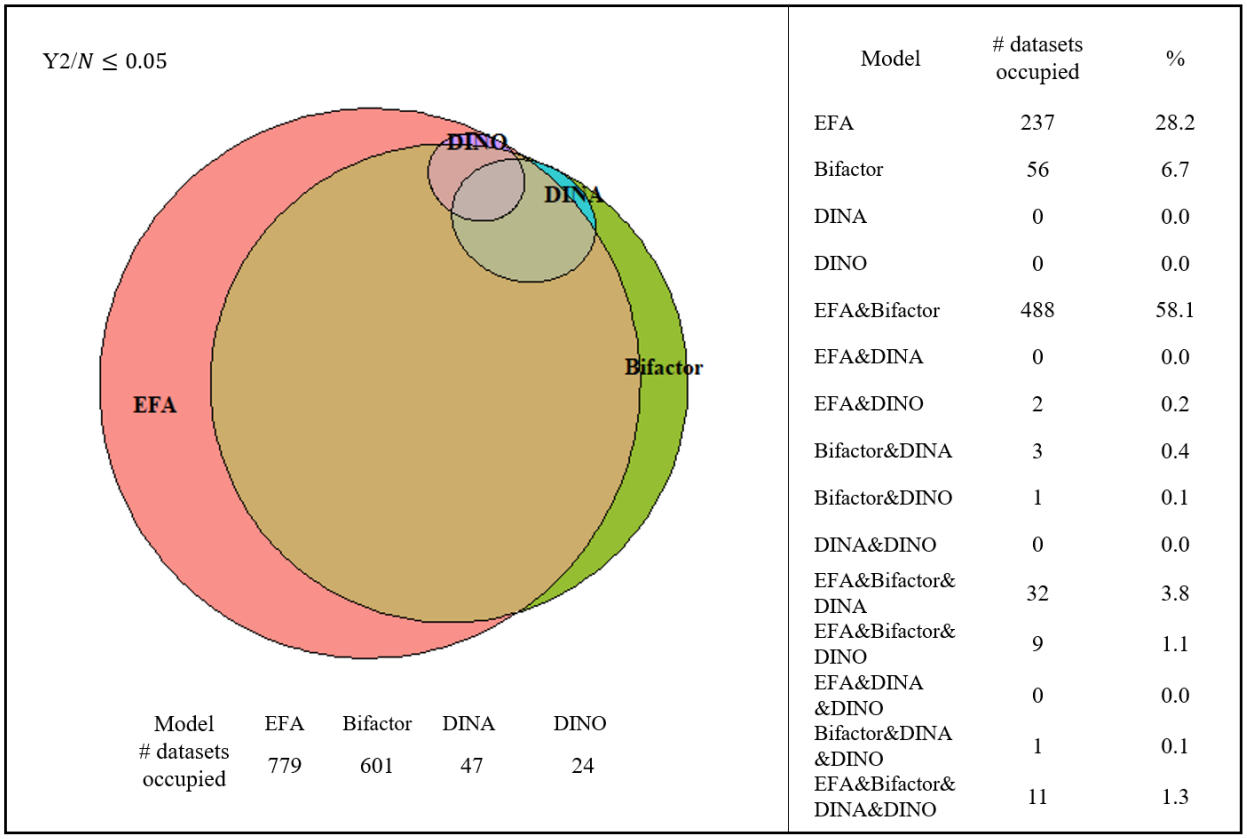


Model	# datasets occupied	%
EFA	225	50.9
Bifactor	55	12.4
DINA	0	0.0
DINO	0	0.0
EFA&Bifactor	159	36.0
EFA&DINA	0	0.0
EFA&DINO	0	0.0
Bifactor&DINA	1	0.2
Bifactor&DINO	0	0.0
DINA&DINO	0	0.0
EFA&Bifactor&DINA	2	0.5
EFA&Bifactor&DINO	0	0.0
EFA&DINA&DINO	0	0.0
Bifactor&DINA&DINO	0	0.0
EFA&Bifactor&DINA&DINO	0	0.0

Figure 5. 14: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 0.03$ Simplex Sampling Method × FIML Estimation

Using the most liberal cutoff of the three (i.e., $Y2/N < 0.05$; Figure 5. 15), all models were introduced. The areas occupied by the DCMs were 5.9% of the complete data space and 9.8% of the fitted data space. The EFA model subsumed most of the area covered by the bifactor model with their degree of overlap 90.2%. The most notable difference was the disentanglement in fit for the DCMs. Not only did the DINA model seem to fit more datasets than the DINO, but they also did not overlap perfectly with

each other (degree of overlap was only 20%). To add, not all DCMs could be subsumed by just one of the bifactor or EFA models.



Model	# datasets occupied	%
EFA	237	28.2
Bifactor	56	6.7
DINA	0	0.0
DINO	0	0.0
EFA&Bifactor	488	58.1
EFA&DINA	0	0.0
EFA&DINO	2	0.2
Bifactor&DINA	3	0.4
Bifactor&DINO	1	0.1
DINA&DINO	0	0.0
EFA&Bifactor&DINA	32	3.8
EFA&Bifactor&DINO	9	1.1
EFA&DINA&DINO	0	0.0
Bifactor&DINA&DINO	1	0.1
EFA&Bifactor&DINA&DINO	11	1.3

Figure 5. 15: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 0.05$ Simplex Sampling Method × FIML Estimation

Summary

Overall, the Y2/*N* results discussed above indicated that the EFA and bifactor models possess high propensities to fit any possible data. Thus, these models displayed high FPs, especially in comparison to the two DCMs with far lower FPs. While the DCMs were indeed comparable to each other, they could occupy different regions of fit, with DINA having slightly better fit than the DINO models overall. As expected, the results conform to those of Bonifay and Cai (2017) and Aytürk Egrin (2020) who replicated the former's results using different statistical packages in R and other fit indices.

5.4.4 FP Results for PML Estimation

The goal of Simulation Study 2 was to test whether Bonifay and Cai's (2017) findings can be replicated for the proposed PML estimation using the same data. The same types of tables and figures used for FIML-based FP results were used.

The descriptive statistics are summarized in Table 5. 15, for which differences in the magnitudes of Y2/*N* values are the most noticeable. Overall, the values were magnified by over ten-told. Nonetheless, the results showed the same pattern as FIML results in Table 5. 14 where the EFA and bifactor model generally had lower Y2/*N* values compared to the DINA and DINO models. On average, the EFA and bifactor models produced Y2/*N* values of 0.6 or lower. The DINA and DINO models tended to have Y2/*N* values of 0.7.

Table 5. 15: Descriptive Statistics of Y2/*N* for Simplex Sampling Method × PML Estimation

Model	M	SD	Min	Max
EFA	0.562	0.342	0.055	3.625
Bifactor	0.596	0.342	0.055	2.234
DINA	0.714	0.398	0.085	3.229
DINO	0.714	0.397	0.085	3.228

Note. M= mean, SD=standard deviation.

The CDF plot for the models (Figure 5. 16) displays similar patterns to FIML results (Figure 5. 12) as well. The EFA model had the highest percentage of fitting datasets for any Y2/N statistic followed by the bifactor model and, at last the DINA and DINO models (which were identical). In sum, the general patterns of FP were not affected

despite the change in magnitude of Y2/*N* values. However, a discernable difference is that the separation between EFA and bifactor models with the DCMs is not nearly as prominent, although we still observed larger differences between the two sets of models (i.e., EFA and bifactor versus DINA and DINO models) than within. When considering the vertical discrepancy between the curves, the differences in the cumulative percentage of datasets at or below a cutoff Y2/*N* value are not as stark as before. Or from the perspective of looking at the horizontal discrepancy between the curves, the differences in Y2/*N* values to reach a certain percentage of fitting models becomes smaller. The gap between the EFA and bifactor model was smaller as well, implying that they were likely to fit more similarly than before. Such results indicate larger regions of coverage for greater overlap between all models but especially for DINA and DINO models than in FIML estimation. The CDF curves for the DINO and DINA models were still indistinguishable, which might be natural given the decrease in the space between them and the EFA and bifactor models. Another characteristic of the CDF plot was the increase in the space between the CDF curves of the EFA and bifactor models with the DCMs as Y2/*N* values grew larger. This implied larger discrepancies in the Y2/*N* values for as cumulative percentages increased.

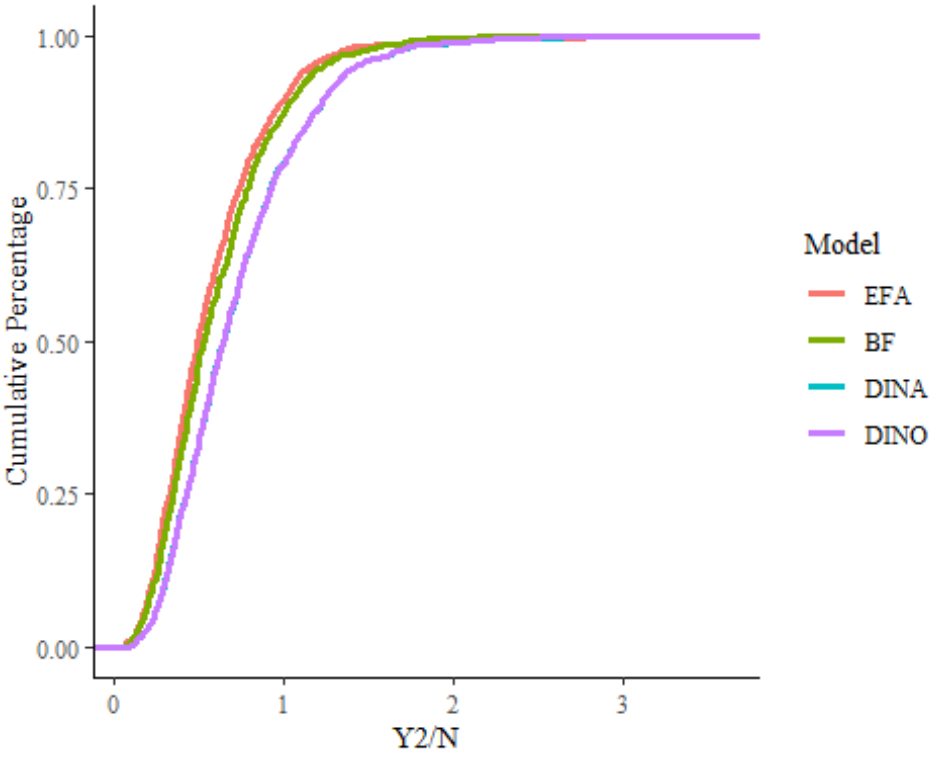
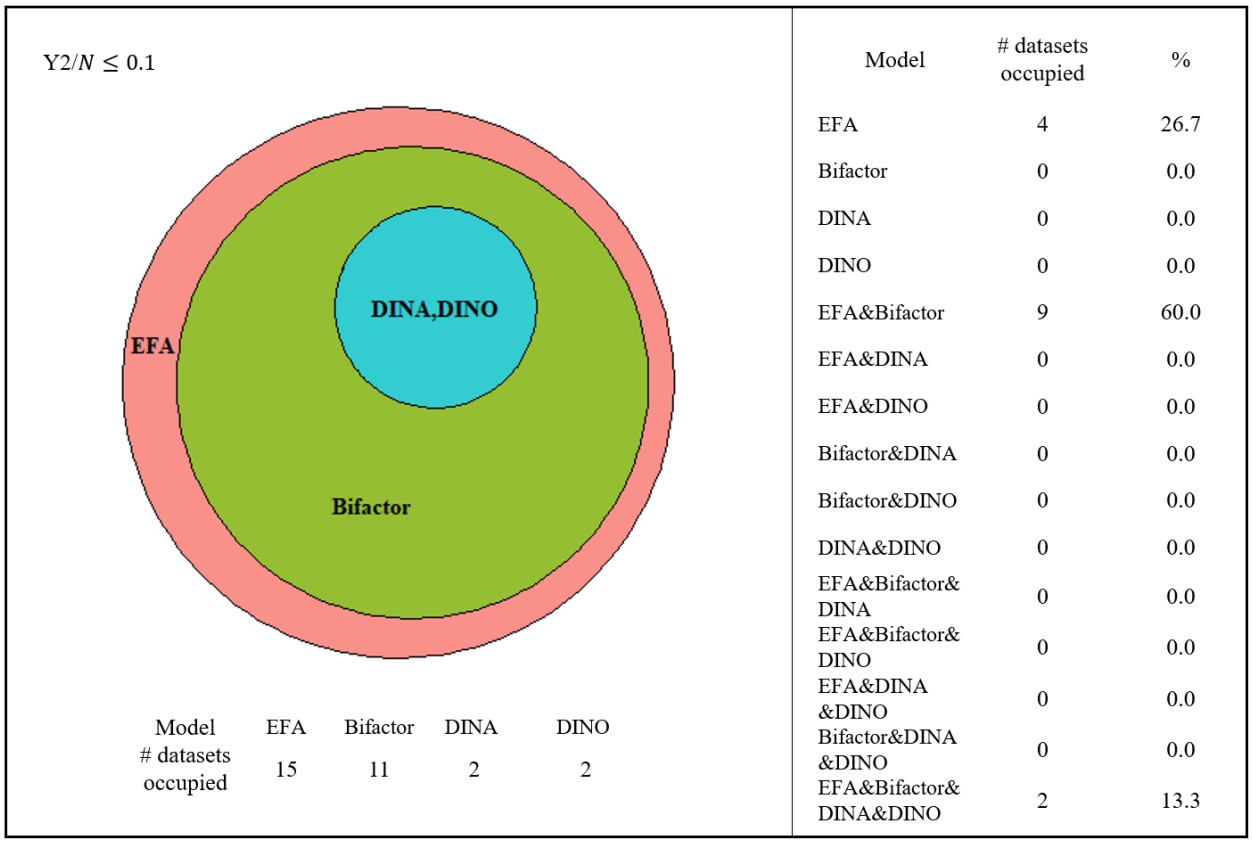


Figure 5. 16: Cumulative Percentage Distributions of the Y2/N statistic for Simplex Sampling Method × PML Estimation

The suspected increase in overlap along with regions of fit was corroborated in Figure 5. 17, Figure 5. 18, and Figure 5. 19 using Euler diagrams. Scale differences with the FIML results meant that the same cutoffs of Y2/N could not be applied. Instead, arbitrary cut points were chosen that best-provided insight about (which provided the best insights into) the four models. The specific cut-points were $Y2/N \leq 0.1$, $Y2/N \leq 0.3$ and $Y2/N \leq 0.5$.

Even under the most stringent cutoff of $Y2/N \leq 0.1$ (Figure 5. 17), the bifactor model covered much of the area for the EFA. Also, areas where the DINA and DINO models fit were not neglectable in comparison to their EFA and bifactor counterparts. While the DCMs consisted of 0.2% of the complete data space, the DCMs accounted for 13.3% of the datasets that fit. The EFA model entirely subsumed the bifactor model,

which was equal to 73.3% of the fitted data space (1.1 of the total data space), meaning that they would give the same result about fit 73.3% of the time. In turn, the bifactor model entirely subsumed both DCMs. No differences in fit were detected between the DINA and DINO model.

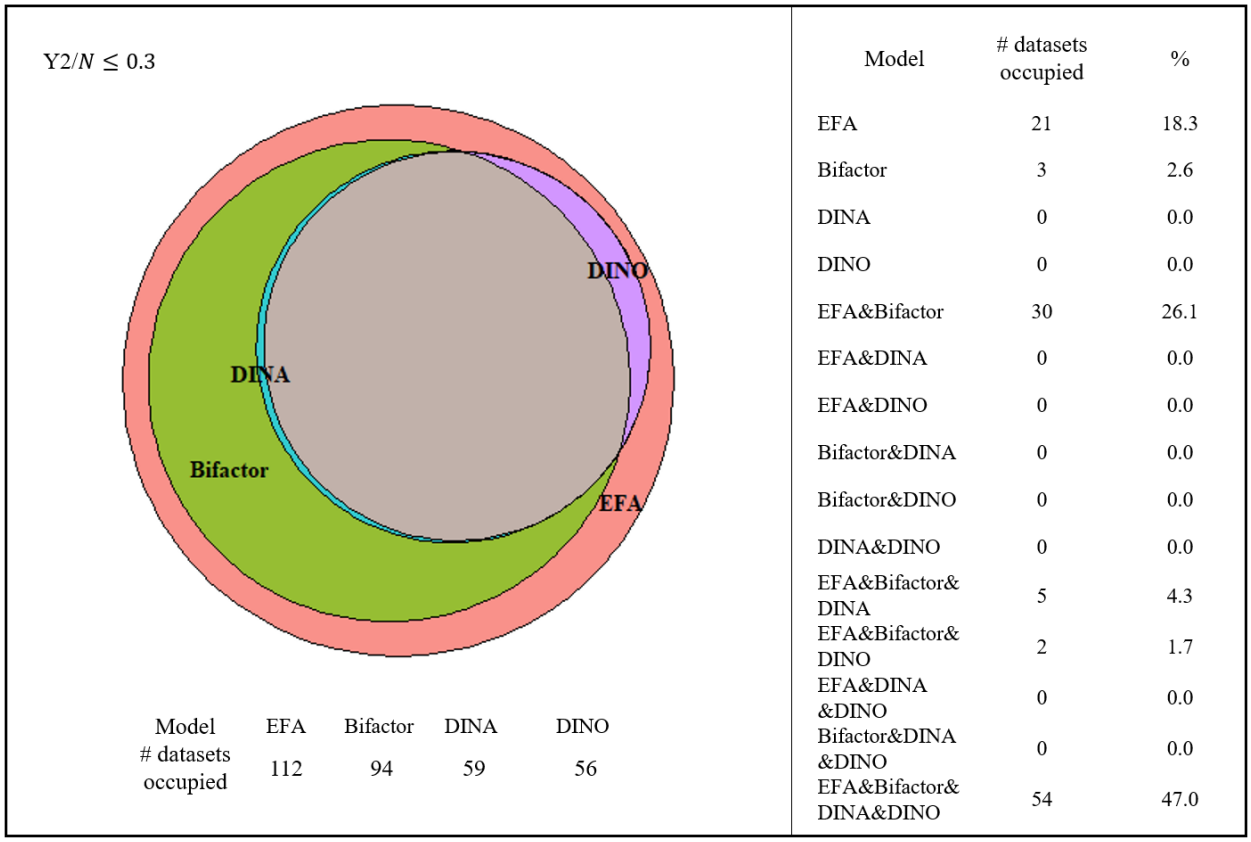


Model	# datasets occupied	%
EFA	4	26.7
Bifactor	0	0.0
DINA	0	0.0
DINO	0	0.0
EFA&Bifactor	9	60.0
EFA&DINA	0	0.0
EFA&DINO	0	0.0
Bifactor&DINA	0	0.0
Bifactor&DINO	0	0.0
DINA&DINO	0	0.0
EFA&Bifactor&DINA	0	0.0
EFA&Bifactor&DINO	0	0.0
EFA&DINA&DINO	0	0.0
Bifactor&DINA&DINO	0	0.0
EFA&Bifactor&DINA&DINO	2	13.3

Figure 5. 17: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 0.1$ for Simplex Sampling Method × PML Estimation

The Euler diagram for $Y2/N \leq 0.3$ is given in Figure 5. 18 and showed a relative increase in growth that was larger for the DINA and DINO models than the EFA and bifactor models. The area covered by both the EFA and bifactor models accounted for 17.3% of the total data space and 76.5% of the fitted data space, similar to Figure 5.10. On the other hand, DCMs accounted for 11.5% of the complete data space and 50.8% of all fitted datasets. Also, while the EFA still completely subsumed the bifactor and both

DCMs the bifactor model only subsumed the DINA model and not the DINO model (the percentage of non-overlap between bifactor and DINA model was 6.1%). The difference between the two, although minor at 2.6%, provides support that their fit to specific datasets may give different conclusions even though the overall percentages of fitted datasets are roughly the same.

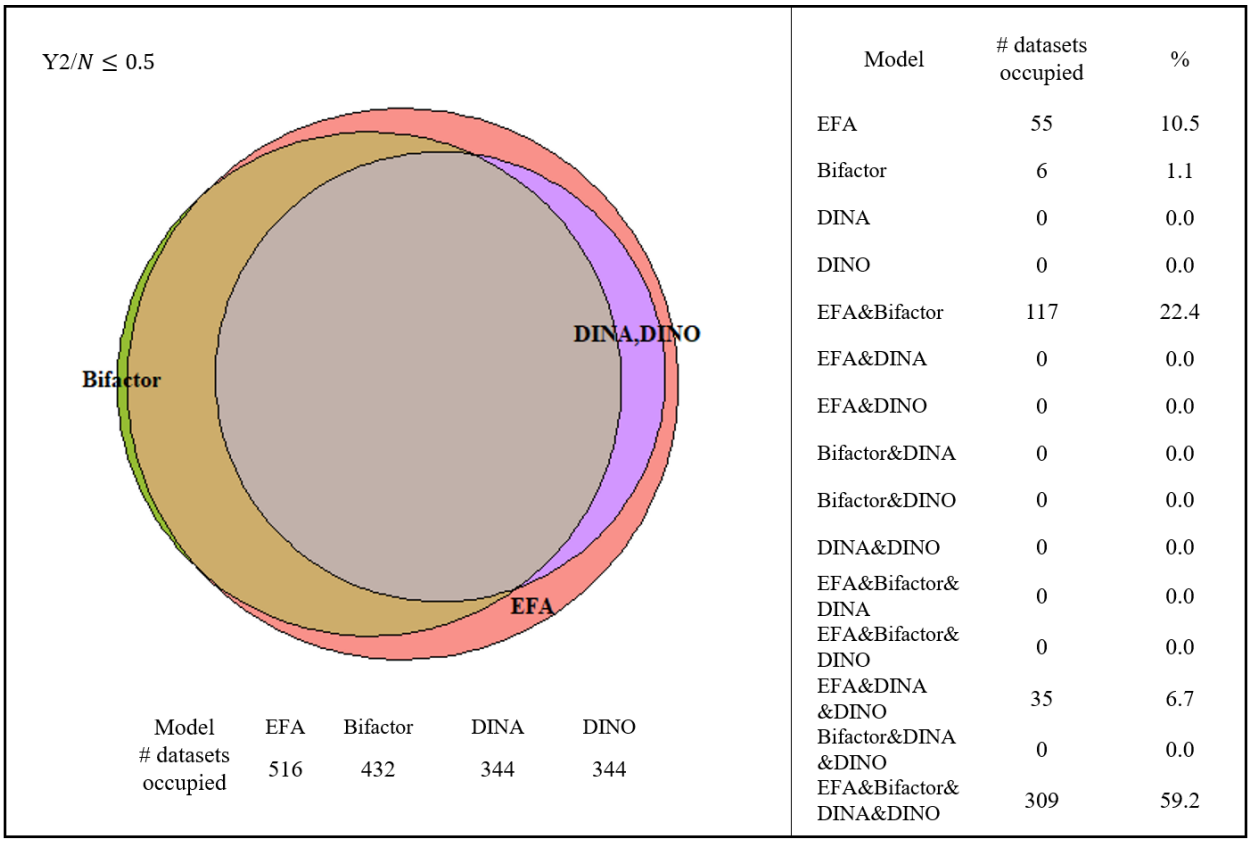


Model	# datasets occupied	%
EFA	21	18.3
Bifactor	3	2.6
DINA	0	0.0
DINO	0	0.0
EFA&Bifactor	30	26.1
EFA&DINA	0	0.0
EFA&DINO	0	0.0
Bifactor&DINA	0	0.0
Bifactor&DINO	0	0.0
DINA&DINO	0	0.0
EFA&Bifactor&DINA	5	4.3
EFA&Bifactor&DINO	2	1.7
EFA&DINA&DINO	0	0.0
Bifactor&DINA&DINO	0	0.0
EFA&Bifactor&DINA&DINO	54	47.0

Figure 5. 18: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 0.3$ for Simplex Sampling Method × PML Estimation

Figure 5. 19 based on $Y2/N \leq 0.5$ shows divergence between the EFA and bifactor model regions of fit. 0.6% of the complete data space or 1.1% of the fitted space showed that the bifactor model fit but the EFA did not. The rate of growth in areas of fit for the DINA and DINO models did not seem to be as large as Figure 5. 18, mirroring the

widening gap between the CDF curves for the EFA and bifactor versus DCMs depicted in Figure 5. 16. In this case, The DINA and DINO regions completely overlapped with each other. The EFA again completely subsumed the DCMs, but some DCMs models showed fit when the bifactor did not (the percentage of non-overlap between bifactor and DCMs was 10.2). All four models fit 30.9% of the complete data space according to $Y2/N \leq 0.5$, consisting of 59.2% of the total number of datasets that showed adequate fit.



Model	# datasets occupied	%
EFA	55	10.5
Bifactor	6	1.1
DINA	0	0.0
DINO	0	0.0
EFA&Bifactor	117	22.4
EFA&DINA	0	0.0
EFA&DINO	0	0.0
Bifactor&DINA	0	0.0
Bifactor&DINO	0	0.0
DINA&DINO	0	0.0
EFA&Bifactor&DINA	0	0.0
EFA&Bifactor&DINO	0	0.0
EFA&DINA&DINO	35	6.7
Bifactor&DINA&DINO	0	0.0
EFA&Bifactor&DINA&DINO	309	59.2

Figure 5. 19: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 0.5$ for Simplex Sampling Method × PML Estimation

Summary

Compared with the FIML estimation, a pronounced characteristic of PML estimation was the decrease in model fit (meaning larger $Y2/N$ values). Another notable difference, clearly apparent in the CDF plot, was the decline in distinguishability between

models when using the PML estimation. This led to higher proportions of all three models being captured in addition to the EFA at various cut points of Y2/N. Using Euler diagrams to further investigate areas of overlap showed that EFA often subsumed the bifactor model area in addition to the ratio of fitted areas between the models being consistently larger than with FIML estimation. In addition, the area in the data space which are fit well by the DINA and DINO models when sliced by Y2/N cutoffs was substantially larger and less distinguishable for the PML approach than the FIML approach. Simulation Study 1 on model recovery of the four models under PML estimation showed that Y2/N values could be considerably smaller for DCMs. Initial investigation implied that Y2/N statistic for DCMs using PML estimation was less sensitive to changes within the model, which could explain the results of Simulation Study 2. Nonetheless, lower Y2/N values were produced with the EFA, followed by the bifactor, and lastly, DCMs across the range of Y2/N as seen by the descriptive statistics and the CDF plot. Also, the EFA and bifactor models (more so for the former) occupied larger areas than the DCMs.

The key findings agreed with that of Bonifay and Cai (2017) and Aytürk Ergin (2020) regarding the difference in FP of models due to the functional form of a model. The EFA and the bifactor model were highly flexible models in general and as well as in comparison to their DCM counterparts, not only above and beyond what can be attributed to its number of parameters (parametric complexity) but also given the model complexity due to the estimation method. This suggested the suitability of the PML estimation approach as a method for investigating FP.

5.5 Simulation Study 3: Investigation of FP in IRT Models using SIS Method and PML Estimation

As verified in Figure 5. 11 of Section 5.4.2, the simplex sampling method does not generate uniformly distributed points over the complete data space defined by the bivariate margins. Rather, it covers only a subspace determined by the highest-order margin (i.e., the full multinomial). Some might argue that trying to capture all areas of the data space (i.e., every possible response pattern) is not worthwhile or relevant if these possible response patterns cannot be plausibly observed in real settings. Nevertheless, there may be consequences in cases where the data comes from outside the subspace (i.e., possible data) and restrict the generalizability of results (Preacher, 2003). Furthermore, investigating the behavior of models under extreme random data can provide more insight into their inherent propensity to fit any possible data (not just plausible).

Therefore, this simulation study examined the FP of models using random data from the complete categorical subspace enabled by the SIS method and PML estimation. The aim was to investigate the fitting behaviors of the four models of Bonifay and Cai (2017) when the data were comprised of more random responses, plausible and non-plausible. In addition, the extent to which Bonifay and Cai's (2017) findings and those from Simulation Study 2 were replicable were also examined.

5.5.1 Simulation Study Setup

Data Generation

A total of 1000 random datasets for seven dichotomously scored items were generated following Table 5. 2. Each dataset consisted of 7 univariate margins and $\frac{7\times6}{2} = 21$ bivariate margins. As probabilities were given, each margin was multiplied by $N = 10000$ to get the number of sample responses in each margin.

Analysis Setup

The four models of the EFA 2PL, bifactor 2PL, DINA, and DINO models were fit to the 1000 random datasets generated via the SIS method using PML estimation. The estimation specifications were kept the same to convergence tolerance of 10^{-6} and 25,000 cycles. The Y2/N indices resulting from fitting each model were recorded and analyzed for all possible replications using CDF plots and Euler diagrams. In addition to comparing the results by model, comparisons across data generation mechanisms were made as well.

5.5.2 Results

Descriptive statistics of Y2/N across replications for the SIS method and PML estimation combination are provided in Table 5. 16. Compared to the descriptive statistics of Table 5. 15, the Y2/N increased significantly. The use of random possible data obtained by uniform sampling over the entire univariate and bivariate margins seemed to have exacerbated problems in using a PML estimation to data that did not fit the model, which is to be expected. Notwithstanding the extremely large sizes of the Y2/N, relative use of Y2/N values in comparing models still showed that EFA on average

and across the quantiles had the lowest values (M=11.41), followed by the bifactor model (M=11.63) and DCMs (M=12.28 for both DINA and DINO) across the range of Y2/N values.

Table 5. 16: Descriptive Statistics of Y2/*N* for SIS Method × PML Estimation

Model	M	SD	Min	Max
EFA	11.408	4.751	2.336	31.291
Bifactor	11.633	4.781	2.346	33.532
DINA	12.282	4.868	2.428	33.245
DINO	12.284	4.854	2.495	32.966

Note. M= mean, SD=standard deviation.

The trends of lower Y2/*N* for the EFA, then bifactor, and, lastly the DCMs were also evident in the CDF plot (Figure 5. 20). However, the CDF curves became even less distinguishable from each other than in Figure 5. 16 where data came from the simplex sampling method. Not only were the EFA and bifactor model CDF curves pushed closer together as were the DINO and DINA model CDF curves (which were nearly identical), but the two sets of model CDF curves (i.e., EFA and bifactor versus DCMs) themselves were closer together than before. This meant the overlap among the four models grew exponentially across a short range of the Y2/N values. For example, Euler diagrams drawn at $Y2/N \leq 12$ (results not provided here), showed that the EFA and bifactor had equivalent size regions with near 100% overlap and the area for each DCM model was over 90% of the entire fitted data space.

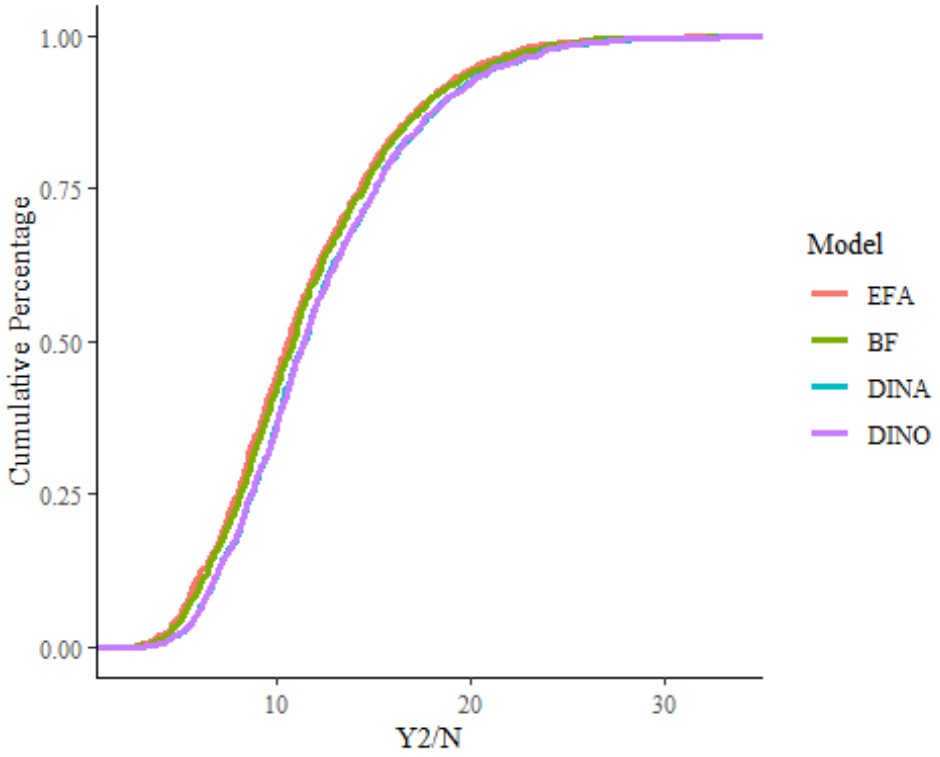
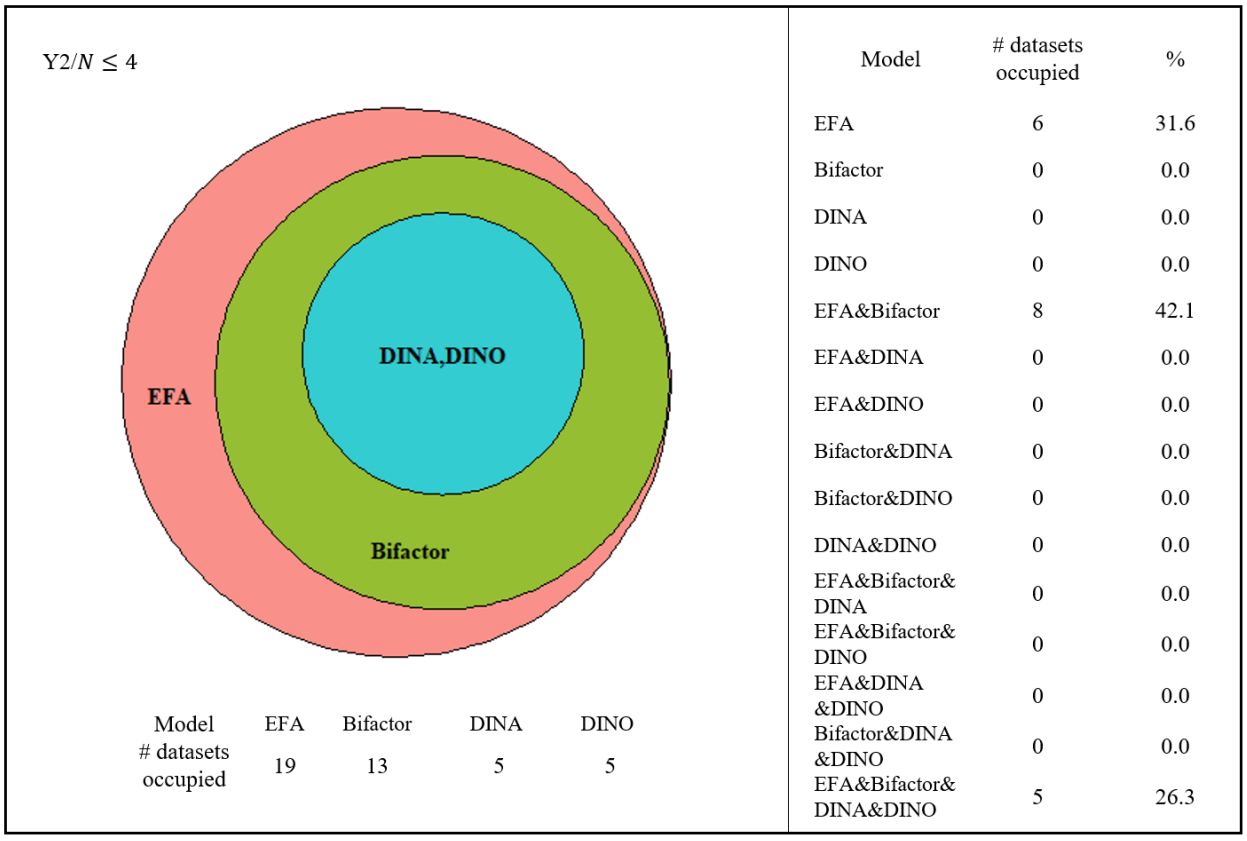


Figure 5. 20: Cumulative Percentage Distributions of the Y2/N statistic for SIS Method × PML Estimation

Figure 5. 20 suggests extreme overlap among the EFA, bifactor, DINA, and DINO models starting from even the lower ends of the Y2/N values. Contrarily, Euler diagrams at smaller cutoffs showed separation between models that could be meaningful. Cutoff values of Y2/N for the diagrams were chosen to be 4, 6, and 8, respectively. For $Y2/N \leq 4$ (Figure 5. 21), the EFA, followed by the bifactor model captured the largest amounts of the data space but the DINA and DINO model also fit a sizable portion of the data space. The EFA model entirely subsumed the bifactor model (with the bifactor model consisting of 68.4% of the data space that fit), which in turn entirely subsumed both DCMs. All four models fit 0.5% of all 1000 datasets which was equal to 26.3% of the fitted area. That is, DINO and DINA fit 26.3% of the fitted area. No differences in fit were

detected between DINA and DINO models as they completely overlapped with each other.

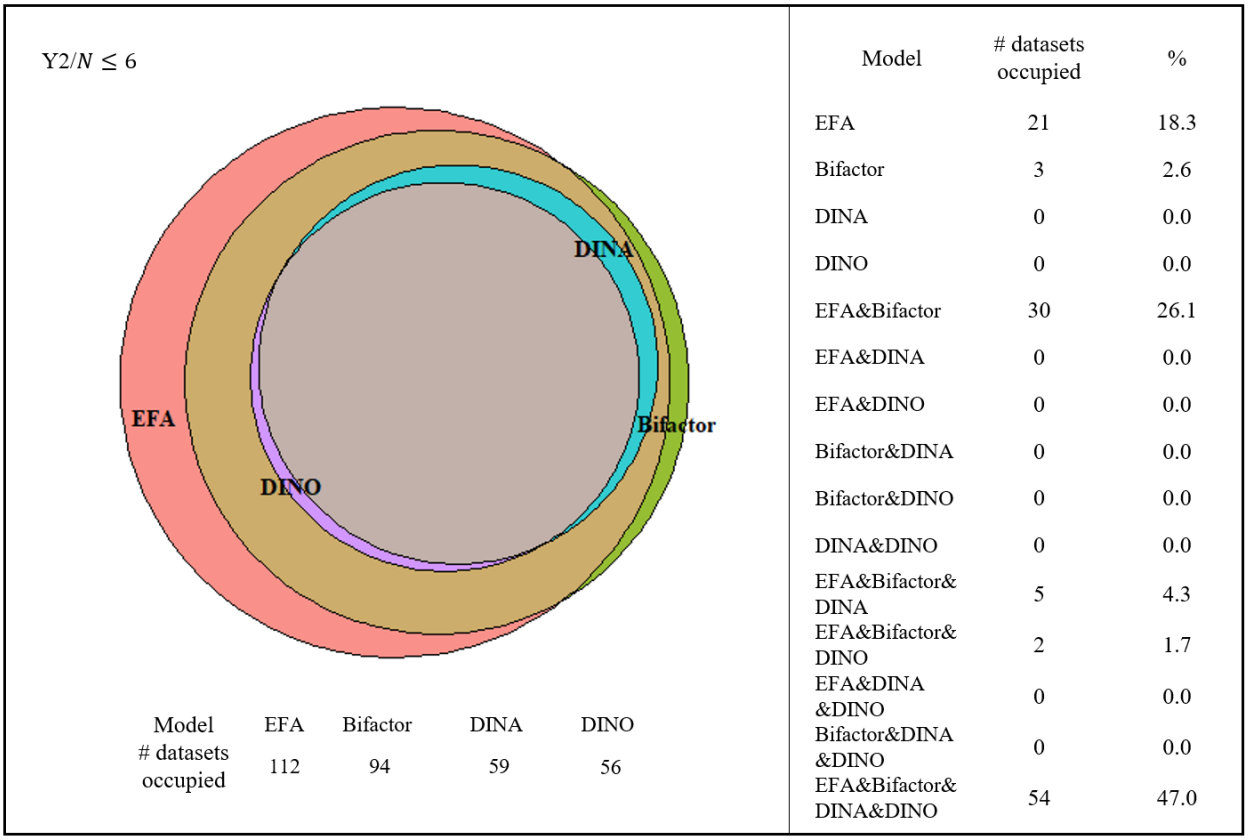


Model	# datasets occupied	%
EFA	6	31.6
Bifactor	0	0.0
DINA	0	0.0
DINO	0	0.0
EFA&Bifactor	8	42.1
EFA&DINA	0	0.0
EFA&DINO	0	0.0
Bifactor&DINA	0	0.0
Bifactor&DINO	0	0.0
DINA&DINO	0	0.0
EFA&Bifactor&DINA	0	0.0
EFA&Bifactor&DINO	0	0.0
EFA&DINA&DINO	0	0.0
Bifactor&DINA&DINO	0	0.0
EFA&Bifactor&DINA&DINO	5	26.3

Figure 5. 21: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 4$ for SIS Method × PML Estimation

The Euler diagram for $Y2/N \leq 6$ was analyzed next (Figure 5. 22). The order of areas of fit was again EFA model > bifactor model > DINA model > DINO model. Bifactor models did not completely overlap, with 3% of the complete and 2.6% of the fitted data space consisting of datasets where the bifactor model fit but the EFA model did not. Both the DINA and DINO models were still subsumed by both the EFA and bifactor models. They in total fit 6.1% of the complete and 53.3% of the fitted data space. However, there was no complete overlap among the DCMs. 4% of the fitted data space

included datasets where the DINA fit but the DINO did not and 2% of the fitted data space showed the opposite results.

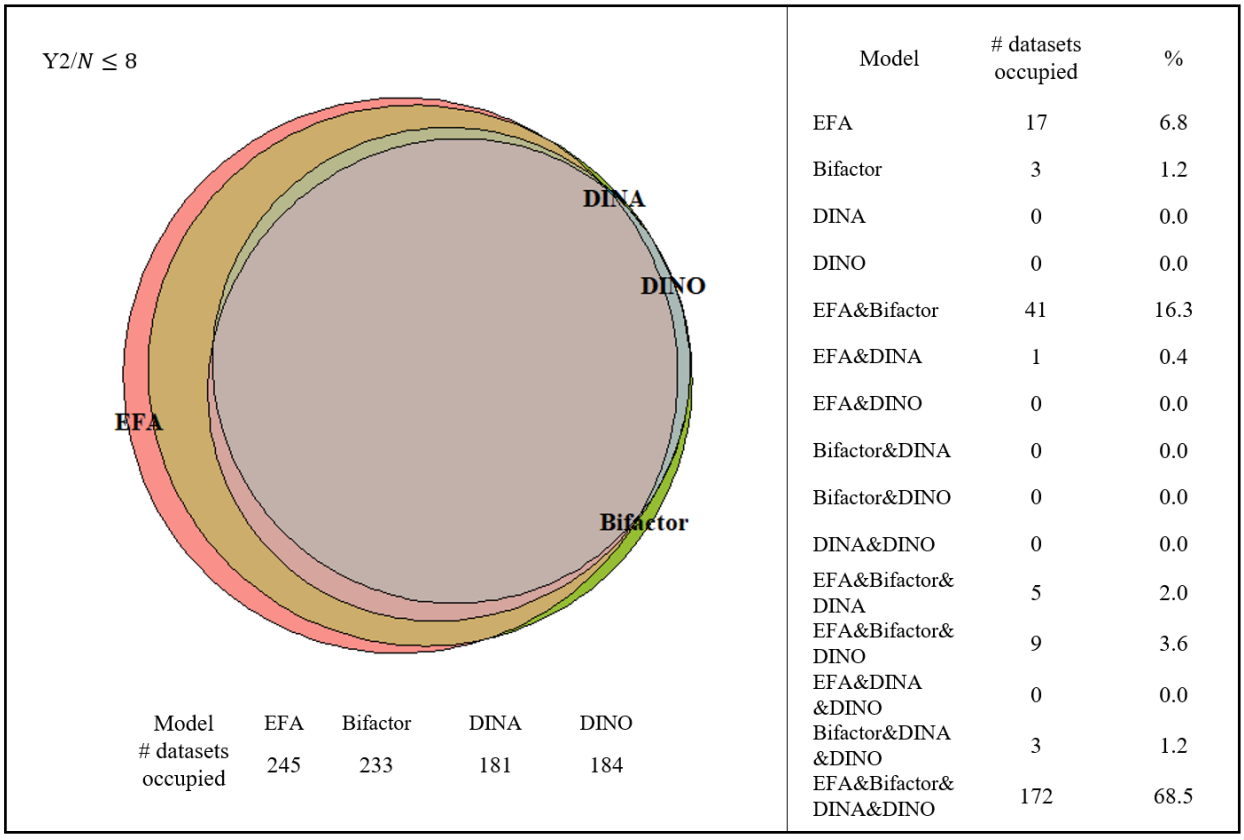


Model	# datasets occupied	%
EFA	21	18.3
Bifactor	3	2.6
DINA	0	0.0
DINO	0	0.0
EFA&Bifactor	30	26.1
EFA&DINA	0	0.0
EFA&DINO	0	0.0
Bifactor&DINA	0	0.0
Bifactor&DINO	0	0.0
DINA&DINO	0	0.0
EFA&Bifactor&DINA	5	4.3
EFA&Bifactor&DINO	2	1.7
EFA&DINA&DINO	0	0.0
Bifactor&DINA&DINO	0	0.0
EFA&Bifactor&DINA&DINO	54	47.0

Figure 5. 22: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 6$ for SIS Method × PML Estimation

Figure 5. 23 of the Euler diagram using $Y2/N \leq 8$ gave similar results as Figure 5. 22 but with an increase in overlap for all models. Conversely, unique regions of fit decreased for all models but were still present. For example, 0.6% of the complete data space or 1.1% of the fitted data space fit the bifactor but not the EFA model. DINA and DINO regions of fit also showed considerable overlap: each of these models occupied 34.4% of the data space. The EFA again completely subsumed the DCMs, but some DCMs models did not always show fit that aligned with the bifactor model (the percentage of

non-overlap between bifactor and DCMs was 10.2%). All four models fit 30.9% of the complete data space according to $Y2/N \leq 8$, which consisted of 59.2% of the total number of datasets that showed adequate fit.



Model	# datasets occupied	%
EFA	17	6.8
Bifactor	3	1.2
DINA	0	0.0
DINO	0	0.0
EFA&Bifactor	41	16.3
EFA&DINA	1	0.4
EFA&DINO	0	0.0
Bifactor&DINA	0	0.0
Bifactor&DINO	0	0.0
DINA&DINO	0	0.0
EFA&Bifactor&DINA	5	2.0
EFA&Bifactor&DINO	9	3.6
EFA&DINA&DINO	0	0.0
Bifactor&DINA&DINO	3	1.2
EFA&Bifactor&DINA&DINO	172	68.5

Figure 5. 23: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 8$ for SIS Method × PML Estimation

5.5.3 Summary

The goal of Simulation Study 3 was to investigate the effects of the data generation method targeting different spaces of data on the FP of the four models. The focus was on how FP might change when we sampled from more random data versus a subspace of more plausible data.

The most notable difference using PML estimation with data sampled from the entire space of the lower-order margins (using the SIS method), compared to more plausible values (under the simplex sampling method), was the exponential decrease in model fit. In other words, the size of the Y2/*N* values for all models not only increased but to an extent larger than when going from the FIML estimation method to the PML estimation method. Notwithstanding, the comparison of models based on descriptive statistics and the CDF plot still showed that the EFA and bifactor model tended to have lower Y2/*N* values than DCMs across the board.

There was even more decreased distinguishability between the models' CDF curves, even though they still indicated more similarity between the EFA and bifactor models than the DCMs (whose curves were again indistinguishable). This led to almost complete overlap between all four models at much earlier ranges of Y2/*N* compared to using the simplex sampling method. This suggested that the Y2/*N* cutoff values had to be small and differences in model fitting patterns would show up at smaller intervals.

Applying cutoffs of Y2/*N* values fitting the criteria and examining the corresponding slices using Euler diagrams showed that the bifactor model tended to dominate more of the space of EFA models than before. Furthermore, a noticeable difference between the Simplex Sampling × PML combination was that there seemed to be distinct regions of fit between the two models, which is more in line with the Simplex Sampling × FIML combination results. This showed that the bifactor fit better than the EFA for a certain number of datasets. It seemed there were higher regions of separation for the SIS × PML combination than the Simplex sampling × FIML, which was corroborated by the Euler diagrams for each (i.e., the former consistently showed regions where only the bifactor model fit, whereas, in the latter, the bifactor model was subsumed by the EFA model).

The data captured by the DCMs was also bigger than the Simplex Sampling × PML combination (and of course, Simplex Sampling × FIML combination). Still, the results were similar in that the areas occupied by the EFA and bifactor models were larger. To add, there was some distinction between the regions of fit between the DINA and DINO models; albeit very sparse.

In general, the results of this study replicated the major findings of Bonifay and Cai (2017) and Aytürk Ergin (2020), as in the case of PML estimation using the simplex sampling method. In fact, the robustness of results in the face of more random data in addition to an estimator prone to fit data worse than FIML only adds to the need to consider the effects of functional form in evaluating model fit.

5.6 Simulation Study 4: Investigation of FP in IRT Models using SIS Method and Iterative Proportional Fitting (IPF)

Simulation Study 2 of Section 5.4.4 displayed that model fit was negatively impacted when using the PML estimation method. Furthermore, Simulation Study 3 in Section 5.5 noted that this was exacerbated as data became more random. Fortunately, the examination of FP does not rely on any one absolute cutoff but on relative comparisons between model fit statistics, whether it be one model or multiple. Results of both simulation studies (i.e., Simulation Studies 2 and 3) showed that the general conclusions about the FP of the four models in question still held. Still, the setup of the simulation study does not allow us to tease apart the combined effects of the SIS × PML estimation.

Accordingly, this study suggested and applied a method for fitting random possible datasets arising from SIS data generation using FIML estimation. The idea was to use the bivariate marginal probabilities produced by the SIS method to reconstruct a

joint distribution that has these probabilities as its bivariate marginals using the iterative proportional fitting procedure (IPFP). The goal was to replicate the results of Simulation Study 2 and 3 to further validate the suitability of the PML estimation method in FP investigation.

5.6.1 Simulation Study Setup

Iterative Proportional Fitting Procedure (IPFP)

The IPFP was first proposed by Deming and Stephan (1940) to estimate cell probabilities in a contingency table subject to certain marginal constraints. Since its conception, the IPFP has been applied to a variety of statistical problems by an equally diverse number of sources (Fienberg, 1970). Among others (and typically employed for fitting log-linear models (Fienberg & Larntz, 1976)), it has been repeatedly suggested for use in simulating multivariate binary data subject to constraints of mainly fixed marginal distributions with specified degrees of association (Lee, 1993; Gange, 1995). Let there be k binary variables $y_1, \cdots, y_k$ with success probabilities $\pi_j = y = (P(y_j = 1)$ for $j = 1, \cdots k)$. As k grows larger, it becomes increasingly infeasible to specify and determine 2^k probabilities. An alternative is to specify k probabilities $\pi_1, \cdots \pi_k$ and the $\frac{(K-1) \times K}{2}$ pairwise-probabilities $\pi_{jj'} = P(Y_j = 1, Y_{j'} = 1), j \neq j')$ and use the IPFP to find a solution of 2^k probabilities where the marginal one- and two-dimensional probabilities satisfy $\{\pi_j\}$ and $\{\pi_{jj'}\}$. There are often many higher-order tables that have the same univariate and bivariate margins, so many solutions are possible. The IPFP ideally converges to one of which of these valid solutions. In comparison to other approaches toward the same goal, the IPFP has the advantage that it produces strictly positive joint probabilities, meaning that none of the theoretically 2^k sequences can be excluded. Furthermore, it can simulate

MVB distributions without assuming an underlying continuous (normal) model so that their restrictions to data, such as positive definite correlation matrices need not be met. This makes the IPFP approach especially attractive in that this study aims to randomly sample from the complete data space to examine FP. Furthermore, we have direct pairwise probabilities to refer to as opposed to having to use correlations or odds ratios as alternatives (Barthelemy & Suesse, 2018).

Data Generation

The univariate and bivariate margins of the datasets used for Simulation Study 3 in Section 5.5 were set as the marginal constraints for the joint distribution of 2^k variables. We start from an array of size ($C_1 \times C_2 \times \ldots \times C_k$) whose cells are all equal to 1. This is the simplest and most uninformative case and starting from a different array would mean adding information that is not available (Ranalli & Rocci, 2016). Then, multiplying by appropriate factors, we adjust the cell probabilities of the joint distribution successively to match the probabilities for each bivariate table. The process is continued until convergence is reached. Convergence is defined as the difference in fitted probabilities between two consecutive iterations being less than an arbitrary $\epsilon > 0$.

Analysis Setup

The 1,000 IPF generated datasets of multinomial probabilities were fit to the EFA, bifactor, DINA, and DINO models using flexMIRT with the same specifications as Simulation Study 2 in Section 5.3. The Y2/*N* index after each model fit was recorded and analyzed for all replications using CDF plots and Euler diagrams. Using IPF along with FIML allowed for examining differences and similarities due to data generation in addition to data estimation methods.

5.6.2 Results

Table 5. 17 displays the descriptive statistics of the Y2/*N* statistic for the four models in question. Once more, the difference in Y2/*N* values was the most perceptible. The descriptive statistics are similar to the results produced by the Simplex Sampling × PML estimation combination; although somewhat higher. They are much smaller in magnitude compared to the SIS × PML estimation combination. Comparing means showed that on average, the EFA and bifactor models produced Y2/*N* values of 0.7 or lower. The DINA and DINO models tended to have Y2/*N* values of 1. The results suggest yet again that the EFA and bifactor model generally had lower Y2/*N* values compared to the DINA and DINO models.

Table 5. 17: Descriptive Statistics of Y2/*N* for SIS Method × FIML Estimation

Model	M	SD	Min	Max
EFA	0.569	0.268	0.087	2.528
Bifactor	0.663	0.289	0.112	1.996
DINA	1.141	0.444	0.190	3.400
DINO	1.126	0.467	0.253	4.090

Note. M= mean, SD=standard deviation.

On the other hand, Figure 5. 24 describing the CDF curves of the Y2/*N* for the models gives results akin to that from FIML estimation using the simplex sampling method. The common factor across all other comparisons (i.e., Simplex sampling × FIML estimation, Simplex sampling × PML, and SIS × PML estimation combinations) is that the EFA model had the lowest Y2/*N* value for any benchmark percentage with increasing values in the order of the bifactor and then DINA and DINO models. The DCMs

consistently had nearly identical curves. However, Figure 5. 24 is much more comparable to FIML estimation concerning the spacing between the three curves for the EFA, bifactor, and DCMs. Despite not being as pronounced as under FIML estimation (i.e., Figure 5. 12), they were still wide enough to gauge sizable differences in Y2/*N*s values between EFA and bifactor models compared to their DCM counterparts in their fit the same amount of data (i.e., less overlap between the former two and latter two models).

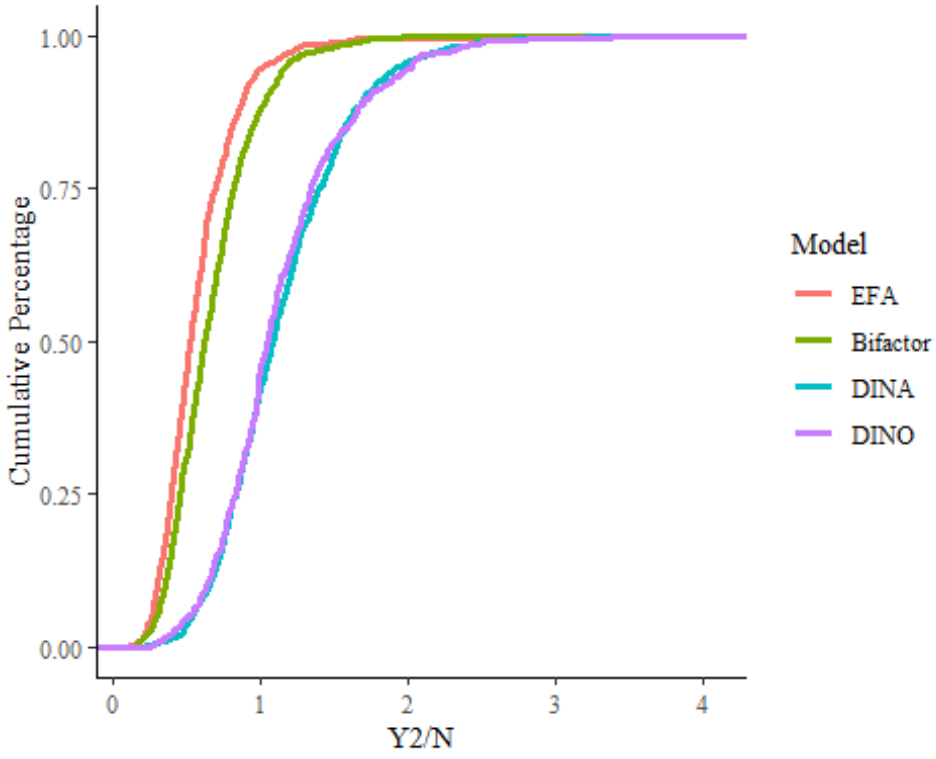
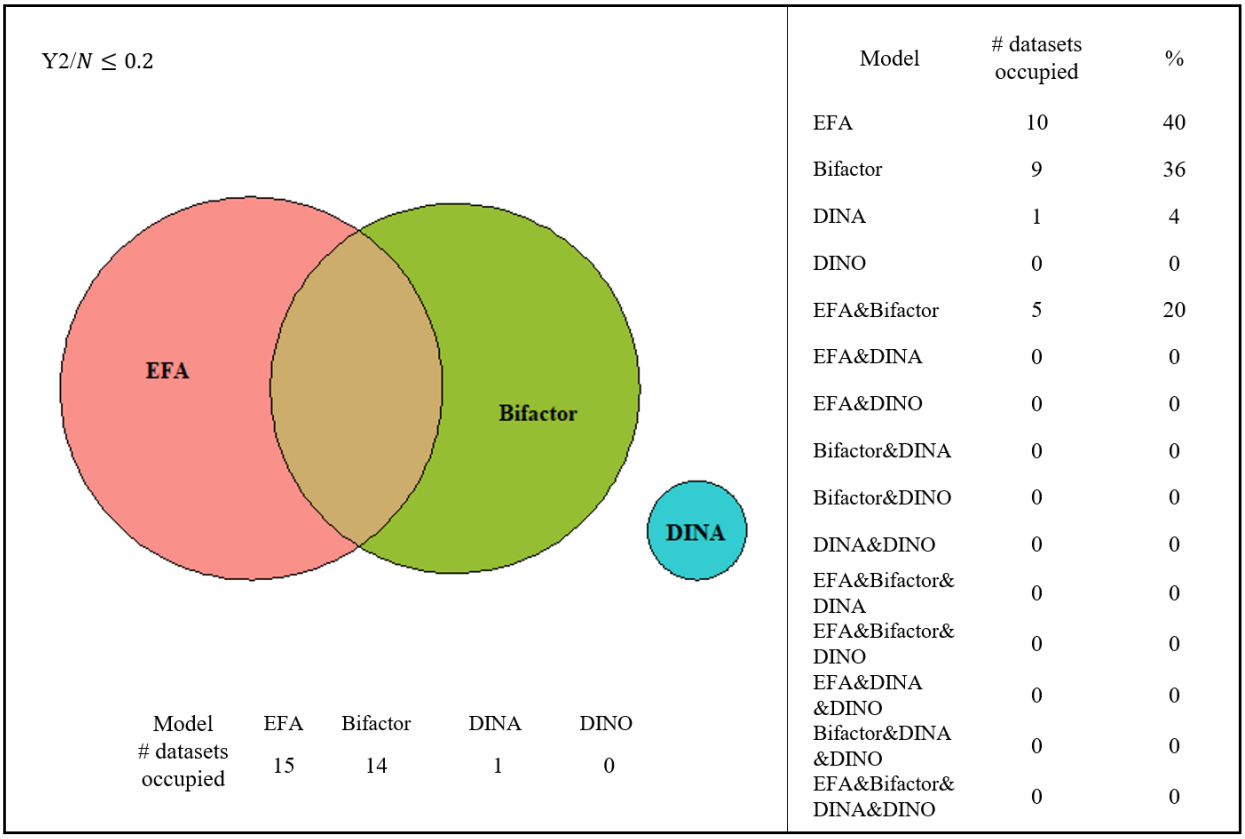


Figure 5. 24: Cumulative Percentage Distributions of the Y2/*N* statistic for SIS Method × FIML Estimation

Euler diagrams from different cut-points of Y2/*N* values were utilized to provide a summary of the growth in the space occupied by models as Y2/*N* increased. The chosen cut-points were $Y2/N \leq 0.2$, $Y2/N \leq 0.4$, and $Y2/N \leq 0.7$. At the $Y2/N \leq 0.2$ (Figure

5. 25), EFA, bifactor models, and a single DINA model were captured. The size of the areas for the EFA and bifactor model rivaled each other with the former fitting 1.5% of the datasets and the latter fitting 1.4% of datasets. The overlap between the two models was 20.8%, meaning that they disagreed with regard to fit more than they agreed. The regions where only EFA fit was 1% of the complete dataset and 40% of the fitted data space and the region where only the bifactor model fit was 0.9% of the complete data space and 36% of the fitted data space. Like this, they had nearly the same number of datasets for which they each fit discrepant regions of data. Most notable was that the single dataset that the DINA model fit was also located in a region not overlapping with either the EFA or bifactor model at even this relatively stringent cutoff. For this dataset, no DINO models fit the criterion $Y2/N \leq 0.2$.

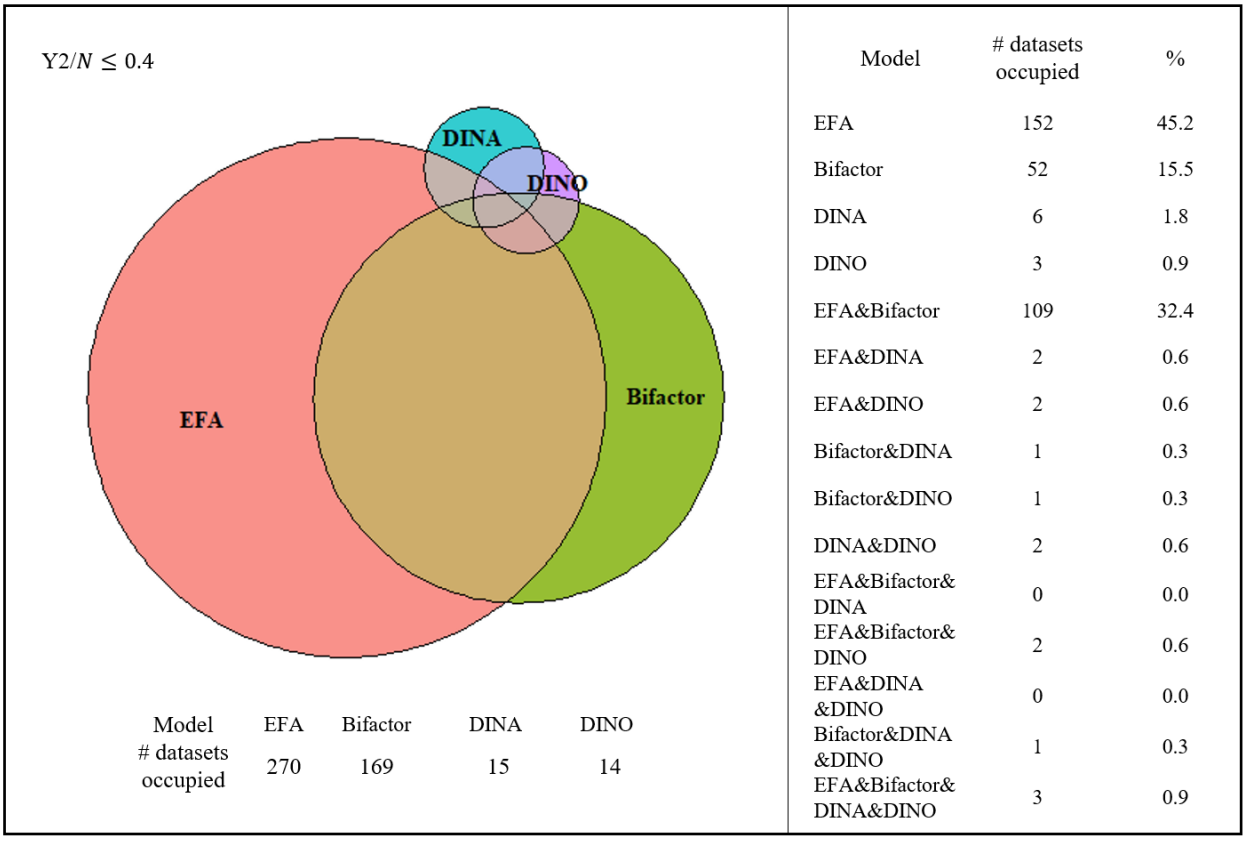


Model	EFA	Bifactor	DINA	DINO
# datasets occupied	15	14	1	0

Model	# datasets occupied	%
EFA	10	40
Bifactor	9	36
DINA	1	4
DINO	0	0
EFA&Bifactor	5	20
EFA&DINA	0	0
EFA&DINO	0	0
Bifactor&DINA	0	0
Bifactor&DINO	0	0
DINA&DINO	0	0
EFA&Bifactor&DINA	0	0
EFA&Bifactor&DINO	0	0
EFA&DINA&DINO	0	0
Bifactor&DINA&DINO	0	0
EFA&Bifactor&DINA&DINO	0	0

Figure 5. 25: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 0.2$ for SIS Method × FIML Estimation

When the cutoff is $Y2/N \leq 0.4$ (Figure 5. 26), all four models showed areas with fit but the regions of fit for the EFA and bifactor model were significantly more prominent. The EFA model fit 27% of the complete data space and 80.3% of the fitted space and the bifactor fit 16.9% of the complete data space and 50.3% of the fitted space, which shows growth for the EFA compared to the bifactor. The two models had more overlap with each other (10.9% of the complete data space and 50.5% of their combined fitted data space), but also had areas where only one of them had $Y2/N$ below 0.4. The DCMs consisted of 2.3% of the complete data space, which was 6.8% of the fitted data space. Although smaller compared to the fitted areas for the EFA and the bifactor model, not only were the two DCMs distinguished in terms of regions of fit, but they both included data spaces where each solely satisfied $Y2/N \leq 0.4$. The degree of overlap between the two was 26.1% (i.e., 6 among 23 datasets). The area where only the DINA model fit was 0.6% of the entire data space and 1.8% of the fitted data space, while the region where only the DINO model fit was half of that. 14 datasets or 60.9% of the fitted datasets that fit the DINA or DINO model fell into the region of fit for either the bifactor or EFA models. The DINA model also fit one more dataset than the DINO model.

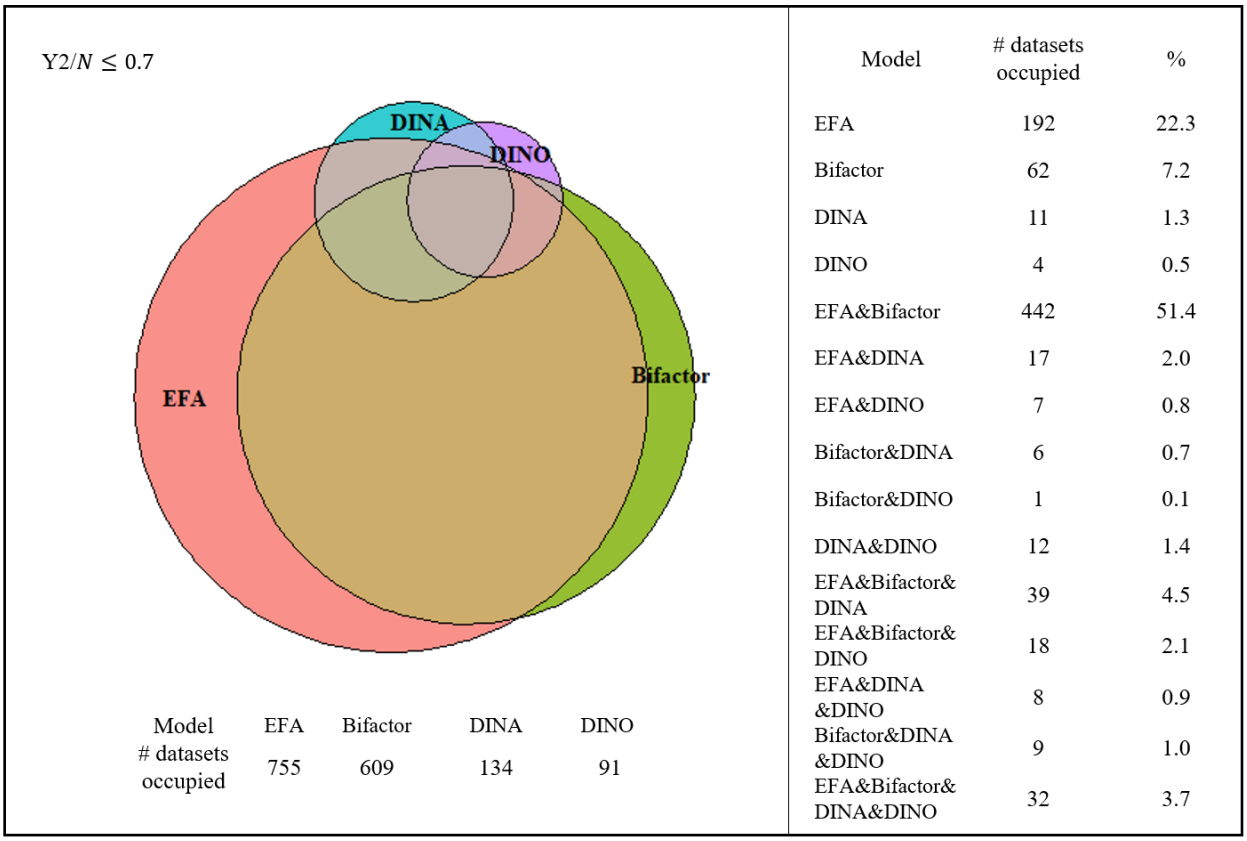


Model	# datasets occupied	%
EFA	152	45.2
Bifactor	52	15.5
DINA	6	1.8
DINO	3	0.9
EFA&Bifactor	109	32.4
EFA&DINA	2	0.6
EFA&DINO	2	0.6
Bifactor&DINA	1	0.3
Bifactor&DINO	1	0.3
DINA&DINO	2	0.6
EFA&Bifactor&DINA	0	0.0
EFA&Bifactor&DINO	2	0.6
EFA&DINA&DINO	0	0.0
Bifactor&DINA&DINO	1	0.3
EFA&Bifactor&DINA&DINO	3	0.9

Figure 5. 26: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 0.4$ for SIS Method × FIML Estimation

The pattern of results for $Y2/N \leq 0.7$, described in Figure 5. 27, was more or less like Figure 5. 26 with $Y2/N \leq 0.4$. When $Y2/N \leq 0.7$, over 75% of the complete datasets are fit by the EFA model and 60.9% are fit by the bifactor model. The two models still show separation, although the area of distinct fit for the bifactor model decreased. Including the areas of discrepant fit between the two, they fit 83.3% of the complete data space, which was 96.9% of the fitted data space. Conversely, this meant that there was still 3.1% of the fitted data space where only the DCMs fit. The two DCMs combined made up 16.1% of the complete data space and 19.1% of the fitted data space. The degree of overlap between the two models was 37.2%, which was larger than before, and 90.9% of their datasets that fit (which was equal to 149 datasets) also fit either the EFA or bifactor.

However, there were still regions of fit where only the DINA or DINO fit. The DINA had larger areas of fit compared to the DINO (13.4% of the complete data space for DINA versus 8.4% of the complete data space for DINO).



Model	EFA	Bifactor	DINA	DINO
# datasets occupied	755	609	134	91

Model	# datasets occupied	%
EFA	192	22.3
Bifactor	62	7.2
DINA	11	1.3
DINO	4	0.5
EFA&Bifactor	442	51.4
EFA&DINA	17	2.0
EFA&DINO	7	0.8
Bifactor&DINA	6	0.7
Bifactor&DINO	1	0.1
DINA&DINO	12	1.4
EFA&Bifactor& DINA	39	4.5
EFA&Bifactor& DINO	18	2.1
EFA&DINA &DINO	8	0.9
Bifactor&DINA &DINO	9	1.0
EFA&Bifactor& DINA&DINO	32	3.7

Figure 5. 27: Hypothetical Approximate Regions of the Captured Data Space at $Y2/N \leq 0.4$ for SIS Method × FIML Estimation

5.6.3 Summary

The results of a SIS method with FIML estimation to investigate FP had overlap with all other possible conditions used to examine the FP of four IRT models (i.e., Simplex Sampling × FIML estimation combination, Simplex Sampling × PML estimation combination, and SIS × PML estimation combination). Among the three, the magnitude of increase in the overall $Y2/N$ values was a little larger than that of the Simplex

Sampling × PML estimation combination. The increase in magnitude when compared with Simplex Sampling × FIML estimation was expected as the datasets were more random. The much larger values of Y2/*N* values from the SIS × PML estimation combination compared to both the Simplex Sampling × PML estimation and SIS × FIML estimation combination implied an interaction effect between the more random data and the LI-based method pushing Y2/N values to more extremes. Nonetheless, the finding across all conditions was still that the EFA, followed by the bifactor had lower Y2/*N* values when compared to the DCMs based on descriptive statistics.

The CDF plots and the Euler diagrams also gave a consistent result with all other conditions that there seemed to be two sets of models at play: the EFA and bifactor models versus the DCMs. The EFA and bifactor models were always more in sync with each other than their DCM counterparts, although still distinguishable from each other. The overlap between the DCMs was often so great that differences rarely seemed detectable in the CDF plots. Regarding the CDF plots, the results for the SIS × FIML estimation combination were most comparable to those using simplex sampling and FIML estimation with substantial distances in the Y2/*N* values among the EFA and bifactor models with the DCMs. This led to disproportionate regions of fit among the two model types in the Euler diagrams that extended well on to higher ranges of Y2/*N*. Regions of fit of the DCMs did not increase much for the SIS × FML estimation combination. Regions of fit increased more rapidly for the DCMs in the Simplex Sampling × PML estimation combination and even more so than their SIS method counterparts, although they both were more skewed towards greater regions of fit for the EFA and bifactor models.

Nevertheless, the results of the SIS × FIML estimation combination were also unique in their heightened ability for DCMs to fit datasets where both the EFA and

bifactor did not fit based on Y2/*N* cutoffs. Bonifay and Cai (2017) also show that this can occur but, that it only occurred only at very high Y2/*N* cutoffs (although not shown, the results of Simplex Sampling × FIML estimation in this study echoed Bonifay and Cai's (2017) finding). Contrarily, such unique regions of fit for the DCMs appeared early on (i.e., at low Y2/N cutoffs). The areas of fit were also larger for the DCMs but not to the extent of either simplex sampling × PML estimation or SIS × PML estimation combination.

Differences aside, the results for the SIS method × FIML estimation were in harmony with those of Bonifay and Cai (2017) and Aytürk Ergin (2020) about differences in FPs of models due to their functional form. For the four models in question, we repeatedly saw the undesirably high fitting propensity of the EFA and especially the bifactor model. The impact of functional form was still salient even when sampling from the entire data space of the lower-order margins. That is, the EFA and bifactor had the flexibility to fit even extreme random data, especially when compared to the DCMs. Also, these findings provide support for the LI-based data sampling method and provide further evidence for the PML estimation as well functioning and suitable alternative to the method suggested by Bonifay and Cai (2017).

Chapter VI

Discussion

6.1 Summary of Results

Model fit evaluation requires a balance between a model's GoF to observed data and their generalizability to fit future unseen data. Following Occam's razor, a model should not only fit the data well, but should do so in the simplest manner possible (Myung et al., 2005). However, in the social sciences, and especially IRT, there is an over-reliance on GoF statistics and hence, a tendency to choose more complex models, of which we consistently find the bifactor model to be part of (Bonifay & Cai, 2017). This can be problematic because such models may not necessarily be the data generating model but simply be highly flexible. Thus, it may produce a false sense of good model fit. In other words, model complexity due to the functional form of models has not been adequately accounted for.

Preacher (2006) suggested fitting propensity (FP) as an alternative measure of model complexity or parsimony that can take into account multiple complexities. FP refers to a model's inherent flexibility to fit diverse patterns of data, all else being equal. However, the requirement of generating all possible item response patterns for IRT models when using a multinomial framework impedes the study of FP of IRT models.

Focusing on LI methods (Bolt, 2005), this study proposed a novel data generation algorithm that generates random datasets for IRT models using solely the lower-order moments. Accompanying estimation methods capable of fitting such data with only lower moments were also derived based on PML estimation. If successful, the LI-based

approach may alleviate the computational burden of the exponential increase in data responses to generate and to fit using models. In turn, this would expand the utility of model complexity metrics as a means for model evaluation. The goal of this dissertation was to explore the relative flexibilities of the functional forms of the IRT models in Bonifay and Cai (2017) when the suggested methodology was applied.

This dissertation first proposed a data generation algorithm using solely the lower-order margins to promote computational efficiency, capitalizing on the fact that IRT models can be formulated using the marginal moments of the MVB distribution (Maydeu-Olivares & Joe, 2014). The sampling of the first- and second-order margins involves probabilities up to only pairs of items, as opposed to the full multinomial probabilities. Therefore, there is a significant decrease in the number of data patterns to be generated.

Inspired by classical works of sampling of $m \times n$ contingency tables with fixed margins (e.g., Fienberg, 1999) and adapting a sequential importance sampling (SIS) approach capable of sampling both two-way and multi-way tables with many rows and columns (Chen, Diaconis, et al., 2005), their fundamental principles were applied to the IRT framework. The algorithm could readily generate random dichotomous and/or polytomous item data of large quantities. The SIS algorithm was theoretically-sound and , when empirically compared with the conventional simplex method of data generation (Bonifay & Cai, 2017) in the case of binary data, showed adequate and comparable coverage of the desired complete categorical data space. Such results indicated that the data generation algorithm was promising.

The generated datasets using the proposed SIS algorithm contain only information regarding the univariate and bivariate margins. Accordingly, item

parameters must be calibrated using only such data, which can be done using LI or CML estimation methods that do not rely on the full multinomial item response patterns (e.g., Joreskog & Moustaki, 2001). The motivation behind CML estimation is to replace the original full likelihood, which is often complex and intractable, with a function that is easier to evaluate, and hence to maximize (Cox & Reid, 2004). Among different flavors of CML estimation, PMML estimation that uses bivariate marginal information from pairs of observations in estimation and factors in possible dependence between variables (Cox & Reid, 2004) is especially popular. The LI-estimation approach has been applied to IRT models under the UV framework assuming continuous variables, which limits their applicability to a small subset of IRT models.

The goal of this dissertation was to provide a general form for PML estimation that can be employed in theory to any IRT model. The general form was used to construct pairwise log-likelihoods for each of the five dichotomous models of Bonifay and Cai (2017) Bonifay and Cai (2017)—the EFA, bifactor, DINA, DINO, and unidimensional 3PL models—which were maximized using conventional algorithms such as the EM algorithm.

While all binary, they reflected various types of IRT models. Except for the unidimensional 3PL models, a small simulation study on the remaining four models showed adequate item parameter recovery for all four of the remaining models. Furthermore, the results of PML were comparable to a corresponding FIML method. Standard error estimates were also similar across the two methods with PML estimation requiring the use of the Godambe information matrix for standard error calculation.

Lastly, the suitability of the proposed LI-methods for quantifying FP was tested under different estimation and data generation conditions. Four conditions of the

Simplex Sampling Method × FIML estimation combination, Simplex Sampling Method × PML estimation combination, SIS Method × PML estimation combination, and SIS Method × FIML estimation combination were investigated. The aim was to determine the extent to which the conditions replicated the general findings of Bonifay & Cai (which used the Simplex Sampling Method × FIML estimation combination) so that comparisons of results when applied to new models would be meaningful. The main goal was to take advantage of the proposed approach to investigate the FP of models too complex under the traditional approach.

All in all, the results of this study corroborated the primary findings of previous studies (and the FIML-based FP results of this dissertation) that the four models are not equal in their propensities to fit, even when controlling for the number of parameters. The EFA and the bifactor can be considered highly flexible models, above and beyond what can be attributed to its number of parameters (parametric complexity). These general conclusions did not change even in the face of added model complexity due to the model estimation method or data generation method (Pitt et al., 2002; Preacher, 2003). The results suggested that both the data generation and PML estimation can be a suitable alternatives to FIML estimation for FP analysis. In addition, this study also presented a method used on the IPFP that can recover the joint probabilities satisfying the marginal probabilities from the SIS method. Then, traditional FIML methods could be readily applied to examine FP as well.

Furthermore, in validating the LI-based approaches, this dissertation also added to the discussion of the impact of model complexity due to data generation and estimation method (Pitt et al., 2002) on model fit. In terms of FP, the results alluded to the importance of functional form over other aspects of complexity as well as parametric complexity. In fact, the robustness of results against estimation methods and data

generation methods in favoring more complex models to a high degree consolidates the importance of accounting for structural complexity in model evaluation.

6.2 Implications

Recall Box (1976) and his statement that "all models are wrong, but some are useful." Models are used to approximate or explain reality so that they cannot perfectly capture every aspect of it. “But some are useful" implies that simplifications of reality can nonetheless be quite useful as they can help us explain, predict, and understand various phenomena. It is the role of researchers to find these useful models. GoF, generalizability, and model complexity define the usefulness of models from a model-fitting perspective.

All models should be able to fit the data at hand. Nonetheless, the consistent conclusion across the studies of FP in this dissertation as well as other previous studies of FP is that models are not equal in their ability to fit data sets. That is, solely relying on GoF indices can result in models that appear to fit well to any potential dataset. A model producing good GoF is not necessarily the one that reflects the true processes underlying the observed data but could be due to the fact that the model’s functional form is too flexible so that it can accommodate any given dataset (Preacher, 2006). This is not to say that complex IRT models should be thrown out completely. Instead, it stresses the importance of the justifiability of a model on theoretical grounds. In addition, it argues that even when the model is theoretically driven, researchers should be aware that various aspects of complexity (structural and parameter complexity alike) of statistical models can hinder falsifiability.

6.3 Future Directions

The proposed data generation algorithm and complementing PML estimation method allow the sampling and estimation of many dichotomous and polytomous items. The next step is to extend the approach for investigating the FP of polytomous IRT models, where not only do the data patterns quickly exceed manageable levels but there is, as far as we know, no confirmed valid random data generation mechanism that exists for sampling from the complete data space. Among polytomous models, the FP of the graded response model (GRM) and the generalized partial credit model (GPCM) are at the center of interest. These two models provide the same parameterization with a discrimination parameter and k-1 boundary parameters (k = the number of response categories) per item so that the number of parameters is the same for both models. Nonetheless, they fall into different model classes to have different functional forms ("difference" model for the GRM versus a "divide-by-total" model (Thissen & Steinberg, 1986)). In addition, the scoring process underlying the GRM (grade response scoring) is conceptually different from that of the GPCM (partial credit scoring).

The issue of model fit comparison of the GRM and GPCM has been an ongoing research topic (e.g., Kang et al., 2009; Ostini et al., 2014), to which FP can contribute with its focus on the effects of the functional form of models. Initial results of the item parameter recovery of PMM estimation to the GRM, GPCM, as well as the nominal response model (NRM), showed promising results. Preliminary results based on the unidimensional GRM and GPCM suggest a tendency of higher FP for the former compared to the latter. These results align with previous research such as Aytürk Ergin (2020). The aim is to explore further whether such findings generalize to multidimensional cases of the GRM and GPCM, such as the bifactor and EFA models.

In addition, the proposed LI-based method can be useful in investigating in more detail the effect of other facets of model complexity. Other than parametric complexity and structural complexity, Pitt et al. (2002) and Preacher (2006) also identified plausible versus possible data space, range of parameter space, sample size, estimation method (specifically in the shape of the probability distribution specified in the likelihood function), and research design, as potential factors contributing to a model's flexibility to fit a wider range of data. As the byproduct of applying the SIS data generation method and/or PML estimation, factors of model complexities regarding data space and estimation method were introduced, and comparisons were made. While the salience of the FP functional form has been detected, it was still possible to see the impact of using these methods in FP investigation. More in-depth investigation regarding the effect of multiple model complexities, both separate and combined, will contribute to a more comprehensive, multi-faceted approach for selecting the most appropriate model in various practical application settings.

Furthermore, the methods and results of this dissertation need not be limited to quantifying FP. For example, the data generation method may help examine the performance of not only the proposed PML estimation but of CML estimation overall in light of model misspecification. The results are divided, with some suggesting robustness compared to FIML methods (Varin et al., 2011; Xu & Reid, 2011), while others show evidence of the opposite situation (Ogden, 2016). The results of this dissertation found a significant decline in fit, ascertained using the $Y2/N$ statistic, when using PML estimation compared to FIML. In addition, this was exacerbated as data became more random. One possible explanation may be the assumptions of the marginal distributions being violated (Ogden, 2016).

Moreover, the DCMs estimated using the PML method seemed to be less impacted in terms of model fit, as indicated by their higher propensity to fit under PML estimation. The EFA and bifactor model and DCMs differ mainly in their assumptions of the latent variable distribution. Thus, this may be the effect of the shape of the probability distribution mentioned by Pitt et al. (2002). Simulation Study 1 on model recovery under PML estimation revealed that for DCMs, the Y2/*N* statistic could be much smaller in PML estimation compared to FIML estimation. Initial investigation showed that the difference in the Y2/*N* statistic for the two estimation methods grew smaller as the data fit became closer to the assumed model, implying that DCMs using PML estimation were less sensitive to changes within the model.

The random data generation method itself has the potential for use in many other simulation studies. Most simulated datasets require a model to get the simulated responses. Depending on the choice of the model, we can unfairly favor some results or models over other contenders. The SIS data generation method can be used to generate random or more realistic data that is not as influenced by a particular model. In addition, the range of applications further increases when the presented SIS method is coupled with the IPFP to generate higher-dimensional data, as shown in this study (i.e., Simulation Study 4). Univariate and bivariate margins are much easier to set or obtain for use in generating data when compared to having to set and generate data using full multinominal probabilities. Furthermore, the bivariate margins can be easily converted into correlations or odds ratios and vice versa, which can be used in the process if needed. Examples of applications include the generation of random data for longitudinal studies and the generation of random correlation matrices.

BIBLIOGRAPHY

Aytürk Ergin, E. (2020). Fitting Propensities of Item Response Theory Models. (Unpublished doctoral dissertation, Fordham University)

Baker, F. B., & Kim, S. H. (Eds.). (2004). *Item response theory: Parameter estimation techniques*. CRC Press.

Bartholomew, D. J., & Leung, S. O. (2002). A goodness of fit test for sparse 2p contingency tables. *British Journal of Mathematical and Statistical Psychology*, *55*(1), 1-15.

Barthélemy, J., & Suesse, T. (2018). mipfp: An R package for multidimensional array fitting and simulating multivariate Bernoulli distributions. *Journal of Statistical Software*, *86*, 1-20.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443-459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model forn dichotomously scored items. *Psychometrika*, *35*(2), 179-197.

Bolt, D. (2005). Limited and full information estimation of item response theory models. In A. Maydeu-Olivares & J.J. McArdle (Eds.), *Contemporary psychometrics* (pp. 27–71). Mahwah: Earlbaum.

Bonifay, W. (2015). *An integrative framework of model evaluation* (Doctoral dissertation, UCLA).

Bonifay, W. (2019). *Multidimensional item response theory*. Sage Publications.

Bonifay, W., & Cai, L. (2017). On the complexity of item response theory models. *Multivariate behavioral research*, *52*(4), 465-484.

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791-799.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581-612.

Cai, L. (2021). flexMIRT® version 3.6.4: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cai, L., Choi, K., Hansen, M., & Harrell, L. (2016). Item response theory. *Annual Review of Statistics and Its Application*, *3*, 297-321.

Cai, L., Maydeu-Olivares, A., Coffman, D. L., & Thissen, D. (2006). Limited-information goodness-of-fit testing of item response theory models for sparse 2P tables. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 173-194.

Cai, L., & Moustaki, I. (2018). Variable Models for Categorical Outcome Variables. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*, 253.

Cai, L., & Thissen, D. (2014). Modern approaches to parameter estimation in item response theory. *Handbook of item response theory modeling: Applications to typical performance assessment*, 41-59.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological methods*, *16*(3), 221.

Cattelan, M., & Sartori, N. (2016). Empirical and simulated adjustments of composite likelihood ratio statistics. *Journal of Statistical Computation and Simulation*, *86*(5), 1056-1067.

Chen, Y., Diaconis, P., Holmes, S. P., & Liu, J. S. (2005). Sequential Monte Carlo methods for statistical analysis of tables. *Journal of the American Statistical Association*, *100*(469), 109-120.

Chen, Y., Dinwoodie, I., Dobra, A., & Huber, M. (2005). Lattice points, contingency tables, and sampling. *Contemporary Mathematics*, *374*, 65-78.

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Cox, D. R., & Reid, N. (2004). A note on pseudolikelihood constructed from marginal densities. *Biometrika*, *91*(3), 729-737.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, *11*(4), 427-444.

Diaconis, P., & Efron, B. (1985). Testing for independence in a two-way table: new interpretations of the chi-square statistic. *The Annals of Statistics*, 845-874.

Diaconis, P., & Efron, B. (1987). Probabilistic-geometric theorems arising from the analysis of contingency tables. In *Contributions to the Theory and Application of Statistics* (pp. 103-125). Academic Press.

Falk, C., & Muthukrishna, M. (2020). Parsimony in Model Selection: Tools for Assessing Fit Propensity. *arXiv preprint arXiv:2007.03699*.

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, *41*(3), 907-917.

Fienberg, S. E., & Gilbert, J. P. (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association*, *65*(330), 694-701.

Fienberg, S. E., & Larntz, K. (1976). Log linear representation for paired and multiple comparisons models. *Biometrika*, *63*(2), 245-254.

Gange, S. J. (1995). Generating multivariate categorical variates using the iterative proportional fitting algorithm. *The American Statistician*, *49*(2), 134-138.

Gao, X., & Song, P. X. K. (2010). Composite likelihood Bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, *105*(492), 1531-1540.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423-436.

Godambe, V. P. (1960). An optimum property of regular maximum likelihood equation. *The Annals of Mathematical Statistics*, 31(4), 1208-1211.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of mathematical psychology*, *44*(1), 133-152.

Grünwald, P. D., Myung, I. J., & Pitt, M. A. (Eds.). (2005). *Advances in minimum description length: Theory and applications*. MIT press.

Hansen, M. P. (2013). *Hierarchical item response models for cognitive diagnosis*. (Doctoral dissertation, UCLA).

Haslbeck, J., & van Bork, R. (2022). Estimating the number of factors in exploratory factor analysis via out-of-sample prediction errors. *Psychological Methods*.

Joe, H., Reid, N., Somg, P. X., Firth, D., & Varin, C. (2012, April). Composite likelihood methods. In *Report on the Workshop on Composite Likelihood*.

Jöreskog, K. G. (1990). New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality and Quantity*, *24*(4), 387-404.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research,*, *36*(3), 347-387.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 User's Reference Guide*. Scientific Software International.

Kang, T., Cohen, A. S., & Sung, H. J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement*, *33*(7), 499-518.

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243-4258.

Kunina-Habenicht, O., Rupp, A. A., & Wilhelm, O. (2012). The impact of model misspecification on parameter estimation and item-fit assessment in log-linear diagnostic classification models. *Journal of Educational Measurement*, *49*(1), 59-81.

Larsson J (2021). *eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses*. R package version 6.1.1, https://CRAN.R-project.org/package=eulerr.

Lee, A. J. (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association. *The American Statistician*, *47*(3), 209-215.

Li, Y., Wen, Z., Hau, K. T., Yuan, K. H., & Peng, Y. (2020). Effects of cross-loadings on determining the number of factors to retain. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(6), 841-863.

Lin, J. (2016). *On The Dirichlet Distribution.* (Master's thesis, Queen's University).

Lindsay, B. (1988), Composite likelihood methods, in N. U. Prabhu, ed., 'Statistical Inference from Stochastic Processes', Providence RI: American Mathematical Society.

Liu, W., Zhang, B., Zhang, Z., Chen, B., & Zhou, X. H. (2015). A pseudo-likelihood approach for estimating diagnostic accuracy of multiple binary medical tests. *Computational Statistics & Data Analysis*, *84*, 85-98.

Markon, K. E., & Krueger, R. F. (2004). An empirical comparison of information-theoretic selection criteria for multivariate behavior genetic models. *Behavior genetics*, *34*(6), 593-610.

Martin, N., Pardo, L., & Zografos, K. (2019). On divergence tests for composite hypotheses under composite likelihood. *Statistical Papers*, *60*(6), 1883-1919.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models. *Measurement: Interdisciplinary Research and Perspectives*, *11*(3), 71-101.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713.

Maydeu-Olivares, A., & Joe, H. (2014). Assessing approximate fit in categorical data analysis. *Multivariate Behavioral Research*, *49*(4), 305-328.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115-132.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of mathematical psychology*, *44*(1), 190-204.

Myung, J. I., & Pitt, M. A. (2004). Model comparison methods. In *Methods in enzymology* (Vol. 383, pp. 351-366). Academic Press.

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. *Handbook of cognition*, 422-436.

Nguyen, T. T., & Sampson, A. R. (1985). The geometry of certain fixed marginal probability distributions. *Linear algebra and its applications*, *70*, 73-87.

Ostini, R., Finkelman, M., & Nering, M. (2014). Selecting among polytomous IRT models. In *Handbook of item response theory modeling* (pp. 303-322). Routledge.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in cognitive sciences*, *6*(10), 421-425.

Preacher, K. J. (2003). *The role of model complexity in the evaluation of structural equation models* (Doctoral dissertation, The Ohio State University).

Preacher, K. J. (2006). Quantifying parsimony in structural equation modeling. *Multivariate Behavioral Research*, *41*(3), 227-259.

Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P., & Green, R. (2019). Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Academic pathology*, *6*, 2374289519873088.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, *16*(1), 19-31.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied psychological measurement*, *9*(4), 401-412.

Ranalli, M., & Rocci, R. (2016). Mixture models for ordinal data: a pairwise likelihood approach. *Statistics and Computing*, *26*(1), 529-547.

Rashidi, H. H., Tran, N. K., Betts, E. V., Howell, L. P., & Green, R. (2019). Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. *Academic pathology*, *6*, 2374289519873088.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological review*, *107*(2), 358.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford press.

Seo, D. G., & Weiss, D. J. (2015). Best design for multidimensional computerized adaptive testing with the bifactor model. *Educational and Psychological Measurement*, *75*(6), 954-978.

Slavkovic, A. B., & Fienberg, S. E. (2009). Algebraic geometry of 2× 2 contingency tables. In *Algebraic and geometric methods in statistics* (pp. 63-82). Cambridge University Press.

Stine, R. A. (2004). Model selection using information theory and the MDL principle. *Sociological Methods & Research*, *33*(2), 230-260.

Teugels, J. L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. *Journal of multivariate analysis*, *32*(2), 256-268.

Thissen, D., & Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika*, *49*(4), 501-519.

Thompson, W. (2018). *Evaluating model estimation processes for diagnostic classification models* (Doctoral dissertation, University of Kansas).

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 5-42.

Varin, C., & Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, *92*(3), 519-528.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological methods*, *12*(1), 58.

Xi, N. (2011). *A composite likelihood approach for factor analyzing ordinal data* (Doctoral dissertation, The Ohio State University).

Xi, N., & Browne, M. W. (2014). Contributions to the Underlying Bivariate Normal Method for Factor Analyzing Ordinal Data. *Journal of Educational and Behavioral Statistics*, *39*(6), 583-611.

Xu, X., & Reid, N. (2011). On the robustness of maximum composite likelihood estimate. *Journal of Statistical Planning and Inference*, *141*(9), 3047-3054.

Yu, L., Wang, S., & Lai, K. K. (2005). An integrated data preparation scheme for neural network data analysis. *IEEE Transactions on Knowledge and Data Engineering*, *18*(2), 217-230.