6-21-2024

Factors Impacting the Empirical Identification of the Bifactor IRT Model of Rating Data

Wenya Chen
Loyola University of Chicago Graduate School

Recommended Citation

Chen, Wenya, "Factors Impacting the Empirical Identification of the Bifactor IRT Model of Rating Data" (2024). *Dissertations*. 4086.
https://ecommons.luc.edu/luc_diss/4086

LOYOLA UNIVERSITY CHICAGO

FACTORS IMPACTING THE EMPIRICAL IDENTIFICATION OF THE BIFACTOR IRT MODEL OF RATING DATA

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE GRADUATE SCHOOL
AS PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

PROGRAM IN RESEARCH METHODOLOGY

BY
WENYA CHEN
CHICAGO, IL
MAY 2024

ACKNOWLEDGMENTS

I extend my profound gratitude to my advisor Dr. Ken Fujimoto. His mentorship was more than I could have ever anticipated. The knowledge and expertise he shared about item response theory models and MCMC method have been invaluable. His unwavering dedication to ensuring I achieve my best, coupled with his patience in guiding me through the inevitable struggles of this process, have left an indelible mark on my academic and personal growth. I am truly privileged to have had him as my guide through this journey.

I would also like to thank the members of my dissertation committee, Dr. Meng-Jia Wu and Dr. Timothy O'Brien, for their invaluable insights, feedback, and time. Their rigorous reviews and suggestions greatly shaped this dissertation.

Last but not least, I want to express my heartfelt appreciation to my husband, Cenfu. His constant love, unwavering support, sacrifices, and belief in my aspirations have been the backbone of this journey. I could not have achieved this without him by my side.

Contents

ACKNOWLEDGMENTS ii

LIST OF FIGURES vi

LIST OF TABLES x

LIST OF ABBREVIATIONS xi

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 LITERATURE REVIEW 5

Dimensional Structures 6

Item Response Theory Model 11

Unidimensional IRT Model 13

Between-Item-Dimensionality IRT Model 14

Within-Item-Dimensionality IRT Model 15

Local Item Independence 18

Determining a bifactor structure 19

Challenges of Fitting a Bifactor Model 21

Parameter Estimation of the Bifactor Model 23

Full-Information Maximum Likelihood Estimation 23

Bayesian Inference 24

Open Questions 25

Conclusion 26

CHAPTER 3 METHOD 28

Study 1: Investigating the Similarity of the Within-Item Discriminations 29

Data Generation 30

Data Analysis 32

Analyses Based on FIML 34

Analyses Based on Bayesian Estimation 34

Prior Assigned to the Item Discriminations 35

Analytic Strategy 38

Study 2: Investigating the Interaction Effect Between Similarity and Magnitude of Item Discriminations 39

Study 3: Investigating the Interaction Effect Between Similarity and Item Targetedness 41

CHAPTER 4 RESULTS **46**

Results of Study 1 46

Acceptable Rates across the Three Estimation Methods 46

Parameter Recovery: Errors of the Parameter Estimates 49

Summary of Study 1 59

Results of Study 2 60

Full-Information Maximum Likelihood Estimation 60

Bayesian Method Using Less Informative Priors 65

Bayesian Method Using Adaptive Informative Priors 73

Summary of Study 2 82

Results of Study 3 83

Full-Information Maximum Likelihood Estimation 83

Bayesian Method Using Less Informative Priors 88

Bayesian Method Using Adaptive Informative Priors 98

Summary of Study 3 102

CHAPTER 5 DISCUSSION **106**

Limitations and Future Directions 110

Conclusions 111

REFERENCES **113**

VITA **121**

LIST OF FIGURES

1 Figure 1. Visualizations of different dimensional structures 7

2 Figure 2. Visualization of the dimensional structure investigated 31

3 Figure 3. Visualizations of the acceptable rates (ARs) under the three estimation methods, from Study 1 48

4 Figure 4. Boxplots summarizing the errors of the item discrimination estimates under full-information maximum likelihood (FIML) by sample size and ratio of Item 1's discriminations, from Study 1 51

5 Figure 5. Boxplots summarizing the errors of the item discrimination estimates under Bayesian method using less informative priors (LessInfo) by sample size and ratio of Item 1's discriminations, from Study 1 54

6 Figure 6. Boxplots summarizing the errors of the item discrimination estimates under Bayesian method using adaptive informative priors (AdptInfo) by sample size and ratio of Item 1's discriminations, from Study 1 57

7 Figure 7. Visualizations of the acceptable rates (ARs) under full-information maximum likelihood estimation (FIML), from Study 2 62

8 Figure 8. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under full-information maximum likelihood (FIML), from Study 2 64

9 Figure 9. Boxplots summarizing the errors of the item discrimination estimates for $N = 1,000$ under full-information maximum likelihood (FIML), from Study 2 66

10 Figure 10. Boxplots summarizing the errors of the item discrimination estimates for N = 2,000 under full-information maximum likelihood (FIML), from Study 2 67

11 Figure 11. Visualizations of the acceptable rates (ARs) under the Bayesian method using less informative priors (LessInfo), from Study 2 69

12 Figure 12. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under the Bayesian method using less informative priors (LessInfo), from Study 2 71

13 Figure 13. Boxplots summarizing the errors of the item discrimination estimates for $N = 1{,}000$ under the Bayesian method using less informative priors (LessInfo), from Study 2 72

14 Figure 14. Boxplots summarizing the errors of the item discrimination estimates for $N = 2{,}000$ under the Bayesian method using less informative priors (LessInfo), from Study 2 74

15 Figure 15. Visualizations of the acceptable rates (ARs) under the Bayesian method using adaptive informative priors (AdptInfo), from Study 2 75

16 Figure 16. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under the Bayesian method using adaptive informative priors (AdptInfo), from Study 2 78

17 Figure 17. Boxplots summarizing the errors of the item discrimination estimates for $N = 1{,}000$ under Bayesian method using adaptive informative priors (AdptInfo), from Study 2 80

18 Figure 18. Boxplots summarizing the errors of the item discrimination estimates for $N = 2{,}000$ under Bayesian method using adaptive informative priors (AdptInfo), from Study 2 81

19 Figure 19. Visualizations of the acceptable rates (ARs) under full-information maximum likelihood estimation (FIML), from Study 3 85

20 Figure 20. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under full-information maximum (FIML) estimation, from Study 3 86

21 Figure 21. Boxplots summarizing the errors of the item discrimination estimates for N = 1,000 under full-information maximum likelihood (FIML), from Study 3 89

22 Figure 22. Boxplots summarizing the errors of the item discrimination estimates for N = 2,000 under full-information maximum likelihood (FIML), from Study 3 90

23 Figure 23. Visualizations of the acceptable rates (ARs) under Bayesian method using less informative priors (LessInfo), from Study 3 92

24 Figure 24. Boxplots summarizing the errors of the item discrimination estimates for N = 500 under Bayesian method using less informative priors (LessInfo), from Study 3 93

25 Figure 25. Boxplots summarizing the errors of the item discrimination estimates for N = 1,000 under Bayesian method using less informative priors (LessInfo), from Study 3 96

26 Figure 26. Boxplots summarizing the errors of the item discrimination estimates for N = 2,000 under Bayesian method using less informative priors (LessInfo), from Study 3 97

27 Figure 27. Visualizations of the acceptable rates (ARs) under Bayesian method using adaptive informative priors (AdptInfo), from Study 3 99

28 Figure 28. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under Bayesian method using adaptive informative priors (AdptInfo), from Study 3 101

29 Figure 29. Boxplots summarizing the errors of the item discrimination estimates for N = 1,000 under Bayesian method using adaptive informative priors (AdptInfo), from Study 3 103

30 Figure 30. Boxplots summarizing the errors of the item discrimination estimates for N = 2,000 under Bayesian method using adaptive informative priors (AdptInfo), from Study 3 104

LIST OF TABLES

1 Table 1. The Values Used for the Item Discriminations and Category Intercepts to Generate the Data in Study 1 33

2 Table 2. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.5 in Study 2 41

3 Table 3. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.3 in Study 2 42

4 Table 4. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.1 in Study 2 43

5 Table 5. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.5 in Study 3 44

6 Table 6. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.3 in Study 3 44

7 Table 7. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.1 in Study 3 45

LIST OF ABBREVIATIONS

AdptInfo	Adaptive Informative Priors
AR	Acceptable Rate
ECV	Explained Common Variance
EM	Expectation-Maximization
FIML	Full-Information Maximum Likelihood
GRM	Graded Response Model
HMC	Hamiltonian Monte Carlo
ICC	Intraclass Correlation
IQR	Interquartile Range
IRT	Item Response Theory
LessInfo	Less Informative Priors
LII	Local Item Independence
LUL	Lower and Upper Limits
MCMC	Markov Chain Monte Carlo
MMLE	Marginal Maximum Likelihood Estimation
NUTS	No-U-Turn Sampler
RSES	Rosenberg’s Self-Esteem Scale
SD	Standard Deviation

CHAPTER 1

INTRODUCTION

Educational and psychological instruments are widely used to measure traits that are not directly measurable (Crocker & Algina, 1986), such as math ability and self-esteem. Depending on the design of the instrument and the nature of the measured trait, response data may represent different dimensional structures. One of the dimensional structures that has been increasingly discussed in the literature is the bifactor structure (e.g., Bolt, 2019; Rodriguez et al., 2016; Wang et al., 2015).

In a bifactor structure, a primary (or general) dimension exists to represent the trait of substantive interest, and this dimension impacts the responses to all items (i.e., the general dimension represents the dependencies across all items). Also, one or more secondary (or specific) dimensions exist to represent the other traits that additionally impact the responses to subsets of items (i.e., the specific dimensions represent the dependencies in subsets of items beyond the dependencies from the general dimension), with these secondary traits usually being specific domains of the primary trait or altogether irrelevant to the primary trait (such as wording method effects; Marsh et al., 2010).

To confirm whether a bifactor structure is represented in the response data, a bifactor item response theory (IRT) model (Gibbons & Hedeker, 1992) is needed. In a bifactor IRT model, the item discriminations distinguish the different sources that impact the item responses. Such a model, then, can determine how much the general and specific dimensions is represented in the data, which contributes to score validity (American Educational Research Association et al., 2014) and provides theoretical insights about the measured trait (details are provided in Chapter 2; Caspi et al., 2014; Marsh, 1996).

Unfortunately, applying the bifactor model to data has its challenges. One challenge is an empirical identification issue (Chen & Fujimoto, 2022; Stone & Zhu, 2015) that is seldom discussed in the literature. This issue occurs when an item's discriminations on the general and specific dimensions (or within-item discriminations) are similar in strength, making it difficult to obtain accurate estimates for those discriminations.

The empirical identification issue was alluded to in Stone and Zhu (2015), where they noted that when "multiple slope parameters are estimated for each item, the likelihood surface may have multiple equivalent modes when the slope parameters are similar in size" (Stone & Zhu, 2015, p. 165), leading to the estimates being indeterminate. Such equivalent modes, then, could result in biased results that should not be interpreted. Recently, Chen and Fujimoto (2022) provided empirical evidence of this issue and concluded that within-item discriminations being similar in strength creates problems in estimating the bifactor model. More importantly, they demonstrated that software may not produce error messages to indicate that the results are impermissible when within-item discriminations being similar creates problematic estimates, thereby leading researchers to form inferences on results that are untrustworthy.

The current evidence regarding the empirical identification issue was shown in only limited situations under full-information maximum likelihood (FIML) estimation method. The extent to which the within-item discriminations have to be similar before estimation issues arise and whether the similarity depends on sample size, strength of the item discriminations, and item targetedness (i.e., how well the items' response categories are targeted to the respondents) are unclear. Also, whether the empirical identification issue occurs under other estimation methods is unknown.

As researchers are using the bifactor model more frequently, being able to apply the model without having concerns about the parameter estimates is critical. If the empirical identification issue with the bifactor model occurs during their analysis without the researchers realizing it is happening, then inaccurate results may be reported and interpreted, which would mislead researchers into a false sense of score validity and misguide their theoretical conclusions about the measured trait. Thus, a thorough investigation on the empirical identification issue is needed, which will inform researchers as to when they should interpret their findings or proceed with caution.

This dissertation fills the aforementioned void. Specifically, I used simulations to investigate how similar the within-item discriminations need to be before estimation issues arise and whether an interaction effect exists between the within-item discriminations being similar and other factors, with these factors including sample size, magnitude of the within-item discriminations, and item targetedness. In addition, I examined the empirical identification issue under FIML and Bayesian estimation methods. The former is the dominant approach to estimating IRT models, with marginal maximum likelihood (MML) via the expectation-maximization (EM) algorithm being commonly used as the estimation technique (DeMars, 2013). Exploring the full-information method informs researchers on how the within-item discriminations being similar affects the discrimination estimates under a commonly used method. With regards to Bayesian estimation, it has been gaining momentum with IRT models (Fox, 2010), as it incorporates prior information that could be helpful in obtaining more accurate results in situations where FIML fails (e.g., Fujimoto, 2019). By examining the Bayesian method, I would show whether the empirical identification issue is a problem that occurs under other methods and whether assigning priors to item discriminations could prevent the issue from happening. If Bayesian estimation does not produce better results than FIML, then I would show that the empirical identification issue is a concern for Bayesian methods as well and assigning priors does not help in estimating the discrimination parameters when they are similar in strength. If Bayesian estimation produces more accurate results, then I would provide a solution to the empirical identification issue of the bifactor model.

In the next chapter, I provide a conceptual overview of three different types of dimensional structure, including unidimensional, between-item-dimensionality, and bifactor structures. I then introduce the technical details of the IRT models that can be used to confirm these dimensional structures, followed by a discussion of a key assumption of the IRT models (i.e., local item independence) and the approaches to determining a bifactor structure. Thereafter, I discuss the challenges of fitting a bifactor model and the two estimation approaches (i.e., FIML and Bayesian estimation) I focused on. Chapter 2 ends with the open questions I plan to answer in this dissertation. Then, in Chapter 3, I provide the details of the method I used to investigate the conditions that lead to the empirical identification issue. In Chapters 4 and 5, I focus on the findings, and discuss the significance and limitations of my findings, respectively.

CHAPTER 2

LITERATURE REVIEW

Educational and psychological instruments (or tests) are often designed to measure latent traits having different dimensional structures, leading to different patterns of the dependencies being displayed in the responses across the items. Investigating the pattern of the dependencies in the item responses, then, allows researchers to determine the dimensional structure represented in the data. Going through the process of identifying the pattern is important because it has implications for score validity as outlined in *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014) and can contribute theoretical insight about the measured trait (e.g., Caspi et al., 2014; Marsh, 1996).

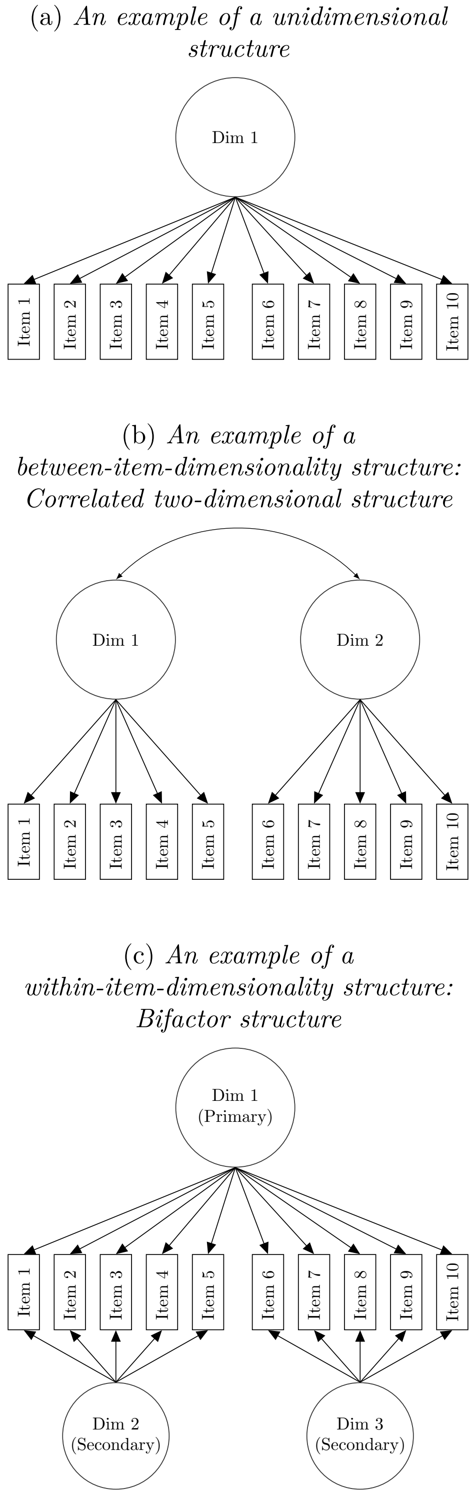
Unfortunately, confirming certain dimensional structures (i.e., confirming certain patterns of the dependencies in the data) could be challenging, leading to inaccurate estimates of the psychometric properties of the data. If researchers interpret the inaccurate results without realizing the inaccuracies, then they would obtain a false sense of score validity and possibly make misleading statements about the theoretical trait being measured.

For this dissertation, I focus on the challenges of confirming the bifactor structure. To get a better understanding of such challenges, next, I start by reviewing two widely-represented dimensional structures—unidimensional and between-item-dimensionality structures—and providing a conceptual overview of the item response theory (IRT) models related to these structures. The reason theses structures are reviewed first is that the bifactor structure can be viewed as an extension of them. I then review the bifactor structure and provide a conceptual overview of the corresponding IRT model, followed by a discussion on why confirming a bifactor structure could be more challenging than confirming other types of structures. The technical details of IRT models are provided thereafter.

Dimensional Structures

The unidimensional structure is the simplest structure that could be represented in data. Figure 1a is a visualization of such a structure represented in data arising from an instrument consisting of ten items (i.e., Item 1 to Item 10, which are represented in the rectangles). All items are related to (or discriminate on) a single common dimension (i.e., Dim 1, which is represented in the circle), as indicated by all items having arrows from Dim 1. Such a structure conveys that the measured latent trait consists of a single dimension, through which the dependencies in the responses across the items can be fully determined. One example of when a unidimensional structure may be represented in data is self-esteem data as measured by the Rosenberg's Self-Esteem Scale (RSES; Rosenberg, 1965). The RSES contains 10 items intended to measure self-esteem. Thus, if the scale measures only this latent trait, then only a unidimensional IRT model is needed for the data—that is, a model that accounts for the dependencies in the responses with a single dimension (Maydeu-Olivares & McArdle, 2005; Reise et al., 2014).

However, a single dimension is rarely able to account for all the dependencies in the data (Reckase, 2009; Reise et al., 2014), leading to the necessity of representing the latent trait with multiple dimensions. The two common forms of multidimensionality are between-item- and within-item- dimensionality (Hartig & Höhler, 2008; Rauch & Hartig, 2010). In a between-item-dimensionality structure (e.g., the structure displayed in Figure 1b), more than one dimension exists (e.g., Dim 1 and Dim 2 represented in the circles in Figure 1b), with each item measuring only one of the dimensions and the dimensions being allowed to correlate. Thus, in Figure 1b, Dim 1 accounts for the dependencies in the responses related to a set of items (i.e., Items 1 through 5), and Dim 2 accounts for the



(a) *An example of a unidimensional structure*

(b) *An example of a between-item-dimensionality structure: Correlated two-dimensional structure*

(c) *An example of a within-item-dimensionality structure: Bifactor structure*

Figure 1. Visualizations of different dimensional structures. In each of the figures, latent trait dimension(s) is (are) represented in circle(s) and items are represented in rectangulars. An arrow pointing from a dimension to an item indicates that the item discriminates on that dimension. A curve with double arrows that links two dimensions means the dimensions are correlated.

dependencies in the responses for another set of items (i.e., Items 6 through 10). One example of this type of structure can be seen in math anxiety, which can consist of two related forms—learning math anxiety and math evaluation anxiety (Hopko, 2003). Accordingly, the latent trait of math anxiety can be represented by two correlated dimensions to account for the dependencies in the subsets of items and also reflect the relationship between the two forms of math anxiety, leading to a structure that is more complex than a unidimensional structure. Of course, a between-item-dimensionality structure can have more dimensions, with each dimension containing more or fewer items than the structure displayed in Figure 1b. To determine whether such a between-item-dimensionality structure is represented in data, a between-item-dimensionality IRT model would be necessary—that is, an IRT model in which each individual has multiple abilities to account for the dependencies in the data, with the abilities being allowed to correlate.

Even though the between-item-dimensionality structure is useful to account for more complex patterns of the dependencies represented in the data than the unidimensional structure, it may not be appropriate in some situations. For example, one situation where a between-item-dimensionality structure may not be appropriate is when the dimensions are highly, but not perfectly, correlated (i.e., large shared variance among the dimensions), suggesting that a common dimension could exist to explain the dependencies in the responses across all items. Simply assuming a unidimensional structure in such a scenario, however, may still result in unexplained dependencies in subsets of items. Another situation in which the between-item-dimensionality is not ideal is when researchers settle on a two-dimensional structure for data from instruments that consist of positively and negatively phrased items. For instance, a correlated two-dimensional structure is frequently examined to represent the positively and negatively worded items of the RSES (e.g., Alessandri et al., 2015; Donnellan et al., 2016; Salerno et al., 2017). However, separating self-esteem into positive and negative forms of self-esteem could be misleading, as positive and negative wordings are methodological artifacts that have been widely used to prevent response bias (Carmines & Zeller, 1979; Marsh et al., 2010) and do not reflect distinct aspects of self-esteem. In other words, the dimensions of positive self-esteem and negative self-esteem mix up the trait of substantive interest (i.e., the self-esteem) and the wording method effect, leading to score validity being questionable in this case. A more appropriate structure for the aforementioned situations could be a nested structure (or within-item-dimensionality structure), of which the bifactor structure is a typical case.

A bifactor structure is for when multiple sources of dependencies are represented in the data, with one general source leading to dependencies across all item responses and extra sources leading to dependencies in the responses to subsets of items that are beyond the dependencies from the general source. The general source is represented by a primary (or general) dimension in the bifactor structure, and the extra sources are represented by secondary (or specific) dimensions. An example of such a structure is displayed in Figure 1c, in which the general source of dependencies is represented by Dim 1 (i.e., the primary dimension), indicated by all items having arrows from it; the additional sources of dependencies are represented by Dim 2 and Dim 3 below the items (i.e., the secondary dimensions), and each one of these has arrows pointing to only a subset of items. Thus, each item has two arrows pointing to it, indicating that two different dimensions influence the responses to the items (i.e., two dimensions are represented in each item response).

A bifactor structure could be represented in the data for different reasons. One reason could be that a primary dimension is being measured that overlies a few specific domains. Certain mental disorders could be represented by a bifactor structure. For example, in psychopathology data, a general psychopathology dimension should account for the dependencies across all diagnostic symptom items, and the specific dimensions may represent underlying styles of psychopathology (e.g., externalization and internalization) that account for additional dependencies that result from different ways of processing the symptoms (e.g., Caspi et al., 2014).

Characteristics of the way an instrument is developed could also lead to a bifactor structure being represented in the data. In educational tests, items are usually grouped into testlets, which is when a subset of items are linked by a common stimulus (DeMars, 2012). For example, in a math test, some items could be equation-oriented and other items could be word problems. The latter type of items, however, may partially measure students' reading ability, an ability irrelevant to math ability. Ignoring the dependencies resulting from "reading" could be problematic—as would be the case if a unidimensional structure is assumed for the data—because then any overall math ability score would actually represent math ability (the focus of a math test) and reading ability (which is irrelevant to math ability).

A bifactor structure could also be represented in the data when the psychological instruments used to gather the data consist of items written in different polarities. One example of this type of instrument is the RSES (Rosenberg, 1965), which I noted earlier. The RSES measures self-esteem using five positively phrased and five negatively phrased items. Including items worded in different polarities could induce a wording method effect (Marsh et al., 2010; Michaelides et al., 2016), leading to additional dependencies being represented in data. RSES data, then, may not be as simple as reflecting a single dimension, as the responses to the positively worded items could have additional dependencies beyond self-esteem, and likewise, the negatively worded items can as well. If a unidimensional structure is assumed for the RSES data, then any overall self-esteem score would actually represent an individual's self-esteem level as well as how they process positive and negative phrases, with the processing of positive and negative phrases being irrelevant to self-esteem. A unidimensional model, then, would create a score validity issue in this situation because the ability estimates would represent more than just the substantive dimension of interest.

Many reasons other than the ones I have discussed may lead to a bifactor structure being represented in the data, although these ones I noted tend to be the common reasons.

To confirm whether such a structure is represented in the data, a bifactor IRT model (Gibbons et al., 2007) is needed. This model can determine how much a general and a specific dimension is represented in each item's responses (i.e., the model separates the sources influencing the responses), through which one can better understand to what degree the dimension of substantive interest (i.e., the general dimension) is represented in the data and, in turn, obtain more accurate estimates of individuals' levels on this general dimension.

Even though the bifactor IRT model's ability to separate the different sources of dependencies provides useful details about how much the general and specific dimensions influence the responses, this feature could lead to an empirical identification issue (Chen & Fujimoto, 2022; Stone & Zhu, 2015). To provide specific details about this issue and which parameters are involved with it, next, I review the technical details of the IRT models based on the dimensional structures I have discussed. I then discuss a key assumption underlying the IRT models—local item independence (LII)—and explain the impact of specifying an inappropriate structure could have on LII. Thereafter, I review the approaches to determining a bifactor structure and the challenges with fitting a bifactor model to data, followed by a discussion on two estimators used to estimate the parameters of a bifactor model. This chapter ends with the open problems I aim to address. For the remainder of this dissertation, I turn to wording method effects for an example when necessary.

Item Response Theory Model

I start with a generic form of a multidimensional item response theory (IRT) model, followed by a discussion of three special cases of the generic form that correspond to the dimensional structures I have discussed.

The general IRT model I discuss is based on the graded response model (GRM; Samejima, 1997), a model that has been widely used to deal with ordered polytomous responses, such as those gathered from administration of educational and psychological rating scales. For the presentation of the IRT models and for general discussion, I use the following notations. Let i represent an individual (where $i = 1, 2, \ldots, N$, with N representing the total number of individuals). Let j represent the item index (where $j = 1, 2, \ldots, J$, with J representing the total number of items). Let d represent a dimension (where $d = 1, 2, \ldots, D$, with D representing the total number of dimensions). Let k represent a category score (where $k = 0, 1, 2, \ldots, m$, with m representing the highest score category).

A GRM based on a logit link function has the following general form for the cumulative probability of individual i endorsing category k or greater on item j:

$$P(Y_{ij} \geq k | \boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \tau_{jk}) = \frac{e^{\boldsymbol{\alpha_j} \boldsymbol{\theta_i}^{\intercal} - \tau_{jk}}}{1 + e^{\boldsymbol{\alpha_j} \boldsymbol{\theta_i}^{\intercal} - \tau_{jk}}}. \tag{1}$$

Accordingly, the conditional probability of a response of k is defined as follows:

$$P(Y_{ij} = k | \boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \tau_{jk}) = \begin{cases} 1 - P(Y_{ij} \geq k + 1 | \boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \tau_{j(k+1)}), \text{ if } k = 0, \\ P(Y_{ij} \geq k | \boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \tau_{jk}) - P(Y_{ij} \geq k + 1 | \boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \tau_{j(k+1)}), \text{ if } 0 < k < m, \\ P(Y_{ij} \geq k | \boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \tau_{jk}), \text{ if } k = m. \end{cases} \tag{2}$$

Regarding the parameters that make up these equations, $\boldsymbol{\theta_i}$ represents individual i's $1 \times D$ vector of latent trait dimensional positions (or abilities), or $\boldsymbol{\theta_i} = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{id}, \theta_{iD})$, with all elements being estimated. The population latent trait dimensional positions are typically assumed to follow a D-variate normal distribution, and formally

$$\boldsymbol{\theta} \sim \boldsymbol{\mathcal{N}}_D(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta), \tag{3}$$

where $\boldsymbol{\mu}_\theta$ is the mean vector and $\boldsymbol{\Sigma}_\theta$ is a $D \times D$ variance–covariance matrix, or

$$\boldsymbol{\Sigma}_\theta = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \ldots & \sigma_{1D} \\ \sigma_{21} & \sigma_{22}^2 & \ldots & \sigma_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{D1} & \sigma_{D2} & \ldots & \sigma_{DD}^2 \end{pmatrix}. \tag{4}$$

The elements along the main diagonal of $\boldsymbol{\Sigma}_\theta$ are variances and the elements off the main diagonal are covariances. $\boldsymbol{\alpha_j}$ is item j's $1 \times D$ vector of discriminations, or $\boldsymbol{\alpha_j} = (\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jd}, \alpha_{jD})$. Finally, τ_{jk} is the intercept for the kth category of item j, with $\tau_{j0} < \tau_{j1} < \ldots < \tau_{j(m+1)}$, $\tau_{j0} \equiv -\infty$, and $\tau_{j(m+1)} \equiv \infty$.

Different designs of $\boldsymbol{\theta}$, $\boldsymbol{\Sigma}_\theta$, and $\boldsymbol{\alpha_j}$ can be used to account for different patterns of the dependencies displayed in the response data, which also lead to different IRT models. As discussed, item j is also described by a set of category intercept parameters (i.e., τ_{jk}), which is determined solely by the number of response categories for item j and is irrelevant to the dimensionality of the model. Thus, τ_{jk} is not a focus of the following discussion on the IRT models. Next, I introduce the design matrices for the item discriminations and specifications for $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}_\theta$ under the unidimensional, between-item-dimensionality, and within-item-dimensionality IRT models to discuss the technical aspects of the structures I provided a conceptual overview of earlier. In doing so, the technical complexity of the structures are easier to see.

Unidimensional IRT Model

As noted earlier, the unidimensional model assumes the latent trait space consists of a single dimension. Accordingly, the vector of latent trait dimensional positions ($\boldsymbol{\theta_i}$) under a unidimensional model contains only one element, making $\boldsymbol{\theta_i}$ a scalar, or $\boldsymbol{\theta_i} = \theta_i$. The latent trait dimensional positions are typically assumed to be distributed as a univariate normal distribution, and formally

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma^2), \tag{5}$$

which indicates that the mean vector and variance–covariance matrix in Equation 3 reduce to scalars. α_j is also a scalar under a unidimensional model in that item j discriminates on a single dimension (e.g., in Figure 1a, each item has only one arrow from the latent dimension, or Dim 1). The item discrimination matrix for the J items becomes

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_{1,1} \\ \alpha_{2,1} \\ \vdots \\ \alpha_{j,1} \\ \alpha_{J,1} \end{pmatrix}, \tag{6}$$

with each α_j describing item j's ability to discriminate individuals with respect to their trait levels. In other words, it indicates the amount of information the item provides to the single dimension that represents the latent trait space, or how strongly related an item is to the dimension.

Between-Item-Dimensionality IRT Model

The between-item-dimensionality IRT model assumes the latent trait space consists of multiple correlated dimensions, with each item discriminating on only one dimension. The vector of latent trait dimensional positions for individual i, or $\boldsymbol{\theta_i}$, then, is a $1 \times D$ vector. Regarding the variance–covariance matrix (i.e, Equation 4), $\boldsymbol{\Sigma}_\theta$ is a $D \times D$ matrix, with all elements either below or above the main diagonal being estimated (e.g., $\sigma_{dd'}$ is estimated for all d and d', where $d > d'$ and $d > 2$ when the lower elements are estimated), indicating the dimensions can be correlated (e.g., in Figure 1b, the two latent trait dimensions are correlated, given the correlation is not estimated to be 0).

With respect to the vector of discriminations for item j, or $\boldsymbol{\alpha_j}$, it is also a $1 \times D$ vector. However, only one element in $\boldsymbol{\alpha_j}$ is nonzero because under a between-item-dimensionality specification, each item discriminates on only one of the D dimensions, meaning the other elements in $\boldsymbol{\alpha_j}$ are fixed to 0. For example, in the between-item-dimensionality structure depicted in Figure 1b, each item has only one arrow from one of the two latent trait dimensions, representing a single estimated discrimination related to each item. Thus, a between-item-dimensionality model that can be used to confirm such a structure would have the following design matrix for the item discriminations:

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_{1,1} & 0 \\ \alpha_{2,1} & 0 \\ \alpha_{3,1} & 0 \\ \alpha_{4,1} & 0 \\ \alpha_{5,1} & 0 \\ 0 & \alpha_{6,2} \\ 0 & \alpha_{7,2} \\ 0 & \alpha_{8,2} \\ 0 & \alpha_{9,2} \\ 0 & \alpha_{10,2} \end{pmatrix}, \tag{7}$$

with each row of the matrix representing an item's 1 × 2 vector of discriminations. Likewise, in a D dimensional situation, where $D > 2$, each row would have D elements, with only one element being nonzero and the other elements being 0 (i.e., $D - 1$ elements being fixed to 0).

Within-Item-Dimensionality IRT Model

The bifactor model is a type of within-item-dimensionality model, in which a general and specific dimensions are used to account for the different sources of dependencies in the data. Regarding the parameters of the bifactor model, the vector of latent trait dimensional positions for individual i, or $\boldsymbol{\theta_i}$, is a $1 \times D$ vector, similar to the between-item-dimensionality model but with D representing the general dimensional plus the secondary dimensions. The variance–covariance matrix is also a $D \times D$ matrix, but all elements off the main diagonal in $\boldsymbol{\Sigma}_\theta$ in Equation 4 are 0 (i.e., $\sigma_{dd'} = 0$ for all d and d', where $d \neq d'$), indicating that the dimensions are orthogonal to each other (e.g., in Figure 1c, no correlations exist among the three latent trait dimensions), which is commonly assumed because the general dimension should account for any correlations among the secondary dimensions.

The design matrix for the item discriminations is also different under the bifactor model in that more than one element in $\boldsymbol{\alpha_j}$ can be nonzero. For discussion's sake, I assume that item j's discrimination on the general dimension is the first element (i.e., $d = 1$) and the item's discriminations on the specific dimensions are the remaining elements (i.e., $d \geq 2$). In a bifactor model, for each item, the first element is nonzero because all items discriminate on the general dimension, and only one of the remaining elements at most is nonzero. For instance, in Figure 1c, each item has two arrows pointing to it, one from the general dimension and another from one of the specific dimensions, indicating that each item discriminates on two dimensions. Thus, the following would be the design matrix for the item discriminations related to the structure in Figure 1c:

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_{1,1} & \alpha_{1,2} & 0 \\ \alpha_{2,1} & \alpha_{2,2} & 0 \\ \alpha_{3,1} & \alpha_{3,2} & 0 \\ \alpha_{4,1} & \alpha_{4,2} & 0 \\ \alpha_{5,1} & \alpha_{5,2} & 0 \\ \alpha_{6,1} & 0 & \alpha_{6,3} \\ \alpha_{7,1} & 0 & \alpha_{7,3} \\ \alpha_{8,1} & 0 & \alpha_{8,3} \\ \alpha_{9,1} & 0 & \alpha_{9,3} \\ \alpha_{10,1} & 0 & \alpha_{10,3} \end{pmatrix}, \tag{8}$$

with each row of the matrix representing an item's 1×3 vector of discriminations. The magnitude of the two non-zero discriminations within an item indicates how much information the item contributes to their corresponding dimensions, which in turn indicates how much these dimensions are represented in the responses to that item. A discrimination of 0 indicates that the corresponding dimension is not represented in the responses.

Estimating two discriminations for each item is how the bifactor model separates the sources of dependencies from the secondary dimensions and the source of dependencies from the primary dimension. One example of when we would need to perform this separation is with Rosenberg's Self-Esteem Scale (RSES; Rosenberg, 1965) data. The RSES has five positively worded items and five negatively worded items. When using a bifactor model to analyze RSES data, two discrimination estimates are obtained for each item, with one of them (i.e., α_{j1}) solely representing how much "self-esteem" is represented in that item's responses and the other (i.e., $\alpha_{jd'}$, where d' is the other dimension on which the item discriminates) representing how much the wording method influence the responses. By doing this, we could remove the effects of the wording method on the item responses and thus obtain an ability estimate that represents only the self-esteem level for an individual. In addition, by separating the wording method effects and the effect of "self-esteem" on the item responses, we could gain theoretical insight about whether the individuals treat positive and negative phrased items in different ways, which could be of interest to some researchers (Marsh et al., 2010).

Unfortunately, having to separate the different sources of dependencies (i.e., how much the general and a specific dimension is represented) in each item's responses is what makes confirming a bifactor structure more complex than confirming a unidimensional or a between-item-dimensionality structure. With the latter two structures, the models have to only be concerned with one source of dependencies in each item (i.e., only one discrimination per item has to be estimated), whereas in a bifactor model, an item may have two nonzero discriminations being estimated, leading to the bifactor model being more challenging to apply.

Even though applying the bifactor model has its challenges, there are a couple reasons for fitting the model. One reason to fit the bifactor model relates to the assumption of local item independence, since violating this assumption has consequences. In addition, as previously noted, researchers may be interested in determining how much of the different sources of dependencies are represented in the data for theoretical reasons.

Local Item Independence

To get a better understanding of local item independence (LII) and the consequences of violating it, I now discuss the LII assumption in more detail. LII is one of the central assumptions of IRT models. According to Embretson and Reise (2013), local independence is obtained when the relations among the item responses are fully characterized by the IRT model, indicating that no remaining dependencies exist in the responses given the IRT model. Formally, the assumption requires that an individual's responses to different items are independent once conditioned on the ability level(s) and the item parameters (Chen & Thissen, 1997; Coulacoglou & Saklofske, 2017; Jiao et al., 2012; Liu & Thissen, 2012) and can be expressed mathematically as

$$P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) = \prod_{i=1}^{N} \prod_{j=1}^{J} P(y_{ij}|\boldsymbol{\theta_i}, \boldsymbol{\alpha_j}, \tau_{jk}), \tag{9}$$

where $P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau})$ represents the joint conditional probability of the data matrix and is also the likelihood of the data, given the latent trait positions $\boldsymbol{\theta}$, the discrimination matrix $\boldsymbol{\alpha}$, and the category intercept matrix $\boldsymbol{\tau}$.

As discussed, if an appropriate IRT model is correctly specified (i.e., the IRT model matches the dimensional structure underlying the data), then all the dependencies in the responses should be accounted for by the model, leading to LII in the data. Consequently, using an IRT model that underspecifies the number of dimensions (e.g., a unidimensional IRT model when the data represents a bifactor structure) would result in a violation of the LII assumption (i.e., local item dependence). Such a violation could lead to biased estimates of individuals' ability levels and item parameters, and inflate the measurement reliability of the latent trait estimates (e.g., DeMars, 2006; Jiao et al., 2012; Yen, 1984), as these are based on the likelihood of the data in which LII is met. For example, if a unidimensional model is used to analyze RSES data, it would be assumed that all the dependencies in the data are because of a single source—general self-esteem. The model, then, would ignore any dependencies in the data related to the wording method effect, which is well known to occur when the items are written in different polarities (Marsh et al., 2010; Michaelides et al., 2016), leading to a violation in LII. Yet, the likelihood of the data would be used to establish the psychometric properties of the data—a likelihood that violates an indispensable assumption of the model. To achieve LII with the RSES data, one may need to use a bifactor model.

Determining a bifactor structure

A common approach to evaluating whether a bifactor structure is represented in the data is by calculating the explained common variance (ECV; Sijtsma, 2009). The ECV has been widely used to quantify the degree to which a general trait accounts for the common variance among items. Thus, it provides insights into whether a unidimensional or bifactor structure is more appropriate for the data because it assesses how dominant the general dimension is in the response process relative to the specific dimensions. Formally, the ECV is defined as the ratio of variance explained by the general dimension to the total variance explained by the general and specific dimensions (Reise et al., 2010; Rodriguez et al., 2016) and is expressed as

$$\text{ECV} = \frac{\boldsymbol{\Sigma\lambda}^2_{\text{GEN}}}{\boldsymbol{\Sigma\lambda}^2_{\text{GEN}} + \boldsymbol{\Sigma\lambda}^2_{\text{SPE}_1} + \boldsymbol{\Sigma\lambda}^2_{\text{SPE}_2} + \ldots + \boldsymbol{\Sigma\lambda}^2_{\text{SPE}_{D-1}}}, \tag{10}$$

where $\boldsymbol{\Sigma\lambda}^2$ is the sum of the square of the standardized factor loadings (or λ_{jd}). The subscript "GEN" denotes the general dimension, and "SPE" denotes the specific dimension. Although these standardized factor loadings are typically used in confirmatory factor analysis, they can be easily derived through a straightforward transformation of the item discriminations (Paek et al., 2018), which is expressed as

$$\lambda_{jd} = \frac{\alpha_{jd}}{S}, \tag{11}$$

where S is a scaling constant equal to 1.7. Thus, the accuracy of the item discrimination estimates can influence the ECV, thereby affecting judgments about how strongly a bifactor structure is represented in the data. Specifically, a high ECV indicates that a large proportion of the explained variance in the responses is accounted for by the general dimension, which supports a strong presence of a single dimension and thus a unidimensional model might be a reasonable fit for the data; a low ECV suggests that the specific dimensions substantially contribute to explaining the variance in the responses, thus indicating that a bifactor model might be more suitable than a unidimensional model.

Another approach to assessing the presence of a bifactor structure involves the intraclass correlation (ICC) at the item level. The ICC indicates the proportion of the variance in the responses for item j that is attributable to the general dimension (Lee & Cho, 2017). Formally, the ICC for item j is expressed as

$$\text{ICC}_j = \frac{\alpha^2_{j,\text{GEN}}}{\alpha^2_{j,\text{GEN}} + \alpha^2_{j,\text{SPE}}}. \tag{12}$$

As indicated in Equation 12, the maximum value of ICC is 1, which implies that all the variance in the responses for item j is perfectly accounted for by the general dimension, with no contribution from the specific dimensions. However, in reality, ICC values typically fall between 0 and 1, primarily due to measurement error and individual differences. Thus, ICC values approaching 1 suggest that a unidimensional structure is strongly represented in the data, as the influence of the specific dimensions is minimal. Conversely, low ICC values indicate that a between-item-dimensionality structure is more appropriate, as the general dimension does not adequately account for the variance in the item responses. Moderate ICC values are indicative of a bifactor structure, where both the general and specific dimensions play substantial roles in accounting for the variance in the responses.

Both ECV and item level ICC highlight the importance of accurate estimation of the item discrimination parameters, as any bias in their estimation could potentially lead to incorrect conclusions regarding whether a bifactor structure is represented in the data. However, estimating the discrimination parameters of a bifactor model has its challenges, as I discuss next.

Challenges of Fitting a Bifactor Model

One of the challenges that exists with the bifactor model but has not been fully investigated is an empirical identification issue (Chen & Fujimoto, 2022; Stone & Zhu, 2015). This issue is different from the well-documented mathematical identification issue.

A mathematical identification issue occurs when multiple equivalent modes in the likelihood distribution exist for the item parameters, indicating more than one set of parameter estimates exist that lead to the same likelihood of the data. That is, no one unique solution exists to maximizing the likelihood. One example of the impact of multiple modes in the likelihood is that the signs of the item discrimination estimates can switch but still lead to the same likelihood. When all the item discriminations are supposed to be oriented in the same direction, they can all be either positive or negative. Depending on the sign of the item discriminations, the estimates for all the other parameters of the model will adjust. The issue occurs regardless of the number of dimensions specified for the IRT model (i.e., all the models I have discussed may be affected) or the sample size that contributed to the data.

The estimation issue arising from the multiple equivalent modes in the likelihood can be addressed by setting the location and metric of the latent trait scale underlying the IRT model (Bafumi et al., 2005; Embretson & Reise, 2013). Specifically, the location is often set by fixing the mean of the population ability distribution (i.e., all elements of $\boldsymbol{\mu}_\theta$ in Equation 3) to a specific value, typically 0. Regarding the metric, it is commonly set in one of two ways. One way is to fix the variances of the dimensional variance–covariance matrix (i.e., the elements along the main diagonal of Equation 4) to 1, and freely estimate all nonzero item discrimination parameters, with all the item discriminations restricted to be oriented in the same direction (usually positive). Another way to set the metric is to fix the first nonzero item discrimination related to each dimension (e.g., $\alpha_{1,1}$, $\alpha_{1,2}$, and $\alpha_{6,3}$ in Equation 8) to some value (typically 1) and freely estimate the remaining nonzero discriminations and the dimensional variances (Reckase, 2009; Reise & Haviland, 2005).

Once the location and the metric of the latent trait scale are set, a unique solution for the parameters should be estimatable.

In contrast, an empirical identification issue occurs when the model is mathematically identified but the parameters still cannot be accurately estimated because the data do not provide enough information (e.g., the sample size is too small or the quality of the data is poor). As noted in the Mplus user's guide, "Mixture models that are in theory identified can in certain samples and with certain starting values be empirically non-identified" (L. K. Muthén & Muthen, 2017, p. 526), indicating that the empirical identification may be determined by the amount of information that can be used in the estimation process. In this dissertation, I focus on the challenges of fitting a bifactor model under an empirical identification issue that may occur when more than one discrimination parameter is estimated for an item, and those discrimination parameters for the item are similar in strength (i.e, the within-item discriminations are similar, e.g., $\alpha_{1,1} \approx \alpha_{1,2}$ in Equation 8). In other words, the empirical identification issue occurs when the item level ICC is around .50 (i.e., about 50% of the variance in the item's responses is attributable to the general dimension and 50% to the specific dimension). For simplicity, I use the ratio of the item's discrimination on the general dimension to its discrimination on the specific dimension, or ratio $= \alpha_{j1}/\alpha_{jd}$ (where $1 < d \leq D$), to represent the similarity of an item's discriminations. A ratio close to 1 then indicates that the item's discriminations are very similar in strength, which in turn means that the ICC for that item is near .50.

Stone and Zhu (2015) noted that the reason within-item discriminations being similar (or ratio ≈ 1) could lead to an empirical identification issue is that "the likelihood surface may have multiple equivalent modes" (p. 165). However, Chen and Fujimoto (2022) suggested that the multiple modes were not exactly equal but were close in height to where the amount of information in the data could determine whether the parameter estimates corresponding to the greater mode could be reached. Therefore, when an item's discriminations are similar in strength, obtaining accurate estimates for those discrimination parameters may be difficult, particularly when conventional estimation methods are used, such as the marginal maximum likelihood estimation (MMLE) (Chen & Fujimoto, 2022) that estimates the parameters rely solely on the information in the data. The empirical identification issue that I am discussing is not a problem for the unidimensional and between-item-dimensionality models, as in these models, each item has only one discrimination to estimate.

Currently, there is only some evidence of the empirical identification issue arising from within-item discriminations being similar in the literature, as the issue has only been investigated in a very limited set of conditions under MMLE (Chen & Fujimoto, 2022). A more thorough investigation on the conditions in which the empirical identification issue could arise is missing, which is the focus of my dissertation. In addition, whether the empirical identification issue is a problem that occurs only under MMLE is unknown. This is of interest as well because other estimation methods can incorporate prior information, which could possibly resolve the empirical identification issue. Next, I discuss two estimation methods that I focus on.

Parameter Estimation of the Bifactor Model

Full-Information Maximum Likelihood Estimation

The full-information maximum likelihood (FIML) estimator I discuss is based on marginal maximum likelihood estimation (MMLE) via the expectation-maximization (EM) algorithm (Bock and Aitkin, 1981) method. This estimator is of interest because it is one of the most commonly used estimation methods for IRT models (DeMars, 2013). If the empirical identification issue appears in a wide range of conditions under FIML, then one should be cautious in using it when the within-item discriminations being similar is a possibility.

MMLE is an iterative estimation procedure, with the iterations successively improving the model parameter estimate by maximizing the marginal probability of a response set (Bock & Aitkin, 1981; Embretson & Reise, 2013; Forero & Maydeu-Olivares, 2009; Gibbons et al., 2007). Formally, this marginal probability is expressed as

$$P(\boldsymbol{Y}|\boldsymbol{\alpha}, \boldsymbol{\tau}) = \int_{-\infty}^{+\infty} P(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{13}$$

where $f(\boldsymbol{\theta})$ is a prior distribution assigned to the latent trait dimensional positions (or abilities), such as the multivariate normal distribution expressed in Equation 3. The parameter estimates for $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ are the values that maximize the above likelihood function. The item parameter estimates, in turn, can be used to estimate the individuals' ability levels, such as through the expected a posteriori method (Bock & Mislevy, 1982, Chen et al., 1998).

As Equations 13 shows, MMLE only requires assigning a prior distribution to the latent trait dimensional positions, and no other information other than the response data is needed to estimate the other parameters. Unfortunately, this could create problems in the estimation process when the within-item discriminations are similar in strength, as the information in the data alone may not be sufficient to differentiate the within-item discriminations.

Bayesian Inference

The Bayesian method is of interest because it has been increasingly applied in IRT modeling (Albert, 1992) and, more importantly, recent studies (e.g., Fujimoto & Neugebauer, 2020; Kieftenbeld & Natesan, 2012) have demonstrated the effectiveness of the Bayesian method in estimating item parameters in situations in which conventional methods are not appropriate.

In Bayesian estimation, all parameters are treated as coming from a probability distribution. Accordingly, Bayes' theorem states that the joint posterior distribution for the parameters is the likelihood of the data combined with the prior, or

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}|\boldsymbol{Y}) &= \frac{P(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) f(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau})}{P(\boldsymbol{Y})} \\ &= \frac{P(\boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) f(\boldsymbol{\theta}) f(\boldsymbol{\alpha}) f(\boldsymbol{\tau})}{P(\boldsymbol{Y})}, \end{aligned} \tag{14}$$

where $P(\boldsymbol{Y})$ is the marginal probability of the response data. The joint prior $f(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau})$ can be rewritten as $f(\boldsymbol{\theta})f(\boldsymbol{\alpha})f(\boldsymbol{\tau})$ when the priors for item parameters and latent ability are assumed to be independent.

The Bayesian method could be a potential solution to the empirical identification issue that has been demonstrated to exist under MMLE because a prior is assigned to all the parameters, including the item discriminations, which are the parameters that lead to the empirical identification issue I have been focusing on. Specifically, assigning priors to the item discriminations would provide more information to their estimation than relying on the data alone. The additional information from the priors could then help in the estimation of the within-item discriminations and, in turn, overcome the empirical identification problem that has been shown to appear under MMLE. Unfortunately, no studies have been conducted that investigated this possibility.

Open Questions

Confirming bifactor structures in the data has increased over the years, as many psychological traits being measured tend to follow a bifactor structure (e.g., Bornovalova et al., 2020; Hendy & Biderman, 2019; Murray et al., 2016; Yeo & Suárez, 2022). Unfortunately, using a bifactor IRT model to analyze data could be more challenging compared with the unidimensional and between-item-dimensionality IRT models, as the bifactor model separates the different sources of the dependencies represented in the data for each item (i.e., separates the contributions of the general and specific dimensions on the responses), whereas the other models I have described assume that the dependencies in each item's responses is from a single source. The process of separating the sources of the dependencies within the items is what could lead to an empirical identification issue for the bifactor model (Stone & Zhu, 2015), which could affect the quality of the evidence the model provides for score validity and could also mislead researchers about how strongly the general and specific dimensions are represented in the data, thereby affecting theoretical interpretation of the latent trait measured.

Unfortunately, to date, the different conditions in which the empirical identification problem arises are unclear. Specifically, there is a lack of evidence on how similar the within-item discriminations have to be before estimation issues arise under the bifactor model. Also, it is unclear whether an interaction effect exists between within-item discriminations being similar and some other factors that have been proven to impact the estimation of the IRT models, such as sample size (e.g., De Ayala, 1994; Forero & Maydeu-Olivares, 2009; Jiang et al., 2016; Kose & Demirtasli, 2012) and item targetedness (e.g., Linacre et al., 2002; Xia & Yang, 2018). In addition, the impact of magnitude of the within-item discriminations on the empirical identification of the bifactor model has not been examined.

Filling in these knowledge gaps in the literature is the focus of my dissertation. Specifically, I used simulations to investigate some of the potential conditions necessary for within-item discriminations to create estimation problems, that is, how similar the within-item discriminations need to be and the factors that affect how similar the within-item discriminations have to be. I explored the effects of within-item discriminations being similar and all the related factors across the methods of FIML and Bayesian estimation. If the Bayesian approach can produce more accurate and stable parameter estimates than FIML, then I will have provided a solution to the empirical identification issue of the bifactor model that I have discussed. The prior distributions for the item parameters are discussed in the next chapter.

Conclusion

The bifactor model has been increasingly used by researchers, as it can contribute to score validity and provide theoretical insights about the measured trait. Thus, it is critical that one is able to perform a bifactor analysis without having to be concerned about the estimates. Otherwise, inaccurate results may be reported and interpreted by researchers, resulting in false evidence of support for score validity. Also, theoretical conclusions about the measured trait will be drawn based on inaccurate results, which may lead the readers to perceive the measured trait in a misleading way.

Fully understanding the conditions that may lead to the empirical identification issue under different estimators will alert researchers to be cautious when certain situations appear and also inform them about how large of a sample size (i.e., how much information in the data) is needed to obtain accurate estimates in their circumstances.

The remainder of this dissertation is organized as follows. In Chapter 3, I discuss the details of the simulation studies that I conducted to investigate the conditions that lead to the empirical identification issue for the bifactor model. In Chapter 4, I present the findings of the simulation studies. In Chapter 5, I discuss the findings, significance, and limitations of my findings.

CHAPTER 3

METHOD

In the previous chapter, I introduced an empirical identification issue of the bifactor IRT model, one that occurs when an item's discriminations on the general and specific dimensions are approximately equal (or $\alpha_{j1}/\alpha_{jd} \approx 1$, where $1 < d \leq D$). Such similarity of the within-item discriminations may lead to difficulties in obtaining accurate estimates for those parameters, which in turn may result in scores that lack validity and inaccurate theoretical conclusions about the measured trait. To date, it is unknown how similar the within-item discriminations must be before this issue arises. It is also unclear whether the extent of the similarity of the within-item discriminations that lead to estimation issues depends on other factors such as sample size, item targetedness, and magnitude of the item discriminations. Additionally, whether the estimation method used matters in the empirical identification issue has not been investigated. This dissertation fills these gaps in the literature through a series of simulation studies.

In this chapter, I provide the details of the simulation studies that were conducted to investigate the empirical identification issue with the bifactor model. The first of these studies (i.e., Study 1) aimed to investigate how similar the within-item discriminations must be (i.e., how close to 1 the ratio α_{j1}/α_{jd}, where $1 < d \leq D$, has to be) before the empirical identification issue arises and whether the similarity depends on sample size. Study 1 established a baseline, as it investigated the empirical identification issue under conditions in which the magnitude of the item discriminations and item targetedness were ideal (i.e., only the within-item discriminations were manipulated). If estimation issues appeared in Study 1, it would also demonstrate that the empirical identification issue exists under fairly ideal conditions.

Study 2 examined whether the similarity of the within-item discriminations that leads to the empirical identification issue depends on the magnitude of the item discriminations. It explored the conditions in which item targetedness was ideal but the similarity of the within-item discriminations and the magnitude of the item discriminations were manipulated. Comparing the results across Studies 1 and 2 indicates whether the magnitude of the item discriminations plays a factor in the empirical identification issue.

Study 3 further inspected whether the similarity of the within-item discriminations depends on item targetedness. Therefore, the magnitude of the item discriminations were ideal but the similarity of the within-item discriminations and item targetedness were manipulated. Similarly, comparing the results across Studies 1 and 3 demonstrates whether an interaction effect exists between similarity of the within-item discriminations and item targetedness.

Bifactor models based on full-information maximum likelihood (FIML), Bayesian method using less informative priors (LessInfo), and Bayesian method using adaptive informative priors (AdptInfo) were used to analyze each data replicate in the three studies to investigate whether the estimation method used matters in the empirical identification issue of focus for this dissertation. The results based on FIML served as the baseline, as it did not assume any prior information about the item parameters. Comparing the results based on FIML and Bayesian method demonstrates whether assigning prior information to the item discrimination parameters can be a solution to the empirical identification issue. Regarding the two Bayesian approaches, they differ with respect to the prior assigned to the item discriminations to establish a pattern of how prior information assigned to the discriminations may impact the estimates.

Study 1: Investigating the Similarity of the Within-Item Discriminations

I conducted a simulation study with a 3 × 3 design (sample size by ratio of within-item discriminations). The sample sizes included 500, 1,000, and 2,000 to determine whether the empirical identification issue depends on sample size. Sample sizes smaller than 500 were not tested because FIML has been shown to produce biased estimates in these conditions (Drasgow, 1989; Forero & Maydeu-Olivares, 2009; Reiser & VandenBerg, 1994). In addition, smaller sample sizes risked the possibility that certain response categories would not be represented in the data. Thus, focusing on moderate to large sample sizes ensured that sample size was not the reason for any estimation difficulties and that all response categories were represented. The ratios of the within-item discriminations (i.e., α_{j1}/α_{jd}, where $1 < d \leq D$) included 1.5, 1.3, and 1.1 to examine how similar the within-item discriminations must be before the empirical identification issue arises. These ratios were selected based on preliminary analyses, which represented item-level ICCs of .69, .63 ,and .55, respectively. A ratio of 1 or smaller (i.e., the discrimination on the secondary dimension is equal to or greater than the discrimination on the primary dimension) were not examined because the primary dimension usually represents the latent trait of central interest and the secondary dimensions represent nuisance traits (Toland et al., 2017), indicating that an item's discriminatory power on the primary dimension should be larger than its discriminatory power on the secondary dimension.

The bifactor structure to which I generated data is the same as that used in Chen & Fujimoto (2022) that initially demonstrated the existence of the empirical identification issue. Specifically, it is the dimensional structure used to control for wording method effects in the Rosenberg's Self-Esteem scale (RSES; Rosenberg, 1965), with 10 items discriminating on the primary dimension (Dim 1), three of the items additionally discriminating on a specific dimension (Dim 2), and five items additionally discriminating on another specific dimension (Dim 3) (Reise et al., 2016). A visualization of this bifactor structure is in Figure 2.

Data Generation

Fifty data sets were generated for each simulation condition, with each data set resembling 4-point ratings to 10 items (similar to RSES data). The latent trait dimensional

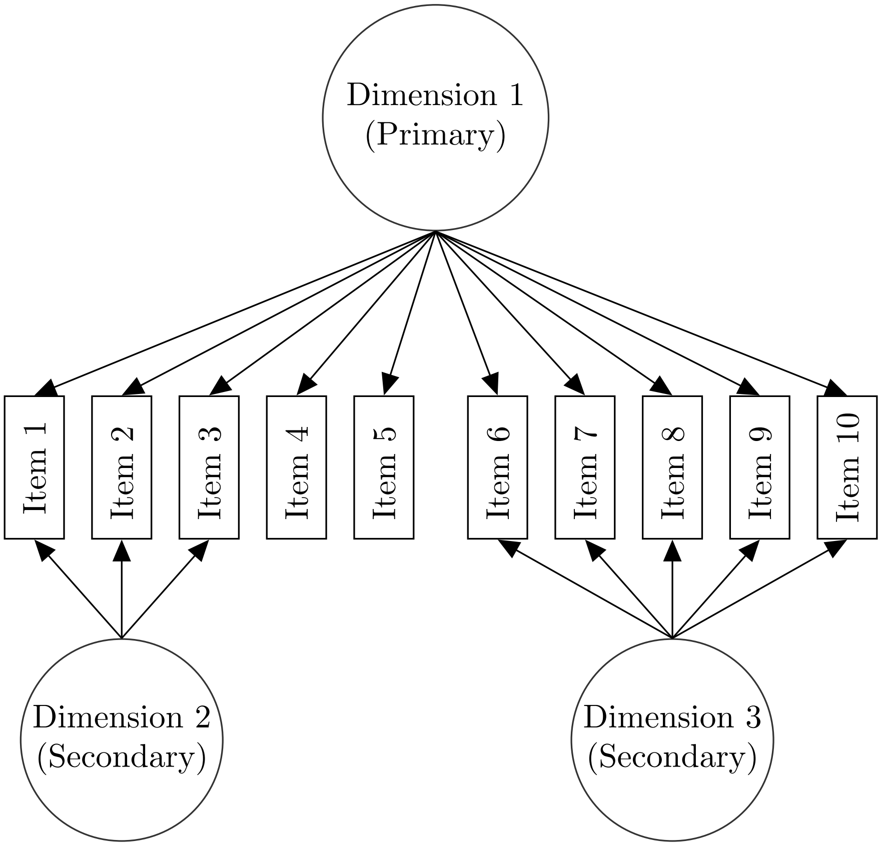


Figure 2. Visualization of the dimensional structure investigated. In the figure, latent trait dimensions are represented in circles and items are represented in rectangulars. An arrow pointing from a dimension to an item indicates that the item discriminates on that dimension.

positions ($\boldsymbol{\theta}_i$) were randomly drawn from a D-variate normal distribution with a mean vector of 0s and an identity matrix for its variance–covariance matrix, $\boldsymbol{\mathcal{N}}_D(\mathbf{0}, \mathbf{I})$. Regarding the item intercepts, τ_{j1} were randomly drawn from a uniform distribution over the interval $(-3.5, -0.5)$, or $U(-3.5, -0.5)$, τ_{j2} were randomly drawn from $U(\tau_{j1} + 1.5, \tau_{j1} + 2.5)$, and τ_{j3} were randomly drawn from $U(\tau_{j2} + 1.5, \tau_{j2} + 2.5)$. The item intercepts obtained through this process ensured that each category was well-represented in the data, given the latent trait dimensional positions.

For the discrimination parameters, the generation values for the item discriminations on the general dimension were obtained by randomly drawing values from $U(1.25, 2.50)$; the item discriminations on the specific dimensions were obtained by taking the items' corresponding discriminations on the primary dimension and dividing them by randomly drawn values from $U(1.5, 1.8)$. These two processes ensured that the resulting discrimination values were distinctly different as well as reasonable in strength, given the latent trait dimensional positions.

Those processes discussed above composed the first stage of the data generation, which led to an ideal situation in that the item responses would represent a bifactor structure without any estimation issues arising because the resulting within-item discrimination values were, as noted earlier, distinct for each item and no categories were under- or over- represented across the items. The second stage of the data generation involved manipulating the resulting values from the first stage to reflect the simulation conditions, which I discuss next.

To investigate the impact of within-item discriminations being similar on parameter estimation, I adjusted Item 1's discrimination on the specific dimension (i.e., $\alpha_{1,2}$) so that it met the ratio condition relative to the item's discrimination on the primary dimension (i.e., $\alpha_{1,1}$). Therefore, $\alpha_{1,2} = \alpha_{1,1}/c$, where c is 1.5, 1.3, or 1.1 depending on the ratio condition. As discussed previously, a ratio of 1.1 leads to more similar within-item discriminations; as the ratio increases, the within-item discriminations become more distinct. I manipulated only one item to ensure that any estimation issue that might occur in this item was not due to similar within-item discriminations in another item and to establish a clear pattern of how the ratio size (i.e., how similar the within-item discriminations were) led to an empirical identification issue in this item and how sample size might influence the necessary ratio size. The data generation values for the item discriminations and category intercepts are in Table 1.

Data Analysis

The bifactor model I fitted is a special case of the general form of the multidimensional IRT model that is expressed in Equations 1 and 2. Three versions of the bifactor model were investigated. The first was based on full-information maximum

Table 1. The Values Used for the Item Discriminations and Category Intercepts to Generate the Data in Study 1

					Ratio = 1.5		Ratio = 1.3		Ratio = 1.1	
	Category intercepts			Primary dimension	Secondary dimensions		Secondary dimensions		Secondary dimensions	
Item	τ_{j1}	τ_{j2}	τ_{j3}	α_{j1}	α_{j2}	α_{j3}	α_{j2}	α_{j3}	α_{j2}	α_{j3}
1	−2.34	−0.53	1.63	2.29	1.53		1.76		2.08	
2	−0.91	0.91	2.53	2.22	1.38		1.38		1.38	
3	−1.05	1.09	2.61	1.67	0.96		0.96		0.96	
4	−2.48	−0.47	1.75	1.91						
5	−1.65	−0.05	2.25	2.15						
6	−1.89	−0.003	1.88	1.64		1.04		1.04		1.04
7	−2.46	−0.26	1.76	1.96		1.19		1.19		1.19
8	−0.96	1.33	3.62	2.34		1.49		1.49		1.49
9	−1.72	0.35	2.53	1.87		1.17		1.17		1.17
10	−2.48	−0.87	1.56	2.26		1.34		1.34		1.34

Note. An empty space indicates a value of 0. The category intercepts and the items' discriminations on the primary dimension remain the same for all ratio conditions in Study 1.

likelihood estimation (Bifactor-FIML), which served as the baseline, as it did not assume any prior information on the item parameters. The other two were based on Bayesian estimation, with one using less informative priors (Bifactor-LessInfo) and the other using adaptive informative priors (Bifactor-AdptInfo). These Bayesian versions were included to demonstrate whether adding prior information can resolve the empirical identification issue, and if so, whether a greater amount of prior information is more effective in resolving this issue. As noted earlier, these two Bayesian approaches differ in terms of the prior assigned to the item discriminations so that a pattern can be established on how different prior information may impact the estimates.

All three versions of the bifactor model were used to analyze each data replicate. The Mplus software (L. Muthén & Muthén, 2016) was used to perform all analyses involving the model based on FIML (i.e., Bifactor-FIML), and RStan (the R interface of Stan; Team et al., 2016) was used to perform all analyses involving the models based on Bayesian estimation (i.e., Bifactor-LessInfo and Bifactor-AdptInfo). Next, I introduce the specifics of the three versions of the bifactor model.

Analyses Based on FIML

When full-information maximum likelihood (FIML) was used to estimate the model, 20 quadrature points were used. Also, the default convergence criterion in Mplus were used. That is, the estimation process stoped when the incremental improvement in the minimization function reached 1E−6. The dimensional positions were assumed to be multivariate normally distributed, and formally

$$\boldsymbol{\theta}_i \sim \boldsymbol{\mathcal{N}}_D(\mathbf{0}, \mathbf{I}). \tag{15}$$

As noted earlier, the mean vector of 0s and the variances being set to 1 (as the identity matrix conveys) establish the location and metric of the underlying scale, respectively. The identity matrix also indicates that all dimensions are set to be orthogonal to each other, as conventionally done for the bifactor model.

Analyses Based on Bayesian Estimation

As discussed in Chapter 2, the joint posterior distribution for the parameters under the Bayesian method is as follows:

$$\begin{aligned} f(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau} | \boldsymbol{Y}) &= \frac{P(\boldsymbol{Y} | \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) f(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau})}{P(\boldsymbol{Y})} \\ &= \frac{P(\boldsymbol{Y} | \boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\tau}) f(\boldsymbol{\theta}) f(\boldsymbol{\alpha}) f(\boldsymbol{\tau})}{P(\boldsymbol{Y})}. \end{aligned} \tag{16}$$

The posterior distribution was estimated using an advanced dynamic Hamiltonian Monte Carlo (HMC) algorithm based on the No-U-Turn sampler (NUTS). NUTS dynamically determines the optimal number of steps during the sampling process (Hoffman, Gelman, et al., 2014), thereby enhancing the efficiency of the standard HMC algorithm. The priors for the item parameters and latent ability were assumed to be independent.

Different priors assigned to the parameters may result in different posterior distributions, which can possibly lead to estimates based on one prior distribution being more accurate than estimates based on another. Next, I describe the two sets of prior distributions that were used — the less informative priors and the adaptive informative priors. These two sets of priors differ only in terms of the prior assigned to the item discriminations. The reason only the prior for the item discriminations are set differently is that the item discriminations are the parameters that are affected by the empirical identification issue. Thus, keeping the priors assigned to the other parameters the same and adjusting only the prior assigned to the item discriminations could establish a clear pattern of how the different priors may result in more accurate item discrimination estimates. I am going to describe the priors for the dimensional positions and category intercepts first, and then discuss how the priors differ for the item discriminations.

Priors Assigned to the Dimensional Positions and the Category Intercepts. The prior assigned to the dimensional positions (i.e., $f(\boldsymbol{\theta})$ in Equation 16) was a multivariate normal distribution that is expressed in Equation 15, with the mean vector of 0s and the elements of the main diagonal of the variance-covariance matrix being fixed to 1. Each category intercept was assigned a univariate normal distribution with a mean of 0 and a standard deviation (*SD*) of 10, or

$$\tau_{jk} \sim \mathcal{N}(0, 10), \tag{17}$$

for all j, and for all k ranging from 1 to 3 with the restriction $\tau_{j1} < \tau_{j2} < \tau_{j3}$.

Prior Assigned to the Item Discriminations

Less Informative Prior. The less informative prior assigned to the item discriminations was a lognormal distribution

$$\alpha_{jd} \sim \text{logn}(\mu_{\alpha_d}, \sigma_{\alpha_d}) \tag{18}$$

for all j and d corresponding to the estimated discriminations, and the mean and *SD* of the lognormal distribution were fixed to 0 and 1, respectively, or

$$\mu_{\alpha_d} = 0.00 \text{ for all } d, \tag{19}$$

and

$$\sigma_{\alpha_d} = 1.00 \text{ for all } d. \tag{20}$$

This prior is less informative compared with the prior I describe next. This less informative prior places support on a narrower range of values than non-informative priors (e.g., a uniform distribution) but a wider range than some stronger priors that may alter results (e.g., strong informative priors that may be needed when the information from the data is not sufficient for parameter estimation). Specifically, a mean of 0 and a *SD* of 1 for the lognormal distribution mainly support values that range from 0.14 to 7.10, which means that about 95% of the values from this lognormal distribution fall within this range.

Adaptive Informative Prior. Under the Bayesian method based on adaptive informative prior, the prior distribution assigned to the item discrimination was a lognormal distribution as expressed in Equation 18, with the *SD* of the lognormal distribution fixed to 0.50, or

$$\sigma_{\alpha_d} = 0.50 \text{ for all d}, \tag{21}$$

and the mean of the lognormal distribution was assigned a hyperprior of

$$\mu_{\alpha_d} \sim \begin{cases} \mathcal{N}(0, 0.40), & \text{when } d \text{ represented a primary dimension,} \\ \mathcal{N}(-0.41, 0.40), & \text{when } d \text{ represented a secondary dimension.} \end{cases} \tag{22}$$

This prior can be beneficial for parameter estimation for two reasons. First, it differentiates the distributions for the discriminations on the primary and secondary dimensions. Recall that with the less informative prior, a fixed mean and *SD* were used, regardless of the dimension. However, as discussed earlier, secondary dimensions usually represent nuisance traits (Toland et al., 2017), leading to the items' discriminatory power on the primary dimension being stronger than that on the secondary dimensions. Thus, assigning the same prior distribution to all discrimination parameters may not provide information to distinguish the two sets of discriminations and thus still lead to the empirical identification issue. The adaptive informative prior differentiates the distributions for the two sets of discriminations by specifying different means for the lognormal distribution, leading to a starting point in which there is a higher probability that the sampled discriminations for the primary dimension are stronger than those for the secondary discriminations.

More importantly, the hyperprior assigned to the mean of the lognormal distribution (i.e., Equation 22) enables the prior distribution in Equation 18 to not only adapt automatically to support the range of values most appropriate to the items' discriminatory power but also adapt only within a reasonable range (Fujimoto & Neugebauer, 2020). For instance, the hyperprior assigned to the mean of the lognormal distribution for the primary dimension (or μ_{α_1}) mainly supports values that range from -1.24 to 1.24, within which approximately 95% of the values of the hyperprior fall. In other words, values smaller than -1.24 or larger than 1.24 will be unlikely to be sampled for μ_{α_1}. In turn, the hyperprior leads to a low probability of sampling values smaller than 0.21 or larger than 4.72 for the item discriminations on the primary dimension. More importantly, the hyperprior assigned to the mean of the lognormal distribution for the secondary dimensions (or μ_{α_d}, where $1 < d \leq D$) leads to a low probability of sampling values smaller than 0.14 or larger than 3.13 for the item discriminations on the secondary dimensions.

Technical Details for the Bayesian Approaches. For the two versions of the bifactor model based on Bayesian estimation, the algorithm used to estimate the posterior distribution for each model consisted of two chains. Each chain consisted of 3,500 sampled values, with the first 1,000 values discarded (i.e., the burn-in samples). In the end, there were 5,000 sampled values on which inferences were formed.

The convergence of the results was evaluated using the $\hat{R}$ statistic and visual inspection. Specifically, an $\hat{R}$ value less than 1.1 (Brooks & Gelman, 1998) and trace plots that appeared to be a "fat hairy caterpillar" without any bends (Lunn et al., 2013; Sorensen & Vasishth, 2015) were considered as adequate convergence.

Next, I introduce the analytic strategies that were used to examine the performance of the three versions of the bifactor model (i.e., Bifactor-FIML, Bifactor-LessInfo, and Bifactor-AdptInfo), which include the acceptable rate and recovery of the discrimination parameters.

Analytic Strategy

Acceptable Rate. The acceptable rate (AR) for each simulation condition was calculated as follows:

$$\text{AR} = \frac{h}{50}, \tag{23}$$

where 50 is the number of data replicates that were generated in each simulation condition. The h is the number of runs in which the model converged and none of the absolute values of the bias in the item discrimination estimates were greater than 1.00. This threshold of 1.00 was selected because it represented one *SD* on the metric of the latent trait scale (when the *SD*s of the latent trait dimensional positions are set to 1), and thus this threshold represented estimates that were inaccurate by over 1 unit of the latent trait scale. The AR, thus, represents the likelihood of obtaining accurate item discrimination estimates, with a noticeably smaller AR indicating that severely inaccurate item discrimination estimates are frequently obtained or a lack of convergence. Comparing the ARs across the simulation conditions under the same estimation method then demonstrates whether accurate estimates are more difficult to be obtained in some conditions than others, and comparing the ARs across the three versions of the bifactor model determines how biased the item discrimination estimates are likely to be under certain estimation method(s).

Parameter Recovery. To investigate how the estimation process varied across the different simulation conditions and the different versions of the bifactor model, I also examined the recovery of the item discriminations with respect to the errors of the discrimination estimates.

Each of these errors represents the difference between the value for item j's discrimination on dimension d that was estimated during the analysis of the rth data replicate ($\hat{\alpha}_{jdr}$) and the data generation value corresponding to that discrimination (α_{jd}), or formally

$$e_r = \hat{\alpha}_{jdr} - \alpha_{jd}, \tag{24}$$

with $\hat{\alpha}_{jdr}$ being the point estimate when FIML estimation is used and being the mean of the posterior distribution when a Bayesian method is used. I used boxplots to provide a visualization of the errors across the data replicates for each simulation condition. The lower and upper limits of the boxplots (i.e., the whiskers) are at most $Q_1 - 1.5\times$ IQR and $Q_3 + 1.5\times$ IQR, respectively, where Q_1 and Q_3 are the first and third quartiles, respectively, and IQR is the interquartile range (i.e., IQR $= Q_3 - Q_1$).

Item discriminations suffering from estimation issues can be reflected in the boxplots in one of two ways. One way is for the median of the errors (i.e., the median bias) being noticeably greater or less than 0. The other way is for the problematic parameters' corresponding IQRs and full ranges (upper limit minus lower limit) being larger and having more outliers than those for the parameters estimated without any issues, regardless of whether the medians are 0; even if the medians are 0, large IQRs and full ranges of the errors across the data replicates can indicate that severe over- and under-estimation occurred, representing inconsistency in the estimation of those parameters across the data replicates. Thus, the full range and IQR can demonstrate the estimation stability across the data replicates.

Study 2: Investigating the Interaction Effect Between Similarity and Magnitude of Item Discriminations

Study 2 used the same parameters as those in Study 1 but Item 1's discriminations were further manipulated to examine whether the similarity of the within-item discriminations that leads to estimation issues depends on the magnitude of the item discriminations (i.e., whether an interaction effect exists between similarity of the within-item discriminations and magnitude of the item discriminations on the empirical identification issue). The magnitude of the item discriminations is of interest because stronger item discriminations are seen in real data. For example, in Rosenberg Self-Esteem Scale (RSES) data, item discriminations could be as large as 3.65, and more importantly, the items in the RSES that were demonstrated to produce questionable results tended to have stronger discriminatory power (Chen & Fujimoto, 2022). However, it is unknown whether the magnitude of the item discriminations plays a role in the estimation issue arising from when the within-item discriminations are similar in size.

The design of this simulation study involved three factors: sample size (three sizes), similarity of the within-item discriminations (three levels), and magnitude of the within-item discriminations (two levels). The conditions for the sample size were the same as those in Study 1 that represented moderate to large samples in typical applications (i.e., 500, 1,000, and 2,000). The interaction between similarity of the within-item discriminations and magnitude of the item discriminations were represented by manipulating Item 1's discriminations. This manipulation involved two steps. First, $\alpha_{1,2}$ (i.e., Item 1's discrimination on the specific dimension) was adjusted so that $\alpha_{1,1}$ relative to $\alpha_{1,2}$ met a specific ratio. More formally, $\alpha_{1,2}$ was obtained through $\alpha_{1,2} = \alpha_{1,1}/c$, where c was 1.5, 1.3, or 1.1 depending on the ratio condition. Second, for each ratio size condition, Item 1's discriminations (i.e., both $\alpha_{1,1}$ and $\alpha_{1,2}$) were amplified by multiplying them by 1.5 and 2 to represent different levels of magnitude. For instance, under the ratio condition of 1.5 and magnitude level of 2, $\alpha_{1,1}$ (i.e., Item 1's discrimination on the primary dimension) was set to $\alpha_{1,1} \times 2$ and $\alpha_{1,2}$ was set to $\alpha_{1,1}/1.5 \times 2$. By adjusting Item 1's discriminatory strength, I was able to examine how the magnitude of the item discriminations impacts how similar the within-item discriminations have to be before the empirical identification issue. For example, when the magnitude of the item discriminations gets larger, a ratio of 1.5 may lead to the empirical identification issue (i.e., the within-item discriminations do not need to be very similar before the issue arises), whereas when the magnitude is smaller, a ratio of 1.3 may be needed (i.e., the within-item discriminations may need to be more similar before the issue arises). The data generation values for the item discriminations and category intercepts are in Tables 2 to 4.

Table 2. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.5 in Study 2

				Magnitude = 1			Magnitude = 1.5			Magnitude = 2		
	Category intercepts			Primary dimension	Secondary dimension		Primary dimension	Secondary dimension		Primary dimension	Secondary dimension	
Item	τ_{j1}	τ_{j2}	τ_{j3}	α_{j1}	α_{j2}	α_{j3}	α_{j1}	α_{j2}	α_{j3}	α_{j1}	α_{j2}	α_{j3}
1	−2.34	−0.53	1.63	2.29	1.53		3.44	2.29		4.58	3.05	
2	−0.91	0.91	2.53	2.22	1.38		2.22	1.38		2.22	1.38	
3	−1.05	1.09	2.61	1.67	0.96		1.67	0.96		1.67	0.96	
4	−2.48	−0.47	1.75	1.91			1.91			1.91		
5	−1.65	−0.05	2.25	2.15			2.15			2.15		
6	−1.89	−0.003	1.88	1.64		1.04	1.64		1.04	1.64		1.04
7	−2.46	−0.26	1.76	1.96		1.19	1.96		1.19	1.96		1.19
8	−0.96	1.33	3.62	2.34		1.49	2.34		1.49	2.34		1.49
9	−1.72	0.35	2.53	1.87		1.17	1.87		1.17	1.87		1.17
10	−2.48	−0.87	1.56	2.26		1.34	2.26		1.34	2.26		1.34

Note. An empty space indicates a value of 0.

Similar to Study 1, for each of the 3 × 3 × 2 design conditions (sample size by ratio of within-item discriminations by magnitude of item discriminations), fifty data sets were generated. The results of Study 2 were compared with those of Study 1, as the latter is equivalent to a magnitude condition in which the multiplying constant is 1. All three versions of the bifactor model (i.e., Bifactor-FIML, Bifactor-LessInfo, and Bifactor-AdptInfo) were used to analyze data under each condition. The same analytic strategies as those used in the previous study were used in this study.

Study 3: Investigating the Interaction Effect Between Similarity and Item Targetedness

Study 3 was performed to determine how item targetedness (i.e., how well each response category is represented in the data) affects how similar the within-item discriminations has to be before the empirical identification issue appears. Item targetedness was investigated because it has been demonstrated to play a factor in the estimation of IRT models (e.g., Linacre et al., 2002; Reiser & VandenBerg, 1994; Xia & Yang, 2018). Unfortunately, no evidence currently exists to show whether items being

Table 3. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.3 in Study 2

				Magnitude = 1			Magnitude = 1.5			Magnitude = 2		
	Category intercepts			Primary dimension	Secondary dimension		Primary dimension	Secondary dimension		Primary dimension	Secondary dimension	
Item	τ_{j1}	τ_{j2}	τ_{j3}	α_{j1}	α_{j2}	α_{j3}	α_{j1}	α_{j2}	α_{j3}	α_{j1}	α_{j2}	α_{j3}
1	−2.34	−0.53	1.63	2.29	1.76		3.44	2.64		4.58	3.52	
2	−0.91	0.91	2.53	2.22	1.38		2.22	1.38		2.22	1.38	
3	−1.05	1.09	2.61	1.67	0.96		1.67	0.96		1.67	0.96	
4	−2.48	−0.47	1.75	1.91			1.91			1.91		
5	−1.65	−0.05	2.25	2.15			2.15			2.15		
6	−1.89	−0.003	1.88	1.64		1.04	1.64		1.04	1.64		1.04
7	−2.46	−0.26	1.76	1.96		1.19	1.96		1.19	1.96		1.19
8	−0.96	1.33	3.62	2.34		1.49	2.34		1.49	2.34		1.49
9	−1.72	0.35	2.53	1.87		1.17	1.87		1.17	1.87		1.17
10	−2.48	−0.87	1.56	2.26		1.34	2.26		1.34	2.26		1.34

Note. An empty space indicates a value of 0.

off-targeted amplifies the estimation difficulty that occurs when within-item discriminations are similar in size.

To demonstrate the role of item targetedness in the empirical identification issue, this simulation study included two item targetedness conditions in addition to the three sample size conditions (i.e., 500, 1,000, and 2,000) and three ratio conditions (i.e., 1.1, 1.3, and 1.5). Similar to the previous studies, only Item 1's parameters were adjusted to display a clear pattern of the potential interaction effect between item targetedness and similarity of the within-item discriminations by sample size. The adjustment of the item discriminations was identical to Studies 1 and 2, that is, $\alpha_{1,2} = \alpha_{1,1}/c$, where c is 1.5, 1.3, or 1.1 depending on the ratio condition. The item targetedness was adjusted in one of two ways. One way was setting Item 1's first ($\tau_{1,1}$), second ($\tau_{1,2}$), and third ($\tau_{1,3}$) intercepts to -6.38, -2.88, and 1.92, respectively, with these values coming from an analysis of real RSES data. These values represent a situation in which Item 1's higher rating categories are noticeably over-represented in the data relative to the lower categories. Another way the item targetedness was adjusted entailed reversing Item 1's intercepts such that $\tau_{1,1}$ became -1.92, $\tau_{1,2}$ became 2.88, and $\tau_{1,3}$ became 6.38. These values represent an opposite situation in which Item 1's lower rating categories are noticeably over-represented in the data relative to the higher categories. The data generation values for the item discriminations

Table 4. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.1 in Study 2

				Magnitude = 1			Magnitude = 1.5			Magnitude = 2		
	Category intercepts			Primary dimension	Secondary dimension		Primary dimension	Secondary dimension		Primary dimension	Secondary dimension	
Item	τ_{j1}	τ_{j2}	τ_{j3}	α_{j1}	α_{j2}	α_{j3}	α_{j1}	α_{j2}	α_{j3}	α_{j1}	α_{j2}	α_{j3}
1	−2.34	−0.53	1.63	2.29	2.08		3.44	3.12		4.58	4.16	
2	−0.91	0.91	2.53	2.22	1.38		2.22	1.38		2.22	1.38	
3	−1.05	1.09	2.61	1.67	0.96		1.67	0.96		1.67	0.96	
4	−2.48	−0.47	1.75	1.91			1.91			1.91		
5	−1.65	−0.05	2.25	2.15			2.15			2.15		
6	−1.89	−0.003	1.88	1.64		1.04	1.64		1.04	1.64		1.04
7	−2.46	−0.26	1.76	1.96		1.19	1.96		1.19	1.96		1.19
8	−0.96	1.33	3.62	2.34		1.49	2.34		1.49	2.34		1.49
9	−1.72	0.35	2.53	1.87		1.17	1.87		1.17	1.87		1.17
10	−2.48	−0.87	1.56	2.26		1.34	2.26		1.34	2.26		1.34

Note. An empty space indicates a value of 0.

and category intercepts are in Tables 5 to 7.

The same number of data replicates were generated for each of the 3 × 3 × 2 conditions (sample size by ratio of within-item discriminations by item targetedness). Also, all three versions of the bifactor model and the same analytic strategies were used to investigate whether the interaction effect between similarity of the within-item discriminations and item targetedness is similar under the three estimation methods.

In the next chapter, I present the findings of the simulation studies that provide insight into how sample size, magnitude of the item discriminations, and item targetedness interact with similarity of the within-item discriminations. Then in Chapter 5, I discuss the findings, significance, and limitations of this dissertation.

Table 5. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.5 in Study 3

				Ideal targetedness			Higher categories over-represented			Lower categories over-represented		
	Primary dimension	Secondary dimension		Category intercepts			Category intercepts			Category intercepts		
Item	α_{j1}	α_{j2}	α_{j3}	τ_{j1}	τ_{j2}	τ_{j3}	τ_{j1}	τ_{j2}	τ_{j3}	τ_{j1}	τ_{j2}	τ_{j3}
1	2.29	1.53		−2.34	−0.53	1.63	−6.38	−2.88	1.92	−1.92	2.88	6.38
2	2.22	1.38		−0.91	0.91	2.53	−0.91	0.91	2.53	−0.91	0.91	2.53
3	1.67	0.96		−1.05	1.09	2.61	−1.05	1.09	2.61	−1.05	1.09	2.61
4	1.91			−2.48	−0.47	1.75	−2.48	−0.47	1.75	−2.48	−0.47	1.75
5	2.15			−1.65	−0.05	2.25	−1.65	−0.05	2.25	−1.65	−0.05	2.25
6	1.64		1.04	−1.89	−0.003	1.88	−1.89	−0.003	1.88	−1.89	−0.003	1.88
7	1.96		1.19	−2.46	−0.26	1.76	−2.46	−0.26	1.76	−2.46	−0.26	1.76
8	2.34		1.49	−0.96	1.33	3.62	−0.96	1.33	3.62	−0.96	1.33	3.62
9	1.87		1.17	−1.71	0.35	2.53	−1.71	0.35	2.53	−1.71	0.35	2.53
10	2.26		1.34	−2.48	−0.87	1.56	−2.48	−0.87	1.56	−2.48	−0.87	1.56

Note. An empty space indicates a value of 0.

Table 6. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.3 in Study 3

				Ideal targetedness			Higher categories over-represented			Lower categories over-represented		
	Primary dimension	Secondary dimension		Category intercepts			Category intercepts			Category intercepts		
Item	α_{j1}	α_{j2}	α_{j3}	τ_{j1}	τ_{j2}	τ_{j3}	τ_{j1}	τ_{j2}	τ_{j3}	τ_{j1}	τ_{j2}	τ_{j3}
1	2.29	1.76		−2.34	−0.53	1.63	−6.38	−2.88	1.92	−1.92	2.88	6.38
2	2.22	1.38		−0.91	0.91	2.53	−0.91	0.91	2.53	−0.91	0.91	2.53
3	1.67	0.96		−1.05	1.09	2.61	−1.05	1.09	2.61	−1.05	1.09	2.61
4	1.91			−2.48	−0.47	1.75	−2.48	−0.47	1.75	−2.48	−0.47	1.75
5	2.15			−1.65	−0.05	2.25	−1.65	−0.05	2.25	−1.65	−0.05	2.25
6	1.64		1.04	−1.89	−0.003	1.88	−1.89	−0.003	1.88	−1.89	−0.003	1.88
7	1.96		1.19	−2.46	−0.26	1.76	−2.46	−0.26	1.76	−2.46	−0.26	1.76
8	2.34		1.49	−0.96	1.33	3.62	−0.96	1.33	3.62	−0.96	1.33	3.62
9	1.87		1.17	−1.71	0.35	2.53	−1.71	0.35	2.53	−1.71	0.35	2.53
10	2.26		1.34	−2.48	−0.87	1.56	−2.48	−0.87	1.56	−2.48	−0.87	1.56

Note. An empty space indicates a value of 0.

Table 7. The Values Used for the Item Discriminations to Generate the Data under Ratio = 1.1 in Study 3

				Ideal targetedness			Higher categories over-represented			Lower categories over-represented		
	Primary dimension	Secondary dimension		Category intercepts			Category intercepts			Category intercepts		
Item	α_{j1}	α_{j2}	α_{j3}	τ_{j1}	τ_{j2}	τ_{j3}	τ_{j1}	τ_{j2}	τ_{j3}	τ_{j1}	τ_{j2}	τ_{j3}
1	2.29	2.08		−2.34	−0.53	1.63	−6.38	−2.88	1.92	−1.92	2.88	6.38
2	2.22	1.38		−0.91	0.91	2.53	−0.91	0.91	2.53	−0.91	0.91	2.53
3	1.67	0.96		−1.05	1.09	2.61	−1.05	1.09	2.61	−1.05	1.09	2.61
4	1.91			−2.48	−0.47	1.75	−2.48	−0.47	1.75	−2.48	−0.47	1.75
5	2.15			−1.65	−0.05	2.25	−1.65	−0.05	2.25	−1.65	−0.05	2.25
6	1.64		1.04	−1.89	−0.003	1.88	−1.89	−0.003	1.88	−1.89	−0.003	1.88
7	1.96		1.19	−2.46	−0.26	1.76	−2.46	−0.26	1.76	−2.46	−0.26	1.76
8	2.34		1.49	−0.96	1.33	3.62	−0.96	1.33	3.62	−0.96	1.33	3.62
9	1.87		1.17	−1.71	0.35	2.53	−1.71	0.35	2.53	−1.71	0.35	2.53
10	2.26		1.34	−2.48	−0.87	1.56	−2.48	−0.87	1.56	−2.48	−0.87	1.56

Note. An empty space indicates a value of 0.

CHAPTER 4

RESULTS

In the previous chapter, I discussed the details of the simulation studies that were conducted to investigate the empirical identification issue with the bifactor model, including whether the estimation method matters. In this chapter, I present the results of the simulation studies conducted. For each study, I discuss the results of acceptable rate and parameter recovery regarding the errors of the parameter estimates by estimation method.

Results of Study 1

To provide a review, this first study was conducted to examine the similarity of the within-item discriminations required before the empirical identification issue arises. It explored the conditions in which the item discrimination similarity (or the ratio of the within-item discriminations) varied, while item targetedness and the magnitude of the item discriminations were ideal. Specifically, the factors manipulated were the sample size (500, 1,000, and 2,000) and the similarity of the within-item discriminations, represented by ratio (1.1, 1.3, and 1.5). Next, I discuss the results by estimation method, followed by a summary of the main findings in Study 1.

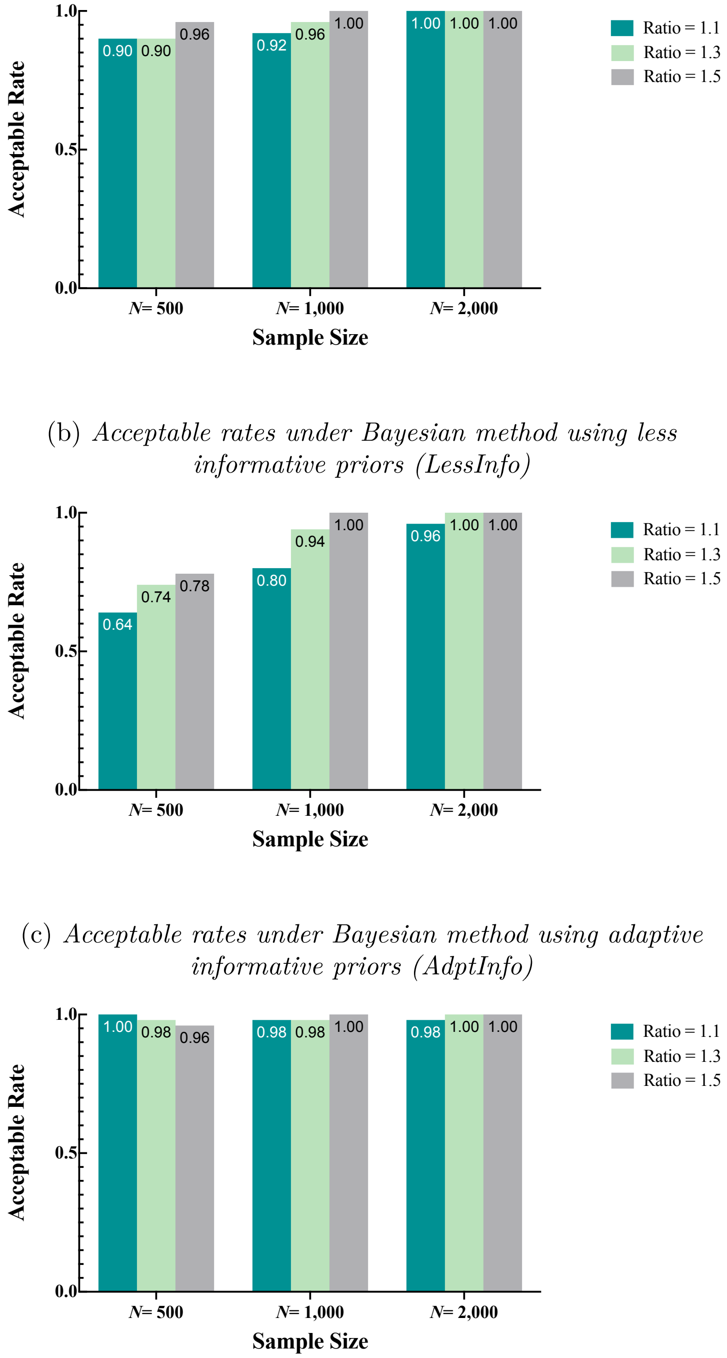
Acceptable Rates across the Three Estimation Methods

Recall that the acceptable rate (AR) represents the proportion of runs that converged and none of the absolute values of the errors in the item discrimination estimates were greater than 1.00. The ARs for the three estimation methods are summarized in Figure 3. In each plot, the sample sizes are represented along the x-axis, the acceptable rates are represented along the y-axis, and the ratios are represented by different colors with smaller ratio indicating that the within-item discriminations are more similar in magnitude and larger ratio indicating the within-item discriminations are more distinct.

Under full-information maximum likelihood (FIML; in Figure 3a), within a sample size, the ARs did not change much across the ratios. For instance, when $N = 500$, the AR was .90 for the ratio of 1.1, .90 for the ratio of 1.3, and .96 for the ratio of 1.5. Similarly, in each of the other two sample sizes, the ARs displayed minor variability across different ratios within the sample size. Additionally, when comparing across the sample sizes under FIML, the ARs were similar for all sample sizes regardless of the ratio. In other words, sample size had minimal impact on obtaining impermissible or extreme estimates regardless of how similar the within-item discriminations were. Specifically, for $N = 500$, the average AR across the ratios was .92, and it was .96 and 1.00 for $N = 1{,}000$ and 2,000, respectively.

In contrast, under Bayesian method using less informative priors (LessInfo; in Figure 3b), the ARs changed more noticeably across the ratio conditions for $N = 500$. The AR was .64 for the ratio of 1.1, .74 for the ratio of 1.3, and .78 for the ratio of 1.5, showing that when the sample size was comparatively small, potentially acceptable parameter estimates were more likely to be obtained under LessInfo when the within-item discriminations were distinct. This trend persisted for $N = 1{,}000$, under which the AR was .80 for the ratio of 1.1, whereas is was .94 and 1.00 for the ratios of 1.1 and 1.5, respectively. However, this pattern did not hold as strongly for $N = 2{,}000$, under which the AR varied less as the ratio increased. Particularly, when $N = 2{,}000$, the ARs ranged from .96 to 1.00. In spite of the trends observed within LessInfo, the ARs were noticeably lower under LessInfo than FIML, especially in the smaller sample sizes and smaller ratio conditions.

When comparing across the sample sizes under LessInfo, the average AR across the ratios was clearly lower for the smaller sample size than those for the larger sample sizes, which was inconsistent with the pattern observed under FIML. For $N = 500$, the average AR was .72 across the ratios, whereas it was .91 and .99 for $N = 1{,}000$ and 2,000, respectively. The average ARs were noticeably lower under LessInfo, particularly for the



(a) *Acceptable rates under full-information maximum likelihood estimation (FIML)*

(b) *Acceptable rates under Bayesian method using less informative priors (LessInfo)*

(c) *Acceptable rates under Bayesian method using adaptive informative priors (AdptInfo)*

Figure 3

Figure 3. Visualizations of the acceptable rates (ARs) under the three estimation methods, from Study 1. In each plot, the sample sizes and the proportion of runs in which the item discrimination estimates were acceptable are represented along the horizontal and vertical axes, respectively. Ratios are represented in different colors.

smaller sample sizes, which suggests that the Bayesian method using less informative priors was more likely to produce unacceptable parameter estimates than FIML when the sample size was small.

With respect to Bayesian method using adaptive informative priors (AdptInfo; in Figure 3c), the ARs were almost perfect for all sample sizes and ratio conditions (AR $\geq .96$), suggesting that AdptInfo consistently produced potentially acceptable parameter estimates. Comparing with FIML, the ARs were higher under AdptInfo than those obtained under FIML for the smallest sample size. Specifically, under AdptInfo, the average AR across the ratios was .98 for $N = 500$, whereas this rate was .92 under FIML. When comparing AdptInfo with LessInfo, the ARs under AdptInfo were noticeably larger than those under LessInfo, particularly at $N = 500$ and $N = 1{,}000$. Specifically, the average ARs at $N = 500$ and $N = 1{,}000$ were .98 and .99, respectively, under AdptInfo, whereas these two rates were .72 and .91, respectively, under LessInfo.

Overall, the AR patterns suggest that non-convergence or extreme parameter estimates were barely obtained under AdptInfo and FIML, regardless of sample size and ratio of within-item discriminations. However, under LessInfo, the ARs were higher for larger sample sizes, suggesting that the sample size played a more impactful role in the parameter estimation of the bifactor model for LessInfo than for the other two methods.

Although the AR showed the frequency of permissible results, it did not demonstrate how accurate the parameter estimates were and whether all runs that converged produced consistent results. The parameter recovery results I review next provide insight into the accuracy and stability of the parameter estimates and inform whether AdptInfo and FIML also provided estimates that were more accurate and consistent than LessInfo for the bifactor model as the within-item discriminations changed.

Parameter Recovery: Errors of the Parameter Estimates

Full-information maximum likelihood. The distribution of the errors related to the item discrimination estimates under FIML is summarized in Figure 4. In each plot, the item discrimination parameters are represented along the x-axis (where α_{jd} is item j's discrimination on dimension d), and the y-axis represents the errors. The plots within the same row differ in sample size and the plots within the same column differ in ratio. If no estimation difficulty exists, then the errors related to each item should be distributed similarly within a sample size and centered around 0, and thus the boxplots (i.e., medians, interquartile ranges [IQRs; i.e. $Q_3 - Q_1$], lower and upper limits of the boxplots [LULs; i.e., $Q_1 - 1.5\times$ IQR and $Q_3 + 1.5\times$], and full ranges [i.e., upper limit minus lower limit]) should be similar across the discriminations. Unfortunately, that is not the case.

Although the median errors were all close to 0 for $N = 500$ (the plots in the left column of Figure 4), the IQRs, LULs, and full ranges varied across the parameters, showing the inconsistency in the accuracy of the estimates for the discriminations. Specifically, when Item 1's discriminations on the primary and secondary dimensions were very similar in magnitude (i.e., ratio = 1.1), the IQR, LUL, and full range of the errors across the data replicates for Item 1 were noticeably greater than those for the other items, especially for the errors related to the primary dimension. For example, the range of the errors was 1.67 with an LUL of (−0.64, 1.03) for $\alpha_{1,1}$, whereas among the other items' discriminations on the primary dimension (i.e., α_{j1} where $j \neq 1$), the largest error range was 1.16 (LUL[−0.43, 0.74]) and the smallest was 0.60 (LUL[−0.34, 0.37]), which were for $\alpha_{8,1}$ and $\alpha_{3,1}$, respectively. A comparable pattern was observed in the errors related to the secondary dimensions. In particular, when the ratio = 1.1, the range of the errors was 2.31 (LUL[−0.69, 1.62]) for $\alpha_{1,2}$, whereas among the other items' discriminations on the secondary dimensions (i.e., $\alpha_{j,d}$ where $j \neq 1$ and $d \neq 1$), the largest error range was 1.13 (LUL[−0.49, 0.64]) and the smallest was 0.81 (LUL[−0.40, 0.40]), which were for $\alpha_{2,2}$ and $\alpha_{3,2}$, respectively.

When Item 1's discriminations became more distinct, the errors associated with $\alpha_{1,1}$ became more similar to those of the other items' discriminations on the primary dimension, whereas the errors associated with $\alpha_{1,2}$ still exhibited larger ranges and LULs than those of

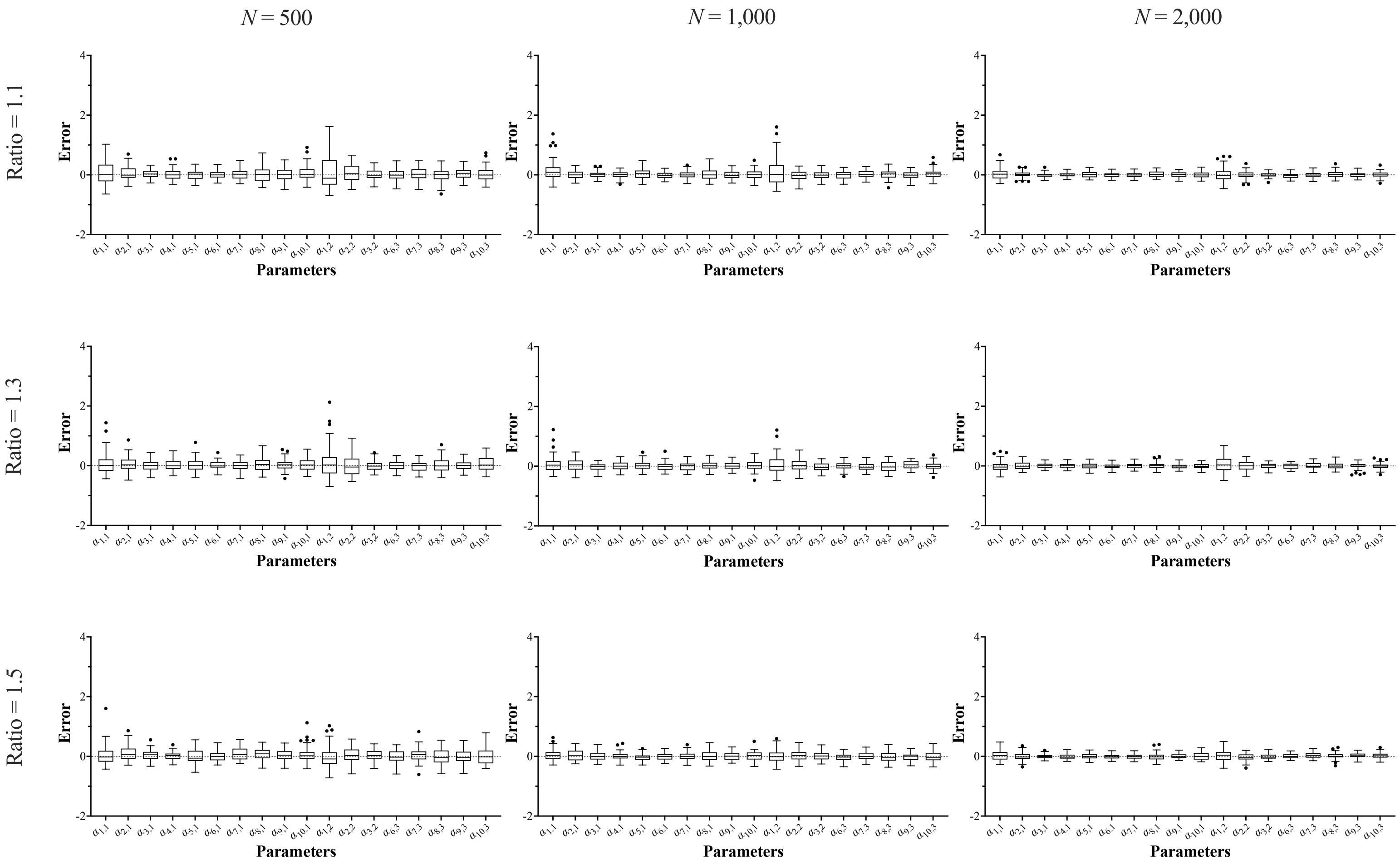


Figure 4

Figure 4. Boxplots summarizing the errors of the item discrimination estimates under full-information maximum likelihood (FIML) by sample size and ratio of Item 1's discriminations, from Study 1. The plots in each row differ in sample size; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

the other items' discriminations on the secondary dimension. Specifically, for ratio = 1.3, the range of the errors was 1.22 (LUL[−0.44, 0.79]) for $\alpha_{1,1}$, which was similar to the largest error range (i.e., 1.14 (LUL[−0.48, 0.66]) for $\alpha_{2,1}$) among the other items' discriminations on the primary dimension. In contrast, the range of the errors for $\alpha_{1,2}$ was 1.81 (LUL[−0.70, 1.11]), whereas among the other items' discriminations on the secondary dimensions, the largest was 1.45 (LUL[−0.53, 0.92]), which was for $\alpha_{2,2}$.

When the ratio of Item 1's discriminations further increased to 1.5, the range of the errors was 1.44 (LUL[−0.72, 0.72]) for $\alpha_{1,2}$, whereas among the other items' discriminations on the secondary dimension, the largest error range was 1.19 (LUL[−0.40, 0.79]), which was for $\alpha_{10,3}$.

As the sample size increased to 1,000 (the plots in the middle column of Figure 4), the ranges of the errors related to Item 1 became smaller compared with those observed at $N = 500$. However, its ranges were still larger than those of the other items, particularly when its within-item discriminations were more similar in strength. For example, the range of the errors was 1.15 (LUL[−0.40, 0.74]) for $\alpha_{1,1}$ under the ratio of 1.1, whereas the largest among the other items' discriminations on the primary dimension was 0.85 (LUL[−0.31, 0.54]), which was for $\alpha_{8,1}$. Likewise, the range of the errors associated with $\alpha_{1,2}$ was larger than that observed among the remaining discriminations on the secondary dimension. However, as the within-item discriminations became more distinct, the differences in the range of the errors between Item 1 and the other items were negligible, regardless of the dimension of the parameter. For example, when ratio = 1.3, the range of the errors was 1.28 (LUL[−0.48, 0.80]) for $\alpha_{1,2}$, whereas the largest among the other items' discriminations on the secondary dimension was 0.95 (LUL[−0.41, 0.54]), which was for $\alpha_{2,2}$.

When the sample size increased to 2,000 (the plots in the right column of Figure 4), the range of the errors related to Item 1 became even smaller. Additionally, the inconsistency in the accuracy of the estimates between Item 1's parameters and those of the other items became minimal, especially when Item 1's discriminations were clearly different. In particular, when $N = 2{,}000$ and ratio $= 1.5$, the range of the errors for $\alpha_{1,1}$ was 0.76 (LUL[-0.28, 0.48]), whereas the ranges in the errors for the other items' discriminations on the primary dimension fell between 0.33 (LUL[-0.15, 0.18] for $\alpha_{3,1}$) and 0.63 (LUL[-0.32, 0.31] for $\alpha_{2,1}$). Similarly, the range of the errors for $\alpha_{1,2}$ was 0.90 (LUL[-0.40, 0.50]), whereas the ranges for the other items' discriminations on the secondary dimensions fell between 0.32 (LUL[-0.14, 0.18] for $\alpha_{6,3}$) and 0.56 (LUL[-0.36, 0.20] for $\alpha_{2,2}$).

Bayesian method using less informative priors. The distribution of the errors related to the item discrimination estimates under LessInfo is summarized in Figure 5. For $N = 500$ (the plots in the left column of Figure 5), the errors associated with Item 1's discriminations showed noticeably larger medians, LULs, and/or full ranges than those observed in the other items, especially when Item 1's discriminations were more similar in magnitude. Specifically, when Item 1's discriminations on the primary and secondary dimensions were very similar in magnitude (i.e., ratio $= 1.1$), the median of the errors for $\alpha_{1,1}$ and $\alpha_{1,2}$ were 0.30 and 0.21, respectively, whereas the median for the other items' discrimination parameters were all close to 0. The full ranges and LULs further demonstrated the inconsistency of estimation accuracy between Item 1 and the other items when $N = 500$. Specifically, the range of the errors was 2.82 with an LUL of (-0.59, 2.22) for $\alpha_{1,1}$ when ratio $= 1.1$, whereas among the other items' discriminations on the primary dimension (i.e., α_{j1} where $j \neq 1$), the largest error range was 1.19 (LUL[-0.44, 0.75]) and the smallest was 0.60 (LUL[-0.29, 0.31]), which were for $\alpha_{8,1}$ and $\alpha_{3,1}$, respectively. For the item discriminations on the secondary dimensions, the range of the errors was 3.32 (LUL[-0.61, 2.71]) for $\alpha_{1,2}$, which was still noticeably larger than those among the other discriminations on the secondary dimensions (i.e., $\alpha_{j,d}$ where $j \neq 1$ and $d \neq 1$) with the largest error range of the latter was 1.30 (LUL[-0.51, 0.79]), which was for $\alpha_{2,2}$, and the smallest was 0.82 (LUL[-0.43, 0.39]) for $\alpha_{3,2}$.

As Item 1's discriminations became more distinct at ratio $= 1.3$, similar patterns

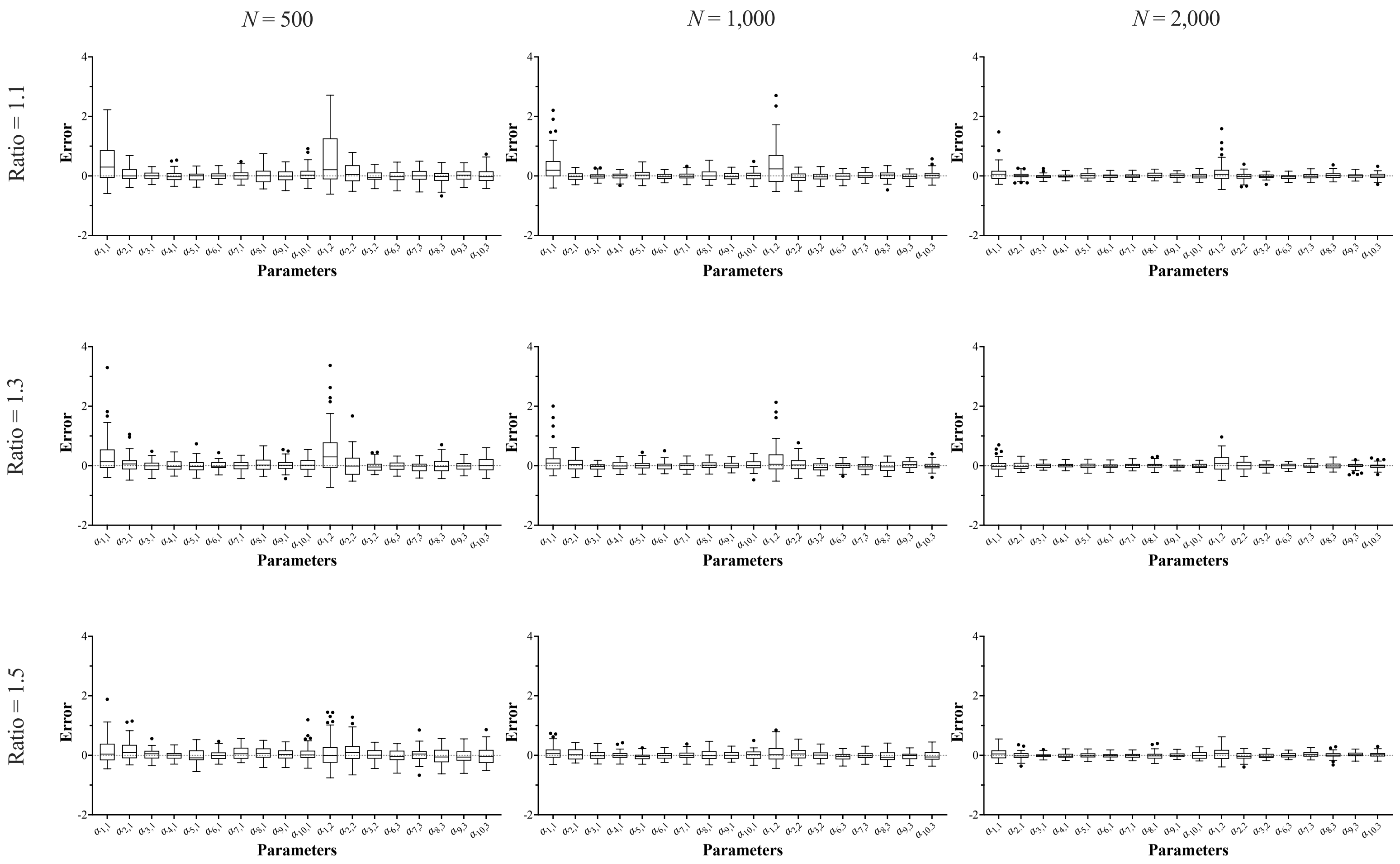


Figure 5

Figure 5. Boxplots summarizing the errors of the item discrimination estimates under Bayesian method using less informative priors (LessInfo) by sample size and ratio of Item 1's discriminations, from Study 1. The plots in each row differ in sample size; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

were observed as those noticed when ratio = 1.1. That is, the medians of the errors related to Item 1's discrimination parameters, especially $\alpha_{1,2}$, were larger than 0, whereas the medians of the errors for the other items' discrimination parameters were all close to 0. Additionally, the full ranges and LULs for Item 1's discrimination parameters were clearly larger than those for the other item discriminations. When the ratio increased to 1.5, the median of the errors related to Item 1's parameter estimates became near 0. Moreover, the disparity in the full ranges and LULs of the errors between Item 1 and the other items became less pronounced compared with what was observed when Item 1's discriminations were relatively similar. Particularly, the range of the errors was 1.69 (LUL[−0.45, 1.23]) for $\alpha_{1,1}$ and 1.83 (LUL[−0.76, 1.07]) for $\alpha_{1,2}$ under ratio = 1.5, whereas among the other item discriminations, the largest error range for the discriminations on the primary dimension was 1.35 (LUL[−0.32, 1.03]), which was for $\alpha_{2,1}$, and the discrimination on the secondary dimensions was 1.62 for $\alpha_{2,2}$ (LUL[−0.66, 0.96]).

When the sample size increased to 1,000, the estimation for all items, including Item 1, became more accurate, although similar trends in estimation inconsistency described for the $N = 500$ were still observed between Item 1 and the other items. For $N =$ 1,000 (the plots in the middle column of Figure 5), the errors for $\alpha_{1,1}$ under the ratio of 1.1 had a median of 0.20, with the range being 1.68 (LUL[−0.41, 1.27]), and the errors for $\alpha_{1,2}$ had a median of 0.24, with the range being 2.58 (LUL[−0.52, 2.06]). However, for the other item discriminations, regardless of the dimension, the largest absolute median of errors was 0.04 (for $\alpha_{2,2}$), with the range being 0.81 (LUL[−0.51, 0.30]). As the ratio increased to 1.3, the medians associated with Item 1 were near 0, but the inconsistency in estimation accuracy persisted. With the ratio further increased to 1.5, the magnitude of the differences in estimation accuracy between Item 1 and the other items became minor, especially for the discriminations on the primary dimension. Specifically, the range of the errors was 0.92 (LUL[−0.31, 0.61]) for $\alpha_{1,1}$, which was similar to those observed for other item discriminations on the primary dimension where the largest and smallest ranges were 0.83

(LUL[−0.34, 0.49]) for $\alpha_{10,1}$ and 0.50 (LUL[−0.23, 0.26]) for $\alpha_{6,1}$, respectively.

When the sample size increased to $N = 2{,}000$ (the plots in the right column of Figure 5), the medians and ranges of the errors related to Item 1 decreased further regardless of the similarity of its within-item discriminations. More importantly, the magnitude of the differences in the accuracy of the estimates between Item 1's parameters and those of the other items became even less noticeable, especially when Item 1's discriminations were clearly different. For instance, when ratio = 1.5, the error range of $\alpha_{1,1}$ was 0.83 (LUL[−0.28, 0.55]), similar to the error ranges of the other items' discrimination estimates on the primary dimension that fell between 0.33 (LUL[−0.16, 0.17]) for $\alpha_{3,1}$ and 0.62 (LUL[−0.31, 0.30]) for $\alpha_{2,1}$.

Bayesian method using adaptive informative priors. The distribution of the errors related to the item discrimination estimates under AdptInfo is summarized in Figure 6. For $N = 500$ (the plots in the left column of Figure 6), the errors associated with Item 1's discrimination parameter on the primary dimension exhibited comparable median, LUL, and full range to those seen for the other items, as observed for all ratios. For instance, when ratio = 1.1, the medians were all close to 0 for the discriminations on the primary dimension, including $\alpha_{1,1}$; with respect to LUL and range, the errors for $\alpha_{1,1}$ had a range of 1.11, with an LUL of (−0.64, 0.47), whereas the largest and smallest ranges for the other parameters on the primary dimension were 1.13 (LUL[−0.42, 0.71] for $\alpha_{8,1}$) and 0.62 (LUL[−0.26, 0.36] for $\alpha_{3,1}$).

Concerning the discrimination parameters on the secondary dimensions, the errors related to $\alpha_{1,2}$ displayed a larger range and LUL than those observed for other items, especially for the ratios of 1.1 and 1.3. For example, when ratio = 1.1, the range of the errors for $\alpha_{1,2}$ was 1.50 (LUL[−0.70, 0.80]), whereas for the other items' discriminations on the secondary dimensions, the largest and smallest ranges were 1.03 (LUL[−0.43, 0.60]) and 0.71 (LUL[−0.31, 0.40]), which were for $\alpha_{2,2}$ and $\alpha_{3,2}$ respectively. As the ratio increased to 1.5, the discrepancies in ranges and LULs between $\alpha_{1,2}$ and the other

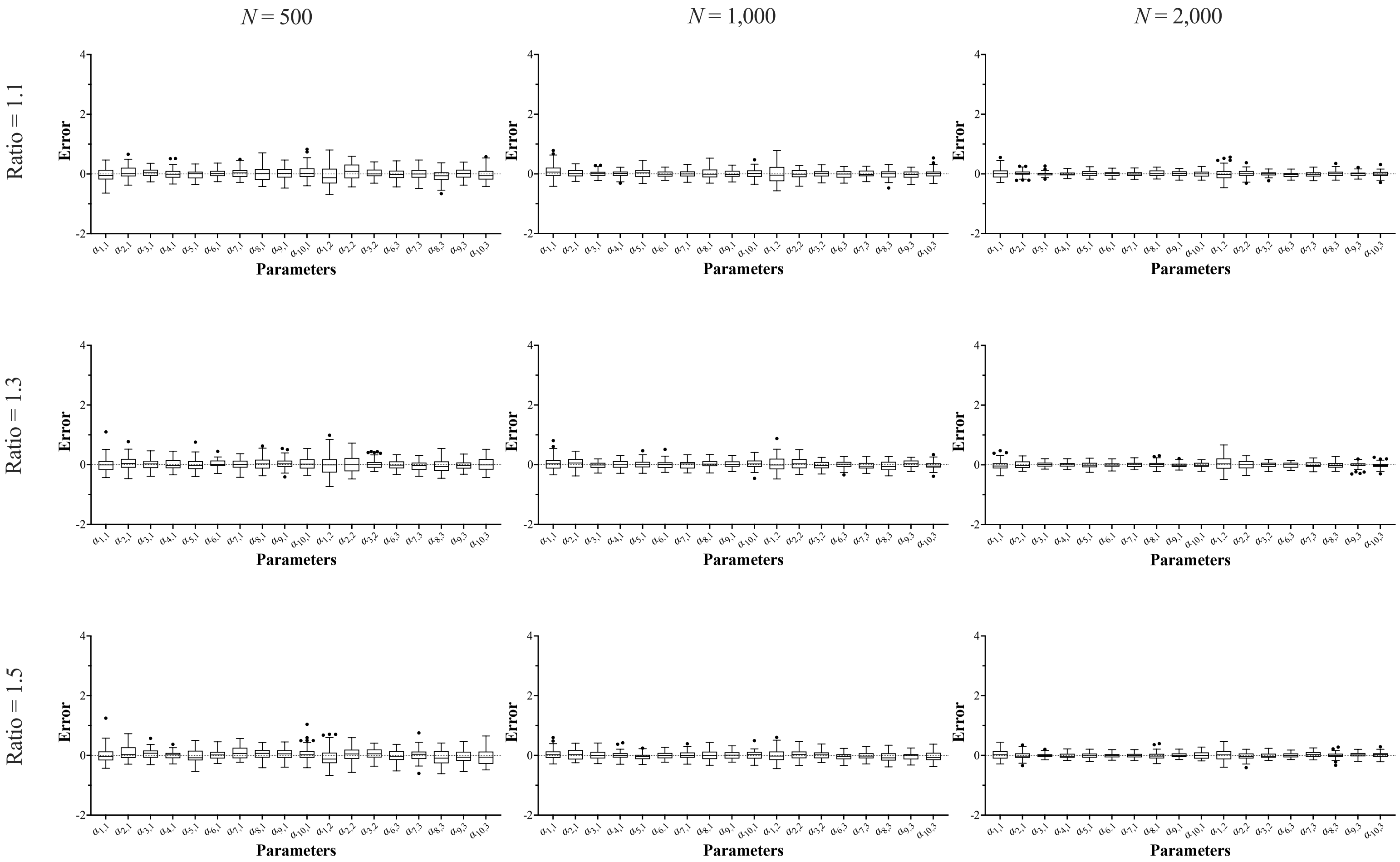


Figure 6

Figure 6. Boxplots summarizing the errors of the item discrimination estimates under Bayesian method using adaptive informative priors (AdptInfo) by sample size and ratio of Item 1's discriminations, from Study 1. The plots in each row differ in sample size; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

discriminations on the secondary dimensions became smaller. Specifically, the range of the errors for $\alpha_{1,2}$ was 1.30 (LUL[−0.67, 0.62]) and the range for the other discriminations on the secondary dimensions fell between 0.77 (LUL[−0.36, 0.41]) for $\alpha_{3,2}$ and 1.16 (LUL[−0.57, 0.59]) for $\alpha_{2,2}$.

For $N = 1{,}000$ (the plots in the middle column of Figure 6), a similar trend was observed across the discrimination parameters on the primary dimension, with no significant differences observed between Item 1 and the other items. With respect to the secondary dimensions, the discrepency in estimation accuracy between Item 1 and the other items still existed, particularly when Item 1's discriminations were similar, albeit to a lesser extent. Specifically, when ratio = 1.1, the range of the errors for $\alpha_{1,2}$ was 1.36 (LUL[−0.57, 0.79]), whereas the largest and smallest ranges were 0.74 (LUL[−0.43, 0.32]) for $\alpha_{8,3}$ and 0.52 (LUL[−0.26, 0.26]) for $\alpha_{7,3}$ among the other item's discriminations on the secondary dimensions.

When the sample size increased to 2,000, all discrimination estimates, including those for Item 1, were noticeably better. With respect to estimation accuracy between Item 1's discriminations and the other discriminations, similar patterns were seen as those for the other two sample size conditions. Specifically, no noticeable differences were observed for $\alpha_{1,1}$, compared with the other discriminations on the primary dimension. However, the estimates for $\alpha_{1,2}$ displayed a larger error range and wider LUL than the other discriminations on the secondary dimensions.

Comparison across the three estimation methods. The empirical identification issue was consistently observed across all three estimation methods. Specifically, when estimating the discrimination parameters of Item 1, the accuracy was noticeably poorer compared with the estimation of the discrimination parameters for the other items. This discrepancy was more pronounced when the sample size was relatively small (for example, with $N = 500$ or 1,000) and when Item 1's discriminations were similar (such as when the ratio was 1.1 or 1.3). However, the degree of the discrepancy in the estimation accuracy varied depending on the estimation method used. For example, the disparity in estimation accuracy between Item 1 and the other items was less pronounced under FIML than LessInfo, showing that FIML could provide more accurate estimates for the bifactor model in the conditions where the sample size is small, the within-item discriminations are similar, or both.

However, when more informative priors were used (i.e., AdptInfo), there was a noticeable improvement in estimating Item 1's discrimination parameters compared with FIML and LessInfo. More specifically, under AdptInfo, the estimation of Item 1's parameter on the primary dimension achieved accuracy levels comparable to those of the other items, and this held true irrespective of the sample size and the similarity of Item 1's discriminations. With regards to Item 1's discrimination on the secondary dimension, its estimation displayed greater inaccuracy and instability compared with the other items, albeit to a noticeably lesser extent than under FIML and LessInfo.

Summary of Study 1

Study 1 demonstrated that within-item discriminations being similar may lead to estimation difficulties. As the within-item discriminations become more similar, the challenges in estimation become increasingly pronounced. Moreover, sample size plays a role in amplifying the estimation challenges. Specifically, for a given level of similarity of the within-item discriminations (i.e., under the same ratio condition), the estimation difficulties intensifies as the sample size decreases.

Additionally, Study 1 evaluated the efficacy of different estimation methods in estimating the discrimination parameters of the bifactor model under challenging situations. The findings show that AdptInfo performs better than FIML and LessInfo, as AdptInfo consistently produced more accurate and stable estimates, particularly when the within-item discriminations were most similar.

In sum, the findings of Study 1 suggest that the empirical identification issue with the bifactor model is affected by both similarity of the within-item discriminations and sample size, and highlight the importance of carefully considering these factors when modeling a bifactor structure. Furthermore, it demonstrates that the choice of estimation method matters, as estimation method can greatly affect the quality of the results.

Results of Study 2

To provide a review, Study 2 further investigated whether the effect of within-item discrimination similarity on the estimates depended on magnitude of the item discriminations. The study explored the conditions in which the item targetedness was ideal but the similarity of the within-item discriminations and the magnitude of the discriminations were varied. Specifically, the factors manipulated in this study were sample size (500, 1,000, and 2,000), ratio of within-item discriminations (1.1, 1.3, and 1.5), and magnitude of the discriminations (1.5 and 2). Next, I discuss the results by estimation method, followed by a comparison of the performances of the three estimation methods and a summary of the main findings. The results of Study 2 were compared with those of Study 1, as the latter is equivalent to a magnitude condition in which the multiplying constant is 1.

Full-Information Maximum Likelihood Estimation

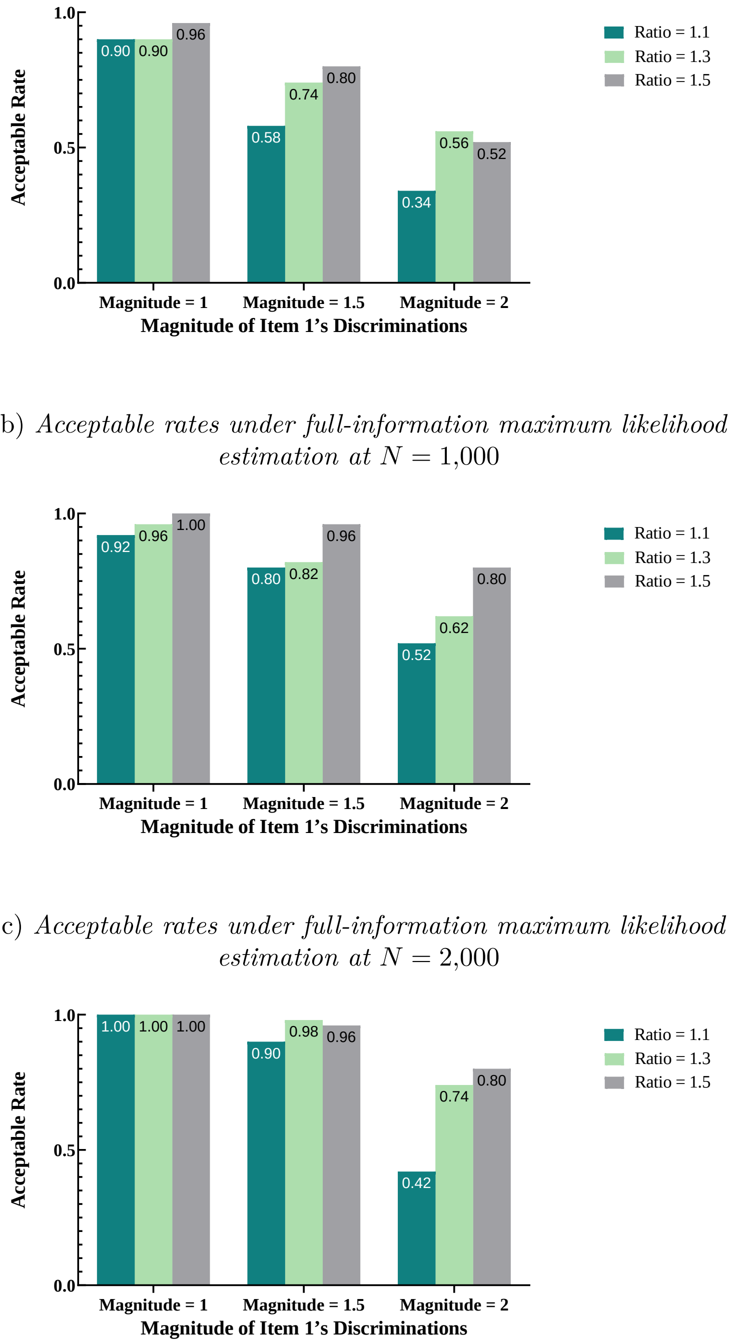
Acceptable Rates. Recall that the acceptable rates (ARs) represent the proportion of runs that converged and none of absolute values of the error in the item discrimination estimates were greater than 1.00. The ARs for full-information maximum likelihood (FIML) estimation are summarized in Figure 7. In each plot, the magnitudes of the within-item discriminations are represented along the x-axis, the acceptable rates are represented along the y-axis, and the ratios are represented by different colors. The results with a magnitude of 1 indicate that they originated from Study 1 and thus serve as a baseline.

When $N = 500$ (as depicted in Figure 7a), the ARs were all acceptable and did not show significant variation compared with the baseline condition (i.e., magnitude = 1, which is in Study 1), regardless of how similar Item 1's discriminations were. However, as the magnitude of Item 1’s discriminations increased, the ARs were reduced noticeably. For instance, when magnitude = 2, the AR for ratio = 1.1 dropped to .34 from the baseline value of .90. Similar patterns were observed under the other two ratio conditions, where the AR noticeably decreased as the magnitude of Item 1’s discriminations increased. This indicates that when dealing with large magnitudes of within-item discrimination, there is a higher likelihood of obtaining extreme estimates, warning messages, or both.

When the sample size increased to 1,000 and 2,000 (as shown in Figures 7b and 7c), a general increase in ARs was observed irrespective of the ratio and the magnitude of Item 1’s discriminations. However, variation in ARs across the different magnitude conditions remained, albeit they were less pronounced than when $N = 500$. In other words, as the magnitude of Item 1’s discriminations increased, the ARs decreased, though not as sharply as when the sample size was smaller. For example, when $N = 1{,}000$, the AR for ratio = 1.1 was .52 when magnitude = 2, compared with .92 at baseline. However, the reduction in ARs (i.e., .40) was not as steep as that observed when $N = 500$ (i.e., .56).

The findings pertaining to ARs suggest that magnitude of within-item discriminations plays a role in the estimation of the bifactor model. The results of the parameter recovery reviewed next further reveal the estimation accuracy and stability of FIML under different magnitude conditions.

Parameter Recovery: Errors of the Parameter Estimates. For $N = 500$, the distribution of the errors related to the item discrimination estimates under FIML is summarized in Figure 8. In each plot, the item discrimination parameters are represented along the x-axis (where α_{jd} is item j’s discrimination on dimension d), and the y-axis represents the errors. The plots within the same row differ in the magnitude of Item 1’s discriminations, while the plots within the same column differ in the ratio of Item 1’s discriminations. If the magnitude of within-item discriminations does not play a role in the empirical identification issue, then the errors related to Item 1 should be distributed similarly across various magnitudes conditions, provided the ratio condition remains


(a) *Acceptable rates under full-information maximum likelihood estimation at $N = 500$*

(b) *Acceptable rates under full-information maximum likelihood estimation at $N = 1{,}000$*

(c) *Acceptable rates under full-information maximum likelihood estimation at $N = 2{,}000$*

Figure 7

Figure 7. Visualizations of the acceptable rates (ARs) under full-information maximum likelihood estimation (FIML), from Study 2. The magnitudes of within-item discriminations are represented along the horizontal axis; the proportion of runs in which the item discrimination estimates were acceptable are represented along the vertical axis. Ratios are represented in different colors.

constant.

Unfortunately, when the within-item discriminations for Item 1 were highly similar (i.e., ratio = 1.1), the discrepancies in the estimation accuracy between Item 1's discrimination parameters and those of the other items became more pronounced as the magnitude of Item 1's discriminations increased, as illustrated in the plots in the first row of Figure 8). Specifically, when magnitude = 1.5, the range of the errors was 3.94 with an LUL of (−1.12, 2.81) for $\alpha_{1,1}$, whereas among the other items' discriminations on the primary dimension (i.e., α_{j1} where $j \neq 1$), the largest error range was 1.08 (LUL[−0.35, 0.73]) and the smallest was 0.62 (LUL[−0.20, 0.41]), which were for $\alpha_{8,1}$ and $\alpha_{9,1}$, respectively. Recall that at baseline (magnitude = 1; from Study 1), the range was 1.67 (LUL[−0.64, 1.03] for $\alpha_{1,1}$, with the largest and smallest ranges across the other parameters on the primary dimension being 1.16 (LUL[−0.43, 0.74]) and 0.60 (LUL[−0.27, 0.32]), respectively. Notice that the largest and smallest ranges of the errors across the other parameters on the primary dimension remained relatively stable; however, the range of the errors for $\alpha_{1,1}$ increased from 1.67 to 3.94 as the magnitude of the within-item discriminations for Item 1 increased from 1 to 1.5. With regards to the discriminations on the secondary dimension, similar patterns were observed. Particularly, the range of the errors for $\alpha_{1,2}$ was 4.55 (LUL[−1.32, 3.24], while the largest and smallest ranges across the other discrimination parameters on the secondary dimension were 1.34 (LUL[−0.52, 0.82]) (for $\alpha_{8,3}$) and 0.68 (LUL[−0.29, 0.39]) (for $\alpha_{3,2}$), respectively. At baseline, that range was 2.31 (LUL[−0.69, 1.62]) for $\alpha_{1,2}$, with the largest and smallest ranges across the other discrimination parameters on the secondary dimension being 1.13 (LUL[−0.49, 0.64]) and 0.81 (LUL[−0.40, 0.40]), respectively. As the magnitude of Item 1's discriminations further increased to 2, the errors associated with Item 1 exhibited even larger ranges compared with those under magnitude = 1 and 1.5. Specifically, when magnitude = 2, the range of the errors increased to 8.62 (LUL [−2.25, 6.37) for $\alpha_{1,1}$ and 10.61 (LUL [−2.49, 8.12) for $\alpha_{2,1}$. It is worth noting that these ranges remained considerably larger than those observed

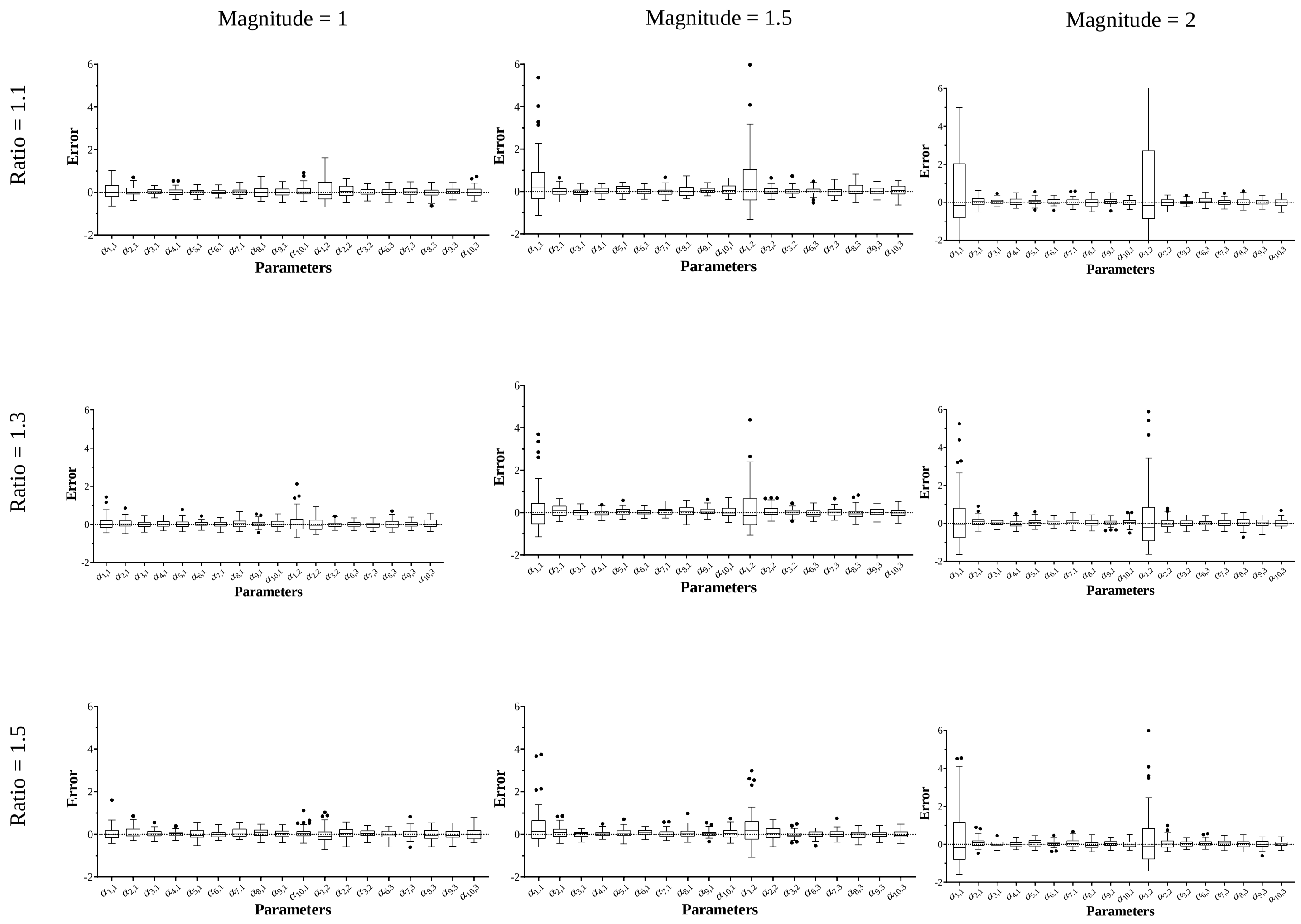


Figure 8

Figure 8. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under full-information maximum likelihood (FIML), from Study 2. The plots in each row differ in magnitude of Item 1's discriminations; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

for other discrimination parameters.

When the within-item discriminations for Item 1 became more differentiated (i.e., ratio = 1.3 and 1.5; shown in the plots in the second and third rows of Figure 8)), the trends observed were similar to those for ratio = 1.1. Specifically, as the magnitude of Item 1's discriminations increased, the full ranges and LULs of the errors related to Item 1's parameter estimates became noticeably larger compared with those observed under magnitude = 1, while the ranges and LULs of the errors related to the parameter estimates for the other items remained stable. This indicates that the magnitude of an item's discriminations can impact the estimation difficulties arising from when its within-item discriminations are similar in size. Larger item discriminations exacerbate the difficulty in estimating that item's discrimiations. In other words, whether an item's discriminations being similar leads to estimation issues depends on the magnitude of the item discriminations.

As the sample size increased to 1,000 (as shown in Figure 9) and 2,000 (as shown in Figure 10), similar patterns persisted under each ratio condition. However, the ranges and LULs associated with Item 1 were mitigated compared with those at $N = 500$. For instance, at $N = 2{,}000$, the error range was 6.47 (LUL[-1.42, 5.05)]) for $\alpha_{1,1}$ and 7.07 (LUL[-1.62, 5.45)]) for $\alpha_{1,2}$ in the most challenging scenario (i.e., ratio = 1.1 and magnitude = 2). In contrast, at $N = 500$, that error range was 8.62 (LUL[-2.25, 6.37)]) for $\alpha_{1,1}$ and 10.61 (LUL[-2.49, 8.12)]) for $\alpha_{1,2}$. This trend held across other combinations of ratio and magnitude conditions, suggesting that a larger sample size can alleviate estimation difficulties that arise when an item's discriminations are large and similar in strength.

Bayesian Method Using Less Informative Priors

Acceptable Rates. The acceptable rates (ARs) for the Bayesian method using less informative priors (LessInfo) are summarized in Figure 11. At $N = 500$ (as shown in Figure 11a), a noticeable decline in ARs was observed as the magnitude of Item 1's discriminations

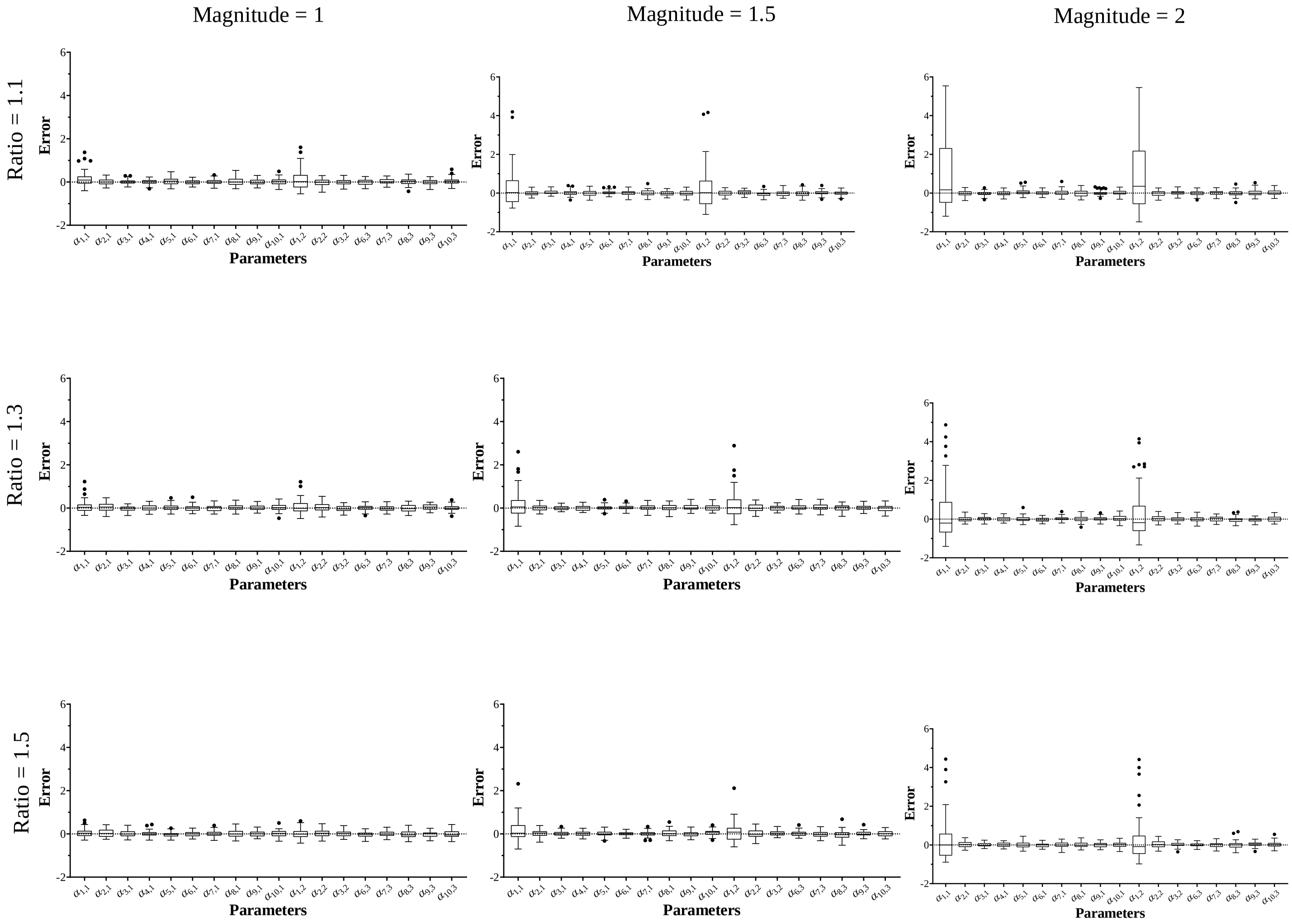


Figure 9

Figure 9. Boxplots summarizing the errors of the item discrimination estimates for $N = 1,000$ under full-information maximum likelihood (FIML), from Study 2. The plots in each row differ in magnitude of Item 1's discriminations; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

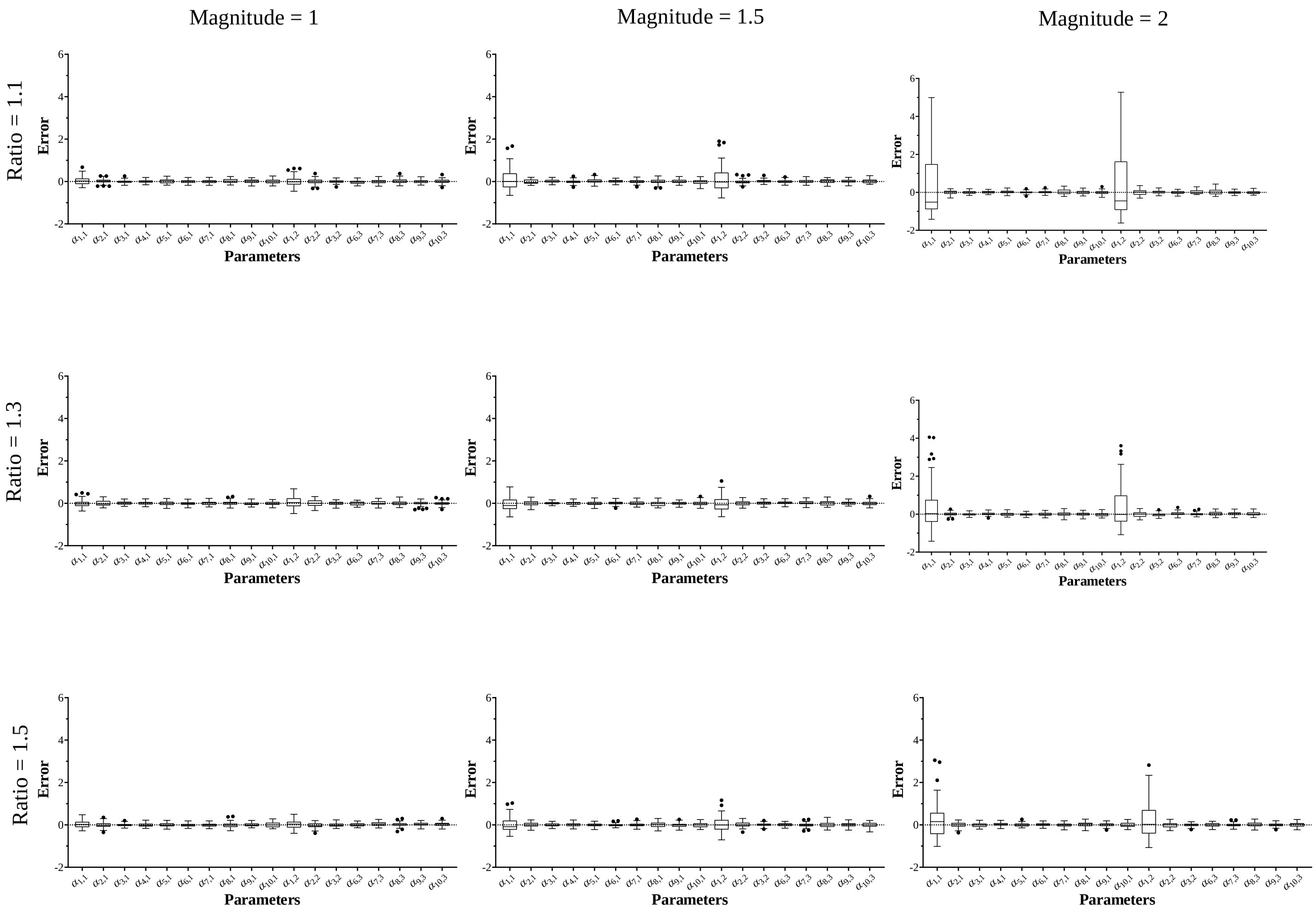


Figure 10

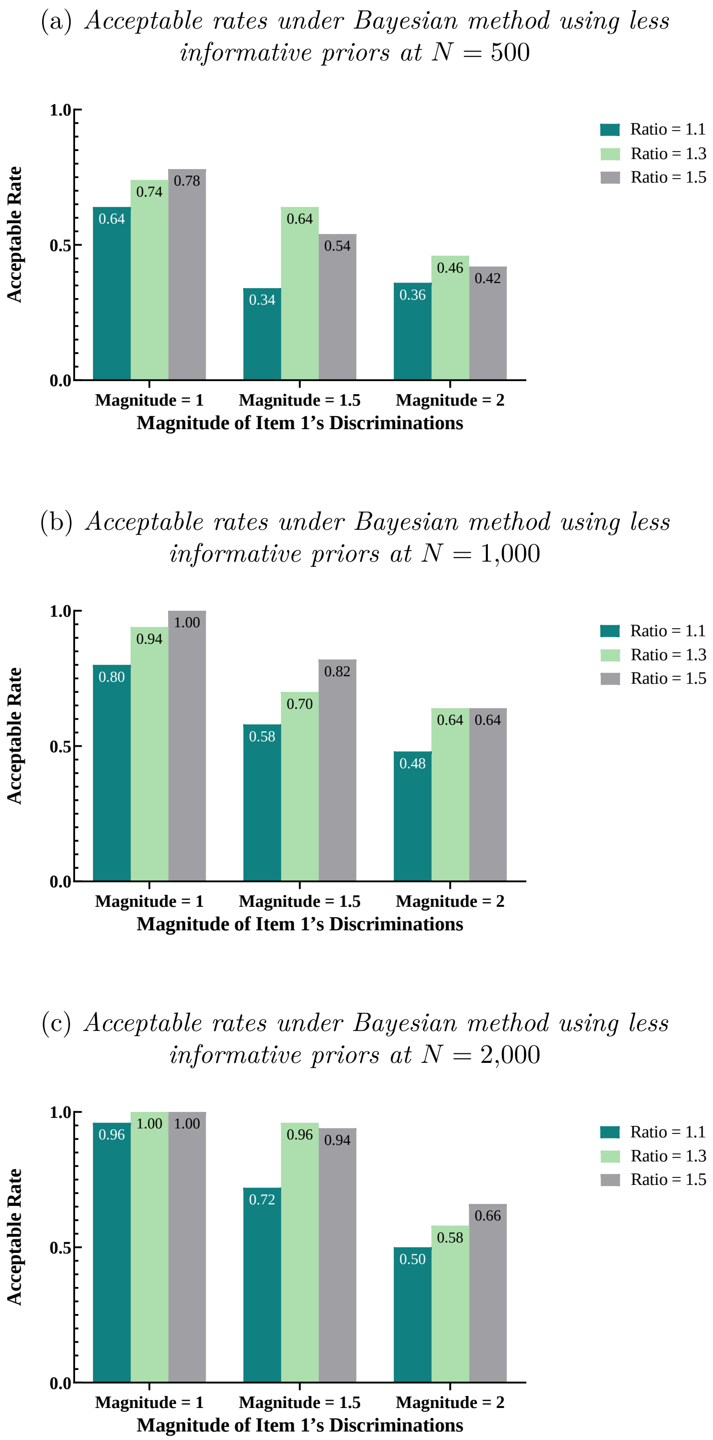
Figure 10. Boxplots summarizing the errors of the item discrimination estimates for $N = 2{,}000$ under full-information maximum likelihood (FIML), from Study 2. The plots in each row differ in magnitude of Item 1's discriminations; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

increased, similar to the pattern seen under FIML. For instance, when ratio = 1.5, the AR dropped to .42 for magnitude = 2, compared with .54 for magnitude = 1.5 and .78 for magnitude = 1. This pattern was consistently observed across the other ratio conditions, indicating that, as the magnitude of Item1's discriminations increased, the ARs decreased.

When the sample size increased to 1,000 and 2,000 (as illustrated in Figures 11b and 11c, respectively), a general increase in ARs was observed, irrespective of the ratio and the magnitude of Item 1's discriminations. Nevertheless, fluctuations in ARs across the magnitude conditions persisted, indicating that the presence of larger item discriminations can exacerbate the estimation challenges arising from when the within-item discriminations are similar.

The observations in ARs under LessInfo provides further support for the influential role that the magnitude of within-item discriminations plays in the parameter estimation of the bifactor model. The results of the parameter recovery reviewed next further reveal the estimation accuracy and stability of less informative priors as the magnitude of Item'1 discrimination changes.

Parameter Recovery: Errors of the Parameter Estimates. For $N = 500$, the distribution of the errors related to the item discrimination estimates under Bayesian method using less informative priors (LessInfo) is summarized in Figure 12. The results revealed that, when the within-item discriminations for Item 1 were very similar (i.e., ratio = 1.1), the difference in estimation accuracy between Item 1's discrimination parameters and those of the other items became more pronounced as Item 1's discriminations increased in magnitude. For example, when magnitude = 1.5, the range of the errors was 5.36 with an LUL of (-1.02, 4.34) for $\alpha_{1,1}$. In contrast, the other items' discriminations on the primary dimension had a largest error range of 1.23 (LUL[-0.53, 0.70]) and the smallest of 0.62 (LUL[-0.22, 0.40]), which were for $\alpha_{8,1}$ and $\alpha_{9,1}$, respectively. When the magnitude increased to 2, the error range for $\alpha_{1,1}$ increased to 5.99 (LUL[-2.20, 3.79]). However, the error ranges for the other items' discriminations on the primary dimension



(a) *Acceptable rates under Bayesian method using less informative priors at* $N = 500$

(b) *Acceptable rates under Bayesian method using less informative priors at* $N = 1{,}000$

(c) *Acceptable rates under Bayesian method using less informative priors at* $N = 2{,}000$

Figure 11

Figure 11. Visualizations of the acceptable rates (ARs) under the Bayesian method using less informative priors (LessInfo), from Study 2. The magnitudes of within-item discriminations are represented along the horizontal axis; the proportion of runs in which the item discrimination estimates were acceptable are represented along the vertical axis. Ratios are represented in different colors.

remained relatively stable. Recall that at baseline (i.e., magnitude = 1), the range was 2.82 (LUL[−0.59, 2.22] for $\alpha_{1,1}$, with the largest and smallest ranges across the other parameters on the primary dimension being 1.19 (LUL[−0.44, 0.75]) and 0.60 (LUL[−0.29, 0.31]), respectively. Moreover, it was observed that the medians of the errors associated with Item 1 also increased at magnitudes of 1.5 and 2 compared with at a magnitude of 1. For example, the median error for $\alpha_{1,1}$ was 1.11 at a magnitude of 1.5, and 0.71 at a magnitude of 2, compared with 0.30 at a magnitude of 1. This suggests that the estimates for Item 1's discrimination parameters are increasingly overestimated under LessInfo as the magnitude of its within-item discriminations rises.

Regarding the discriminations on the secondary dimension, similar trends were evident. For example, with a magnitude of 2, the error range for $\alpha_{1,2}$ was 6.33 (LUL[−2.41, 3.92]), whereas the largest and smallest error ranges for the other discrimination parameters on the secondary dimension were 1.11 (LUL[−0.55, 0.56]) for $\alpha_{2,2}$ and 0.57 (LUL[−0.26, 0.31]) for $\alpha_{3,2}$, respectively. In contract, at baseline, the range was 3.32 (LUL[−0.61, 2.71]) for $\alpha_{1,2}$. Additionally, the median of the errors associated with $\alpha_{1,2}$ was 1.18 for magnitude = 1.5 and 0.56 for magnitude = 2, as opposed to 0.21 for magnitude = 1, which further confirmed the observation that as the magnitude of Item 1's discriminations increased, the overestimation of Item 1's discriminations became greater.

When the within-item discriminations for Item 1 became more differentiated (ratios of 1.3 and 1.5, as shown in the second and third rows of Figure 12), the trends observed aligned with those at ratio = 1.1. Specifically, as the magnitude of Item 1's discriminations increased, both the full ranges and LULs of the errors associated with Item 1's parameter estimates increased noticeably compared with that observed at the baseline magnitude of 1. Conversely, the ranges and LULs of the errors concerning the parameter estimates for the other items remained stable. This indicates that the magnitude of an item's discriminations has a tangible impact on the estimation difficulties arising from when its within-item discriminations are similar.

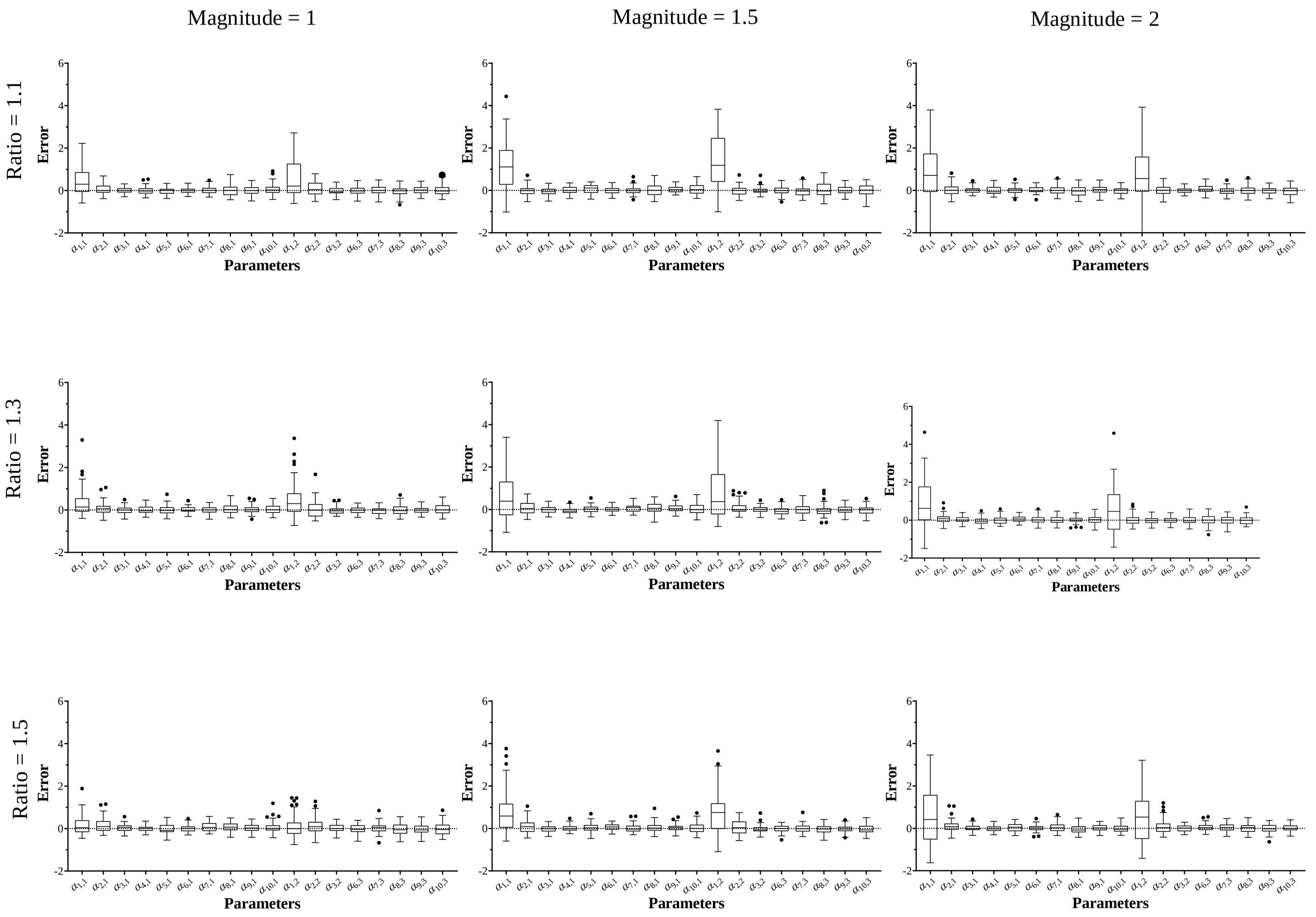


Figure 12

Figure 12. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under the Bayesian method using less informative priors (LessInfo), from Study 2. The plots in each row differ in magnitude of Item 1's discriminations; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

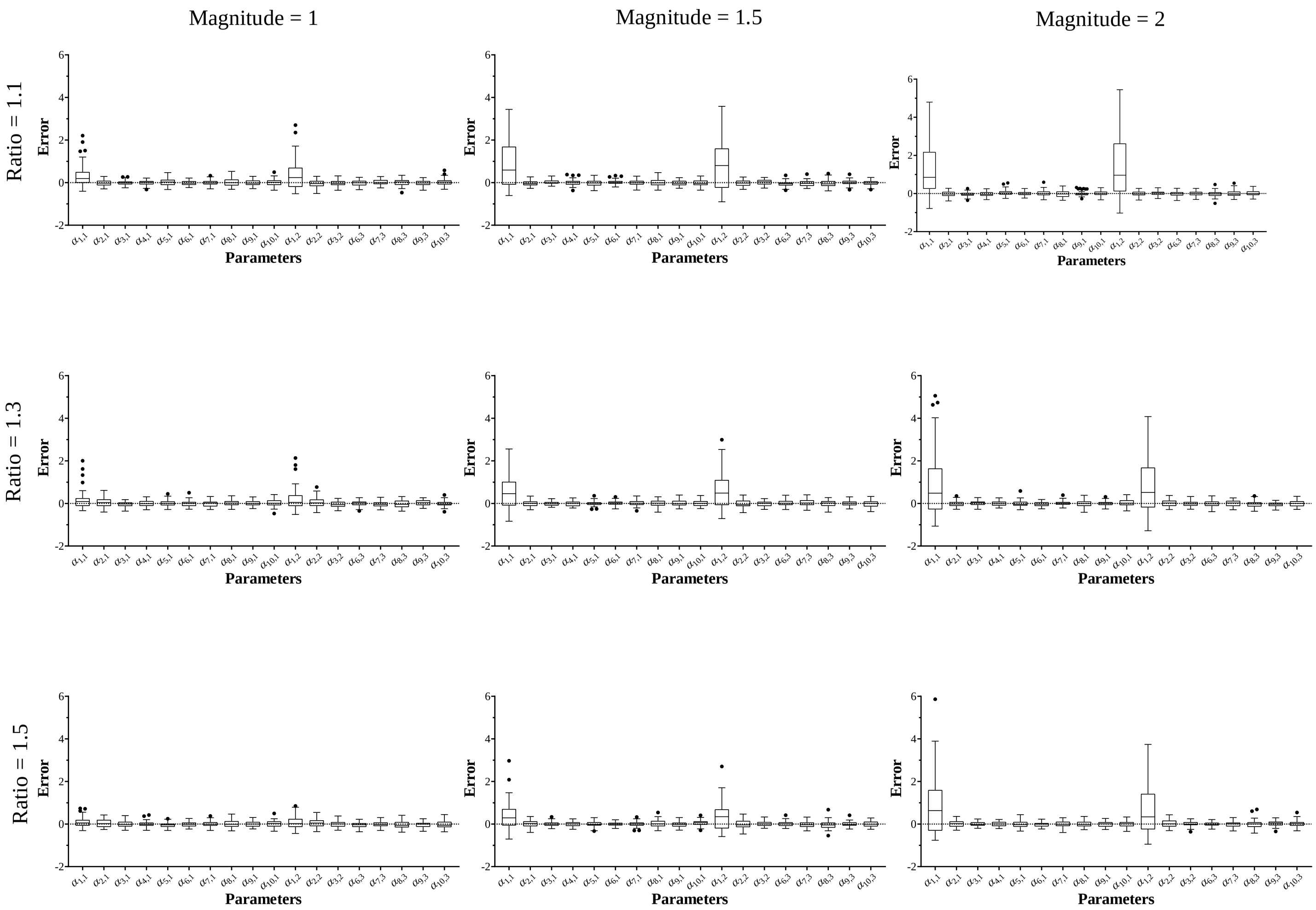


Figure 13

Figure 13. Boxplots summarizing the errors of the item discrimination estimates for $N = 1{,}000$ under the Bayesian method using less informative priors (LessInfo), from Study 2. The plots in each row differ in magnitude of Item 1's discriminations; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

As the sample size increased to 1,000 (as shown in Figure 13) and 2,000 (as shown in Figure 14), the patterns observed under each ratio condition remained consistent. However, there was a noticeable reduction in the ranges and LULs related to Item 1 compared with those at $N = 500$. For instance, when $N = 2,000$, the error range was 3.29 (LUL[−0.58, 2.70)]) for $\alpha_{1,1}$ and 3.79 (LUL[−0.72, 3.06)]) for $\alpha_{1,2}$ for ratio = 1.1 and magnitude = 1.5. In contrast, when $N = 500$, that error range was 5.36 (LUL[−1.02, 4.34)]) for $\alpha_{1,1}$ and 4.83 (LUL[−1.01, 3.82)]) for $\alpha_{1,2}$; and when $N =$ 1,000, that error range was 4.05 (LUL[−0.61, 3.44)]) for $\alpha_{1,1}$ and 4.48 (LUL[−0.90, 3.58)]) for $\alpha_{1,2}$. This trend was generally observed across various combinations of ratio and magnitude conditions, suggesting that an increase in sample size has the potential to mitigate the estimation challenges encountered when an item's discriminations are similar and large in magnitude.

Bayesian Method Using Adaptive Informative Priors

Acceptable Rates. The acceptable rates (ARs) for the Bayesian method using adaptive informative priors (AdptInfo) are summarized in Figure 15. At $N = 500$ (as shown in Figure 15a), slight decrease in AR was observed when the magnitude of Item 1's discriminations increased from 1 to 1.5. Specifically, the average AR across the three ratio conditions was .98 and .94 for magnitude = 1 and 1.5, respectively. However, when the magnitude of Item 1's discriminations further increased to 2, a noticeable reduction in ARs was seen, with the averages AR across the ratio conditions reducing to .43. The most noticeable drop was observed at the ratio of 1.1, indicating that the magnitude and the similarity of Item 1's discriminations jointly affected obtaining potentially acceptable results.

When the sample size increased to 1,000 (as illustrated in Figures 15b), similar trends were observed. To be specific, under magnitude = 1 and 1.5, the ARs were similar and nearly perfect (i.e. $\geq .96$), regardless of the ratio of Item 1's discriminations. However, for magnitude = 2, noticeable decreases were seen in ARs, especially at ratio = 1.1, where the AR dropped to .58, in contrast to .96 at magnitude = 1.5 and .98 at magnitude = 1.

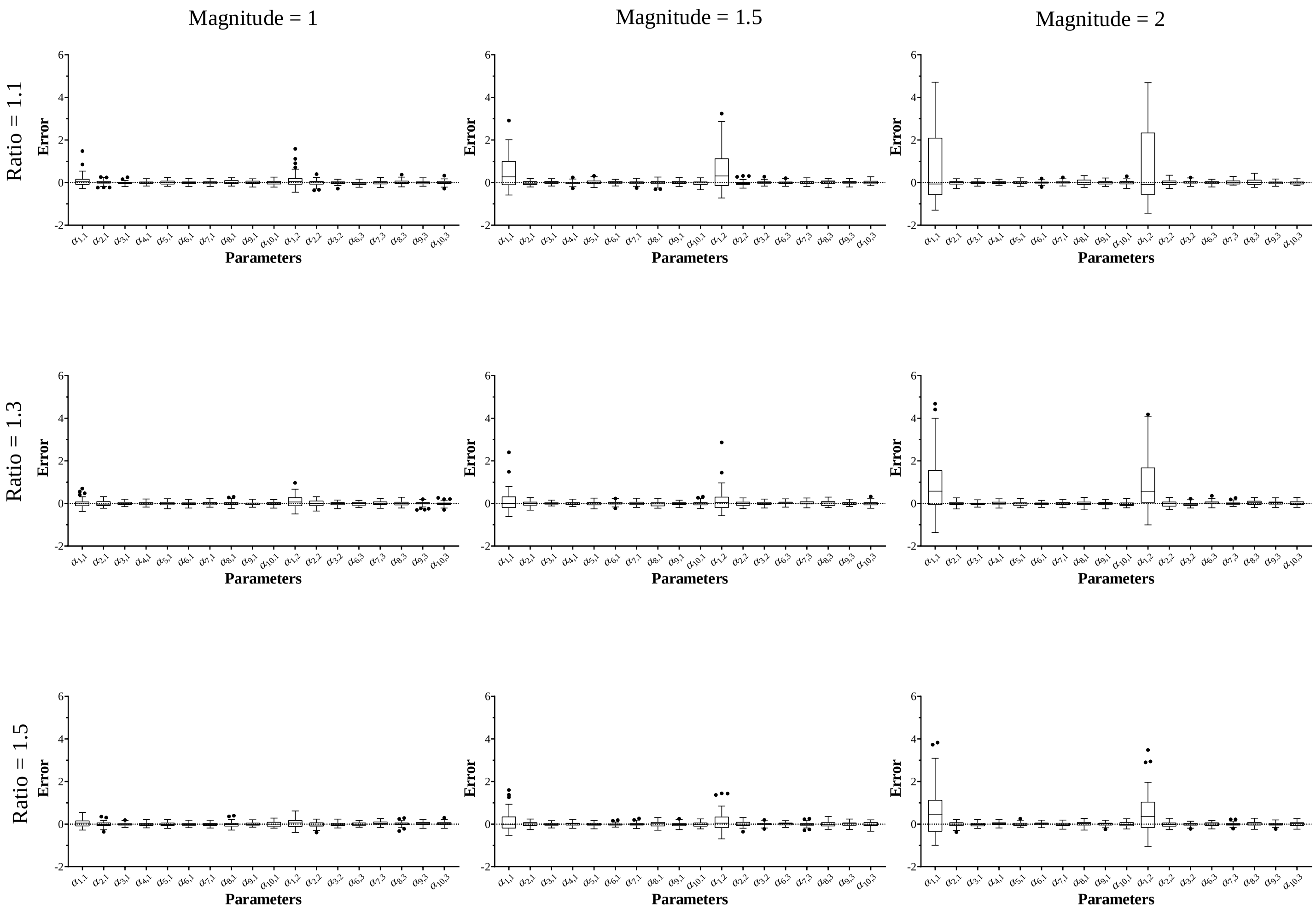
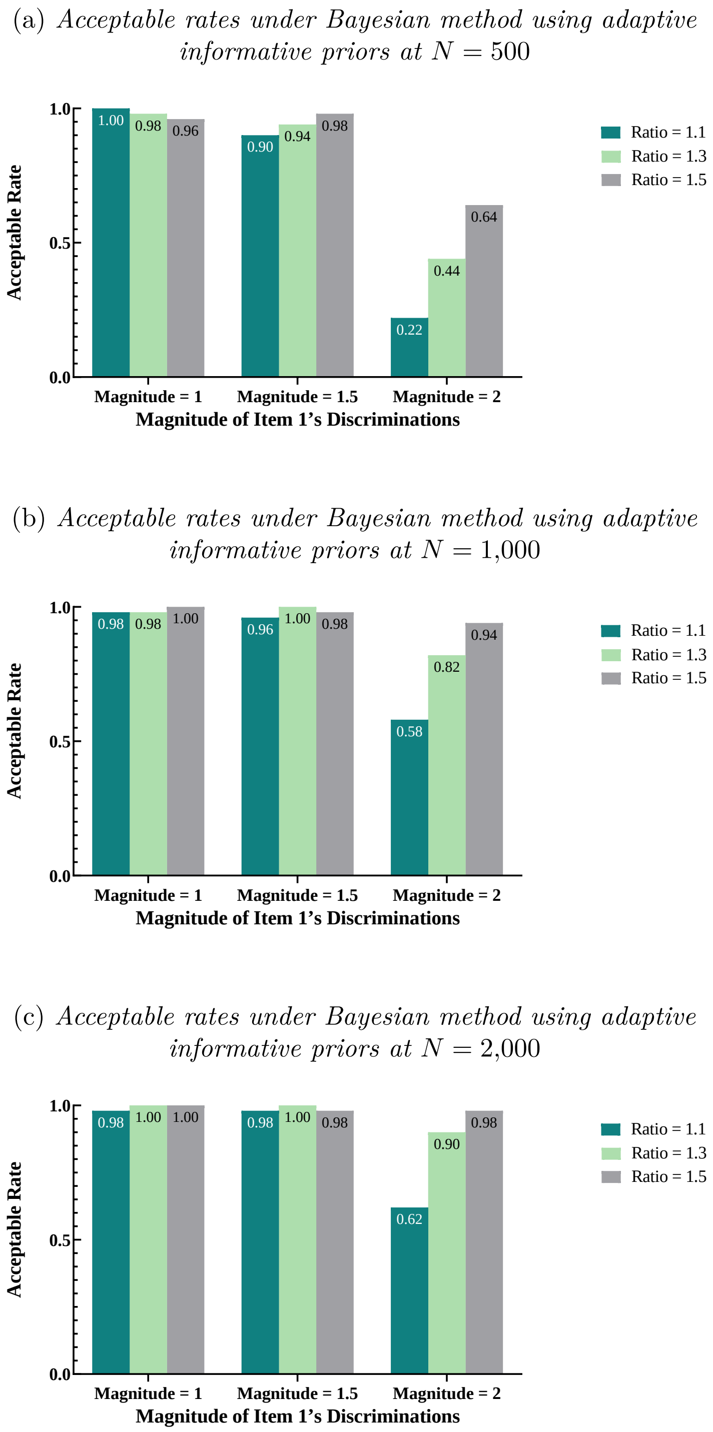


Figure 14

Figure 14. Boxplots summarizing the errors of the item discrimination estimates for $N = 2{,}000$ under the Bayesian method using less informative priors (LessInfo), from Study 2. The plots in each row differ in magnitude of Item 1's discriminations; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.



(a) *Acceptable rates under Bayesian method using adaptive informative priors at* $N = 500$

(b) *Acceptable rates under Bayesian method using adaptive informative priors at* $N = 1{,}000$

(c) *Acceptable rates under Bayesian method using adaptive informative priors at* $N = 2{,}000$

Figure 15

Figure 15. Visualizations of the acceptable rates (ARs) under the Bayesian method using adaptive informative priors (AdptInfo), from Study 2. The magnitudes of within-item discriminations are represented along the horizontal axis; the proportion of runs in which the item discrimination estimates were acceptable are represented along the vertical axis. Ratios are represented in different colors.

Nonetheless, the decreases observed were less pronounced compared with those seen when $N = 500$.

As the sample size further increased to 2,000 (as shown in Figures 15c), the ARs exhibited further improvement across all magnitude and ratio conditions. However, it is critical to note that when magnitude = 2, the AR declined to .62 for ratio = 1.1, in contrast to .98 when magnitude = 1.5 and .98 when magnitude = 1, confirming the existence of an interaction effect between the magnitude and the similarity of Item 1's discriminations on obtaining acceptable parameter estimates.

Although the observations in ARs under AdptInfo further supported the importance of the magnitude of within-item discriminations in the parameter estimation of the bifactor model, the results suggest that the role of the magnitude is negligible when magnitude = 1 or 1.5, regardless of the sample size and the ratio of within-item discriminations. It is only when the within-item discriminations are considerably large in magnitude (for instance, magnitude = 2) that ARs are affected, which diverges from the patterns observed under FIML and LessInfo. The review of parameter recovery results next reveals whether the findings in ARs hold for the estimation accuracy and stability under AdptInfo.

Parameter Recovery: Errors of the Parameter Estimates. For $N = 500$, the distribution of the errors related to the item discrimination estimates under the Bayesian method using adptive-informative priors (AdptInfo) is summarized in Figure 16. When the ratio of the within-item discriminations for Item 1 was 1.1, the difference in estimation accuracy between Item 1's discrimination parameters and those of the other items became more noticeable as Item 1's discriminations increased in magnitude. For example, when magnitude = 1.5, the range of the errors for $\alpha_{1,1}$ was 2.00 with an LUL of $(-1.16, 0.84)$, whereas the largest error range for the other items' discriminations on the primary dimension was 1.17 (LUL$[-0.51, 0.66]$) for $\alpha_{8,1}$ and the smallest was 0.62 (LUL$[-0.20, 0.41]$) for $\alpha_{9,1}$. As the magnitude increased to 2, the error range for $\alpha_{1,1}$ increased to 2.30 (LUL$[-2.26, 0.04]$), whereas the error ranges for the other items' discriminations on the primary dimension remained relatively stable. Recall that at baseline (i.e., magnitude = 1; from Study 1), the range of the errors for $\alpha_{1,1}$ was 1.11 (LUL[−0.64, 0.47]), with no noticeable differences in the ranges of the errors and LULs between $\alpha_{1,1}$ and the other items' discriminations on the primary dimension. However, Item 1's discrimination parameters were progressively underestimated at magnitudes of 1.5 and 2 in comparison to magnitude = 1. For example, the median error for $\alpha_{1,1}$ was −0.27 for magnitude = 1.5, and −1.06 for magnitude = 2, compared with −0.05 at a magnitude of 1. These observations indicate that, under AdptInfo, the estimates for Item 1's discrimination parameters were progressively underestimated as the magnitude of its discriminations became stronger.

Regarding the discrimination parameters on the secondary dimension, similar trends were observed. In particular, as the magnitude of Item 1's discriminations increased, the errors and LULs associated with $\alpha_{1,2}$ became more noticeablely underestimated relative to the baseline. For example, when the magnitude was 2, the range of the errors for $\alpha_{1,2}$ was 2.09 (LUL[−2.50, −0.42]), whereas the largest and smallest ranges of the errors for the other discrimination parameters on the secondary dimension were 1.03 (LUL[−0.44, 0.59]) for $\alpha_{2,2}$ and 0.54 (LUL[−0.18, 0.36]) for $\alpha_{3,2}$, respectively. In contrast, at baseline, the range was 1.50 (LUL[−0.70, 0.80]) for $\alpha_{1,2}$. Moreover, the median error corresponding to $\alpha_{1,2}$ was −0.45 for magnitude = 1.5 and −1.32 for magnitude = 2, as opposed to −0.13 for magnitude = 1. These findings further confirmed the observation that, as the magnitude of Item 1's discriminations increased, the estimates for Item 1's discrimination parameters under AdptInfo were subject to more pronounced underestimation.

When the within-item discriminations for Item 1 became more differentiated at ratios of 1.3 and 1.5 (as shown in the second and third rows of Figure 16), the trends observed aligned with those of ratio = 1.1. Specifically, with an increase in the magnitude of Item 1's discriminations, both the full ranges and LULs of the errors associated with Item 1's parameter estimates increased compared with the baseline magnitude of 1. In contrast, the ranges and LULs of the errors concerning the parameter estimates for the other items

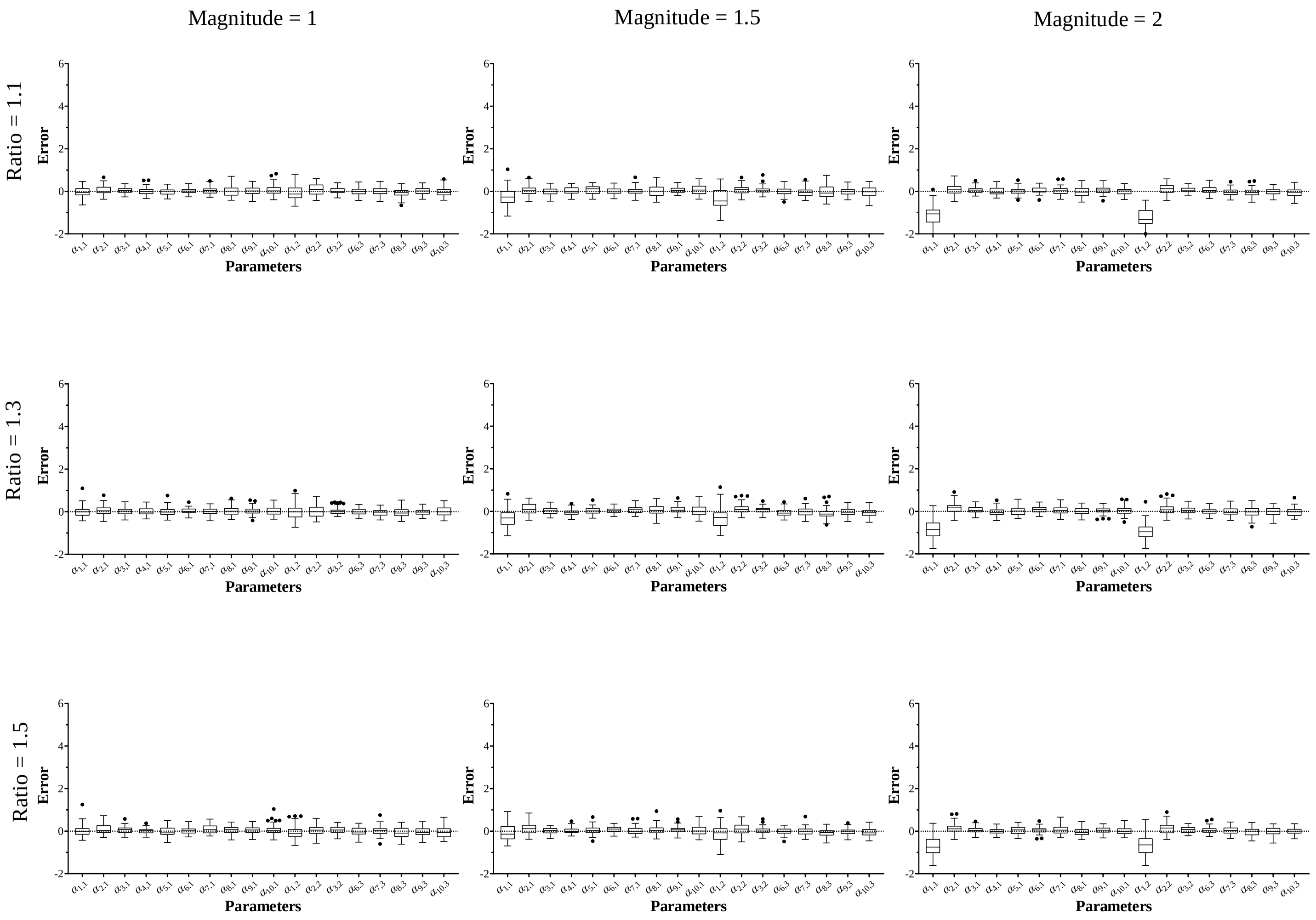


Figure 16

Figure 16. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under the Bayesian method using adaptive informative priors (AdptInfo), from Study 2. The plots in each row differ in magnitude of Item 1's discriminations; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

remained stable.

While Item 1's discrimination parameters tended to be underestimated as the magnitude of its discriminations increased, the underestimation issue was reduced as the ratio of Item 1's discriminations became more differentiated, suggesting that both the ratio and magnitude of the item's discriminations play a role in the estimation of a bifactor model based on adaptive informative priors.

As the sample size increased to 1,000 (as shown in Figure 17) and 2,000 (Figure 18), the estimation accuracy improved for all discrimination parameters, including those for Item 1. However, a discrepancy in estimation accuracy between Item 1 and the other items persisted, especially when Item 1's discriminations were similar in strength and large in magnitude. For example, when $N = 2{,}000$, the range of the errors was 1.42 (LUL[-0.70, 0.72]) for $\alpha_{1,1}$ and 1.69 (LUL[-0.80, 0.89]) for $\alpha_{1,2}$ when ratio = 1.1 and magnitude = 1.5. In comparison, when $N = 1{,}000$, the range of the errors was 1.40 (LUL[-0.81, 0.59]) for $\alpha_{1,1}$ and 2.04 (LUL[-1.16, 0.87]) for $\alpha_{1,2}$. When contrasted with the case where $N = 500$, where the range of the errors was 2.00 (LUL[-1.16, 0.84]) for $\alpha_{1,1}$ and 1.95 (LUL[-1.37, 0.58]) for $\alpha_{1,2}$, it is evident that larger sample sizes helped to improve the estimation of the problematic item. This pattern was consistently observed across almost all combinations of ratio and magnitude conditions, indicating that a larger sample size can mitigate underestimation under AdptInfo when an item's discriminations are large in magnitude and similar in strength.

Comparison across the three estimation methods. The results based on the three estimation methods consistently demonstrate that the estimation challenges arising from when an item's discriminations being similar become more evident as the magnitude of the items' discriminations increased. Specifically, for a fixed sample size and ratio of within-item discriminations, the estimation of Item 1's discrimination parameters became increasingly worse than the remaining items as the magnitude of Item 1' discrimination parameters increased. However, the extent of the discrepancy in estimation

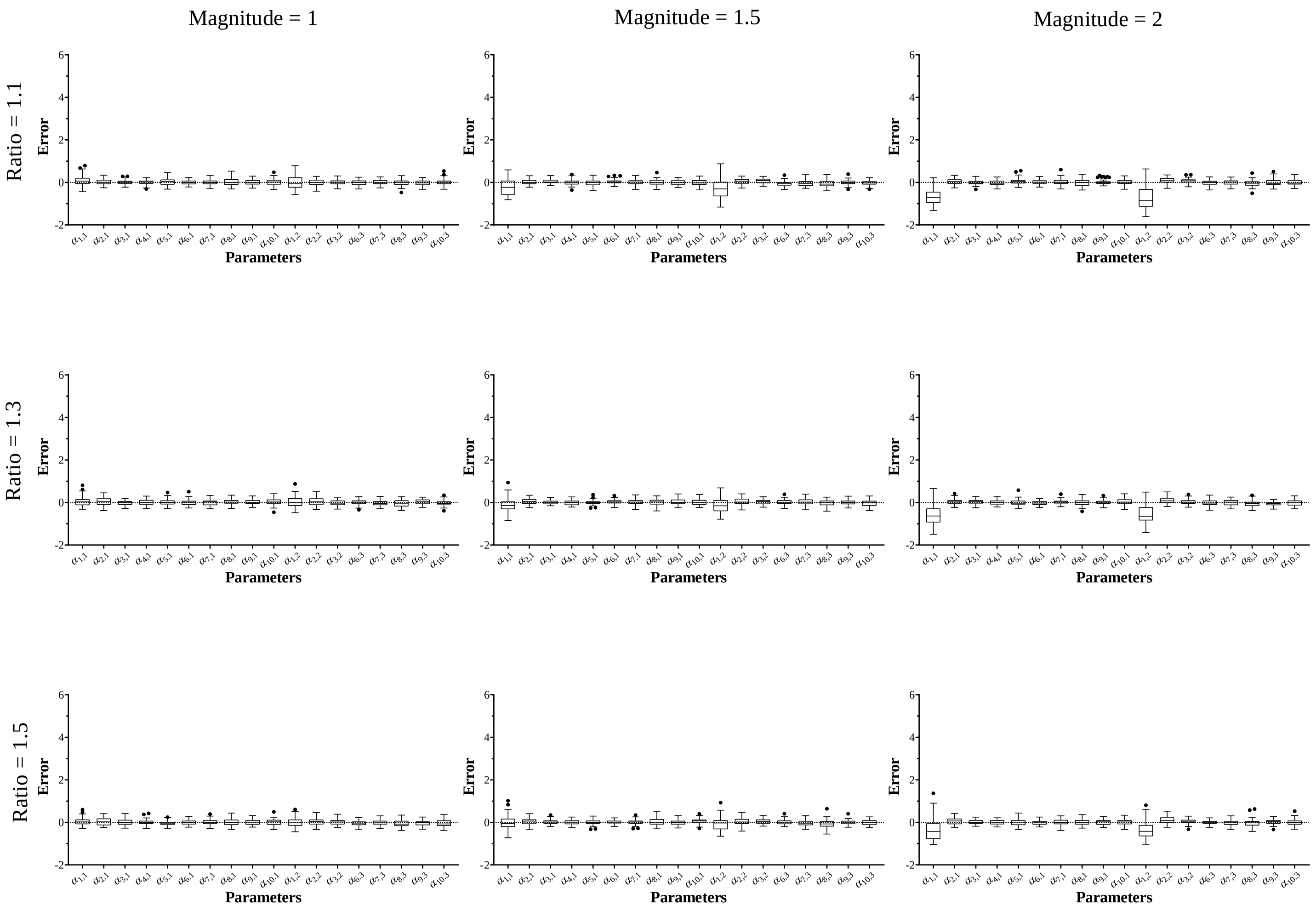


Figure 17

Figure 17. Boxplots summarizing the errors of the item discrimination estimates for $N = 1{,}000$ under Bayesian method using adaptive informative priors (AdptInfo), from Study 2. The plots in each row differ in magnitude of Item 1's discriminations; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

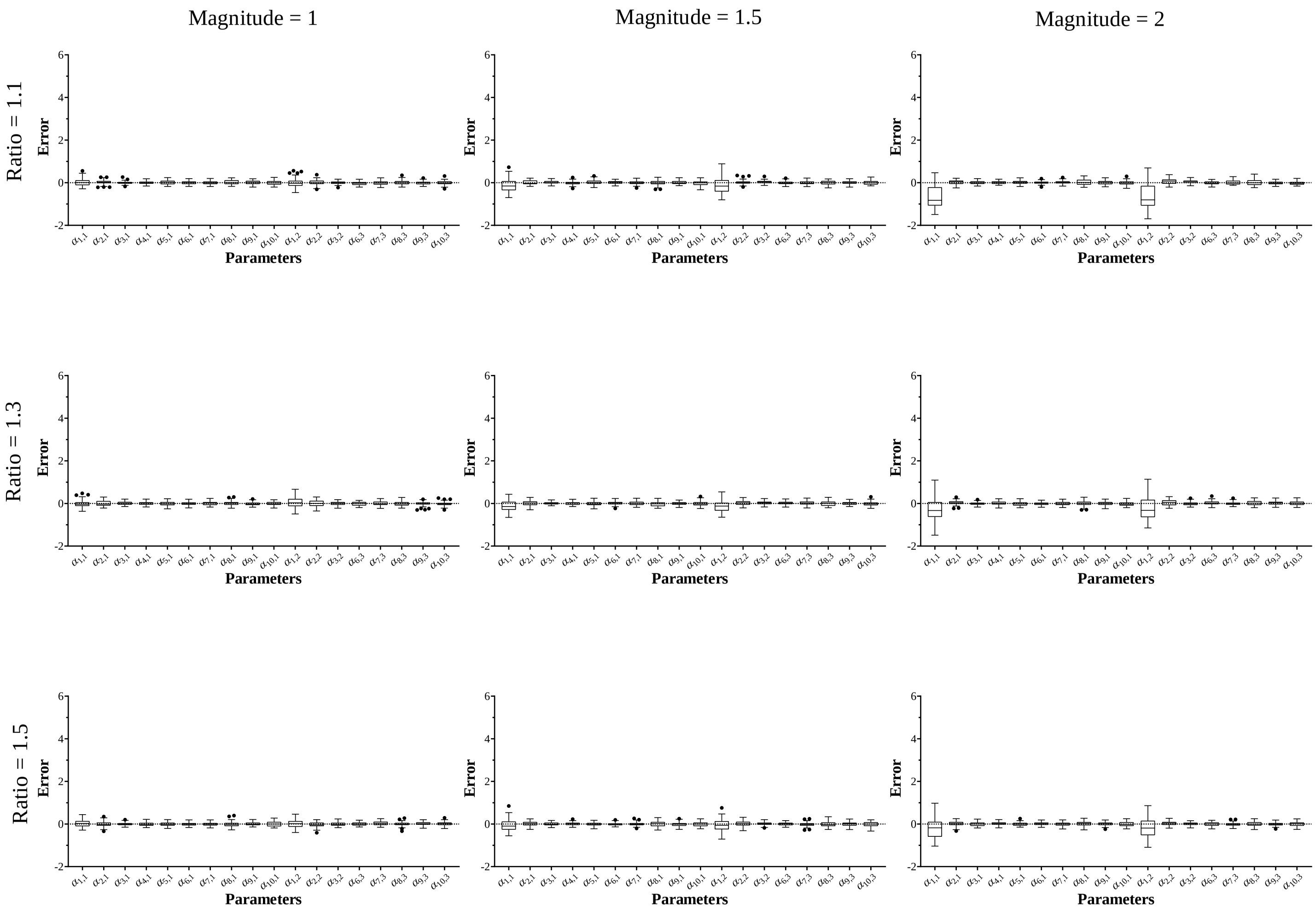


Figure 18

Figure 18. Boxplots summarizing the errors of the item discrimination estimates for $N = 2{,}000$ under Bayesian method using adaptive informative priors (AdptInfo), from Study 2. The plots in each row differ in magnitude of Item 1's discriminations; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

between Item 1 and the other items varied depending on the estimation method employed. Noticeably, the disparity in estimation accuracy was less severe under FIML compared with LessInfo, indicating that FIML could yield more accurate estimates for the bifactor model across a range of ratio, magnitude, and sample size conditions.

While the estimation of Item 1's discrimination parameters worsened as the magnitude of these parameters increased, the use of more informative priors under AdptInfo led to a significant improvement in the estimation accuracy of these parameters compared with using FIML and LessInfo. However, it is important to note that the parameters tended to be underestimated with AdptInfo.

Summary of Study 2

Study 2 demonstrated that an increase in the magnitude of an items' discriminations can exacerbate the estimation difficulties that arise when the item's discriminations are similar in strength. Furthermore, sample size influences the severity of these estimation challenges. Specifically, for a given level of similarity and magnitude of within-item discriminations, the estimation difficulties amplify as the sample size decreases. In addition, Study 2 evaluated the effectiveness of different estimation methods in estimating the discrimination parameters of the bifactor model under more challenging situations than those seen in Study 1. The results highlight some of the advantages of AdptInfo over both FIML and LessInfo, as AdptInfo consistently yielded more accurate and stable estimates. However, it is critical to note that AdptInfo tended to underestimate the discrimination parameters for the problematic items. In summary, the findings of Study 2 suggest that the estimation accuracy of the bifactor model is affected by the similarity and magnitude of an items' discriminations, and sample size. Consistent with Study 1, Study 2 reinforced that the choice of estimation method matters, as it can have an impact on the accuracy of the results.

Results of Study 3

Study 3 further explored whether the effect of within-item discrimination similarity on the estimates depended on item targetedness. It examined the conditions in which the magnitude of the item discriminations was ideal but the similarity of the within-item discriminations and the item targetedness were manipulated. Specifically, the factors manipulated in this study included the sample size (500, 1,000, and 2,000), the ratio of Item 1's discriminations (1.1, 1.3, and 1.5), and the item targetedness of Item 1 (the higher categories being more represented in the responses than the lower categories and vice versa). Next, I discuss the results by estimation methods, followed by a comparison of the performance of the three estimation methods and a summary of the main findings. The results of this study were compared with those of Study 1 because the latter is equivalent to an item targetedness condition in which all items including Item 1 are ideally targeted to the respondents.

Full-Information Maximum Likelihood Estimation

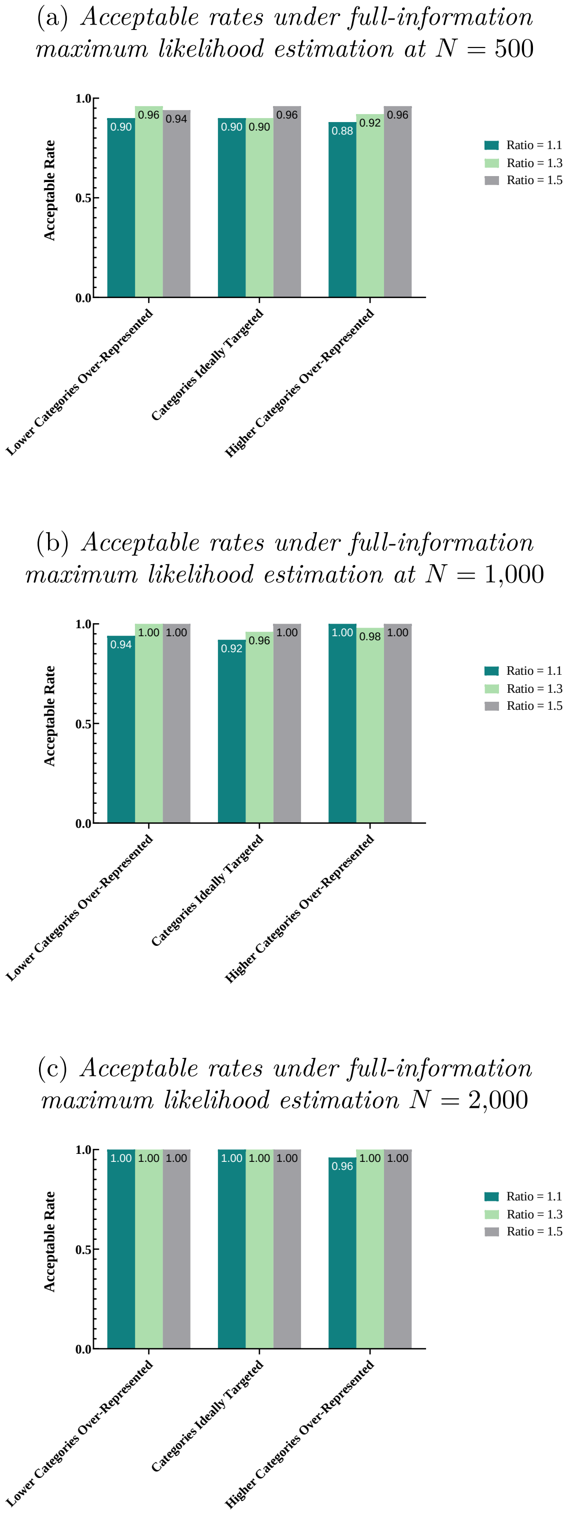
Acceptable Rates. The ARs for full-information maximum likelihood (FIML) estimation are summarized in Figure 19. In each plot, how well Item 1 was targeted to the respondents is represented along the x-axis, the acceptable rates are represented along the y-axis, and the ratios are represented by different colors. The results for when Item 1 was ideally targeted originate from Study 1 and serve as the baseline.

At $N = 500$, the average ARs across the ratios were nearly the same for the three item targetedness conditions (as shown in Figure 19a). Specifically, at baseline, the average AR across the three ratios was .92, with a range of .90 to .94; when the lower categories had a greater representation in the data than the higher categories for Item 1, the average AR was .93, with a range of .90 to .96; when the higher categories had a greater representation in the data than then lower categories for Item 1, the average AR was .92, with a range of .88 to .96. These observations suggest that the variation in the targetedness of Item 1 had a negligible effect on obtaining extreme estimates, warning messages, or both for the bifactor model. When the sample size increased to 1,000 and 2,000 (as depicted in Figures 19b and 19c, respectively), similar patterns were observed. That is, for any given ratio, the ARs did not change noticeably across the three targetedness conditions.

In sum, the AR findings suggest that item targetedness played a minimal role in obtaining convergence warnings, extreme estimates, or both. The results of the parameter recovery reviewed next further reveal whether the item targetedness affects the estimation accuracy and stability of the bifactor model under FIML.

Parameter Recovery: Errors of the Parameter Estimates. For $N = 500$, the distribution of the errors related to the item discrimination estimates under FIML is summarized in Figure 20. In each plot, the item discrimination parameters are represented along the x-axis (where α_{jd} is item j's discrimination on dimension d), and the errors are represented along the y-axis. The plots within the same row differ in Item 1's targetedness, while the plots within the same column differ in how similar Item 1's discriminations are. If item targetedness does not matter, then the distributions of the errors should look similar across the different item targetedness conditions.

When the ratio of Item 1's discriminations was 1.1, the discrepancies in the estimation accuracy between Item 1's discrimination parameters and those of the other items appeared stable despite changes in Item 1's targetedness, as demonstrated by in the plots in the first row of Figure 20). Specifically, when Item 1's lower categories had a greater representation in the data than the higher categories (shown in the first column), the range of the errors was 1.65, with an LUL of $(-0.57, 1.08)$, for $\alpha_{1,1}$, whereas among the other items' discriminations on the primary dimension (i.e., α_{j1} where $j \neq 1$), the largest error range was 1.14 (LUL$[-0.52, 0.63]$) and the smallest was 0.73 (LUL$[-0.34, 0.39]$), which were for $\alpha_{5,1}$ and $\alpha_{3,1}$, respectively. Similarly, when Item 1's higher categories had a greater representation in the data than the lower categories (shown in the third column), the range of the errors for $\alpha_{1,1}$ was 1.51, with an LUL of $(-0.53, 0.98)$, whereas the ranges of the errors for the other items' discriminations on the primary dimension fell between



(a) *Acceptable rates under full-information maximum likelihood estimation at $N = 500$*

(b) *Acceptable rates under full-information maximum likelihood estimation at $N = 1{,}000$*

(c) *Acceptable rates under full-information maximum likelihood estimation $N = 2{,}000$*

Figure 19

Figure 19. Visualizations of the acceptable rates (ARs) under full-information maximum likelihood estimation (FIML), from Study 3. The targetedness of Item 1 is represented along the horizontal axis; the proportion of runs in which the item discrimination estimates were acceptable are represented along the vertical axis. Ratios are represented in different colors.

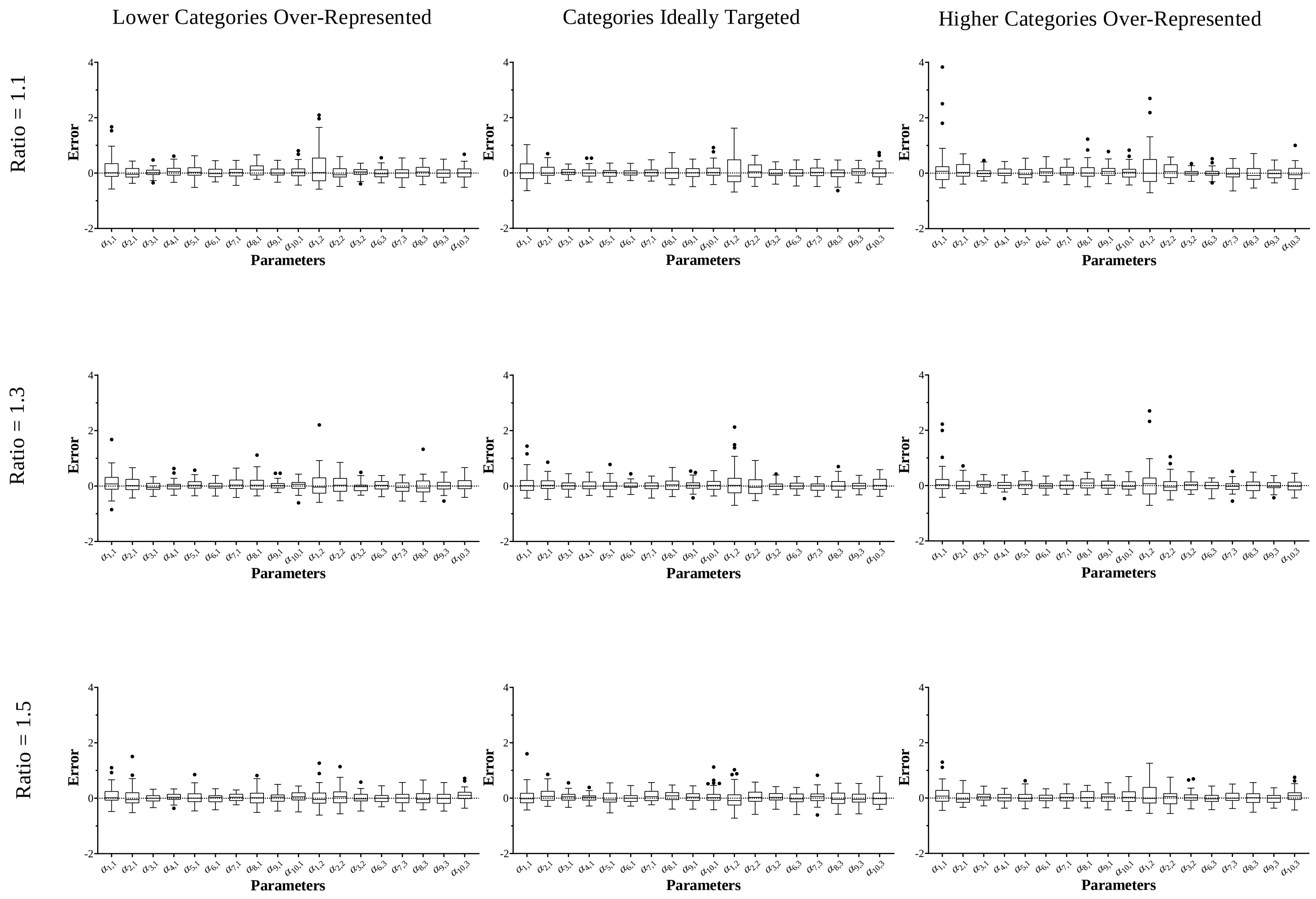


Figure 20

Figure 20. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under full-information maximum (FIML) estimation, from Study 3. The plots in each row differ in Item 1's targetedness; the plots in each column differ in similarity of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

0.74 (LUL[−0.28, 0.45]) for $\alpha_{3,1}$ and 1.20 (LUL[−0.49, 0.71]) for $\alpha_{8,1}$. Notice that the range of the errors related to $\alpha_{1,1}$, as well as the largest and smallest ranges of the errors across the other parameters on the primary dimension, were all similar between the two targetedness conditions. More importantly, comparable values were seen at baseline (shown in the middle column).

With regards to the discriminations on the secondary dimension, similar patterns were observed. Particularly, when Item 1's lower categories had a greater representation in the data (shown in the first column), the range of the errors for $\alpha_{1,2}$ was 2.40 (LUL[−0.58, 1.82], while the largest and smallest ranges across the other discrimination parameters for the secondary dimension were 1.15 (LUL[−0.51, 0.64]) for $\alpha_{10,3}$ and 0.71 (LUL[−0.36, 0.36]) for $\alpha_{3,2}$, respectively. When Item 1's higher categories had a greater representation in the data (shown in the third column), the range of the errors for $\alpha_{1,2}$ was 2.44 (LUL[−0.71, 1.73], while the largest and smallest ranges across the other discrimination parameters for the secondary dimension were 1.36 (LUL[−0.58, 0.77]) for $\alpha_{10,3}$ and 0.59 (LUL[−0.30, 0.30]) for $\alpha_{3,2}$, respectively. Similar to these two targetedness conditions, when Item 1's categories were ideally targeted to the respondents (shown in the middle column), the range of the errors was 2.31 (LUL[−0.69, 1.62]) for $\alpha_{1,2}$, with the largest and smallest ranges across the other discrimination parameters for the secondary dimension being 1.04 (LUL[−0.41, 0.64]) and 0.81 (LUL[−0.40, 0.40]), respectively. Overall, the results observed under ratio = 1.1 indicate that Item 1's targetedness did not have a critical effect on the estimation challenges posed by the empirical identification issue.

When the within-item discriminations for Item 1 became more differentiated, as seen at ratios of 1.3 and 1.5 (shown in the plots in the second and third rows of Figure 20), the trends observed were similar to those for ratio = 1.1. Specifically, even with changes in Item 1's targetedness, the full ranges and LULs of the errors related to Item 1's parameters, as well as those related to other items' parameters, remained stable. These observations confirmed that the targetedness of an item might not be a contributing factor to the estimation difficulties that emerge when the item's discriminations are similar in size. In other words, whether an item's discriminations being similar leads to estimation issues does not depend on the item's targetedness.

As the sample size increased to 1,000 (as shown in Figure 21) and 2,000 (see Figure 22), similar patterns continued to be observed under each ratio condition, albeit with an improvement in the estimation accuracy for all discrimination parameters. The results suggest that the targetedness of an item does not play a noticeable role in the empirical identification issue, reinforcing the conclusion that increasing the sample size can mitigate the estimation challenges that arise from when an item's discriminations are similar in strength.

Bayesian Method Using Less Informative Priors

Acceptable Rates. The ARs for Bayesian method using less informative priors (LessInfo) are summarized in Figure 23. At $N = 500$ (see Figure 23a), the average ARs across the ratios were approximately similar between the two off-targetedness conditions. Specifically, under the condition where Item 1's lower categories had a greater representation in the data than the higher categories (shown in the first column), the average AR across the three ratio conditions was .85, with a range of .82 to .90. Similarly, when Item 1's higher categories had a greater representation in the data than the lower categories (shown in the third column), the average AR across the three ratios was .85, with a range of .80 to .88. In contrast, at baseline (shown in the middle column), the average AR was .72, ranging from .64 to .78. As the sample size increased, the discrepancies in ARs between the baseline and the two off-targetedness scenarios diminished. In particular, when $N = 2{,}000$, the ARs were almost indistinguishable among the three targetedness conditions for any given ratio.

The findings regarding ARs indicate that, for LessInfo, the item targetedness plays a more impactful role in obtaining acceptable results compared with FIML, especially when dealing with smaller sample size. The results of the parameter recovery reviewed next

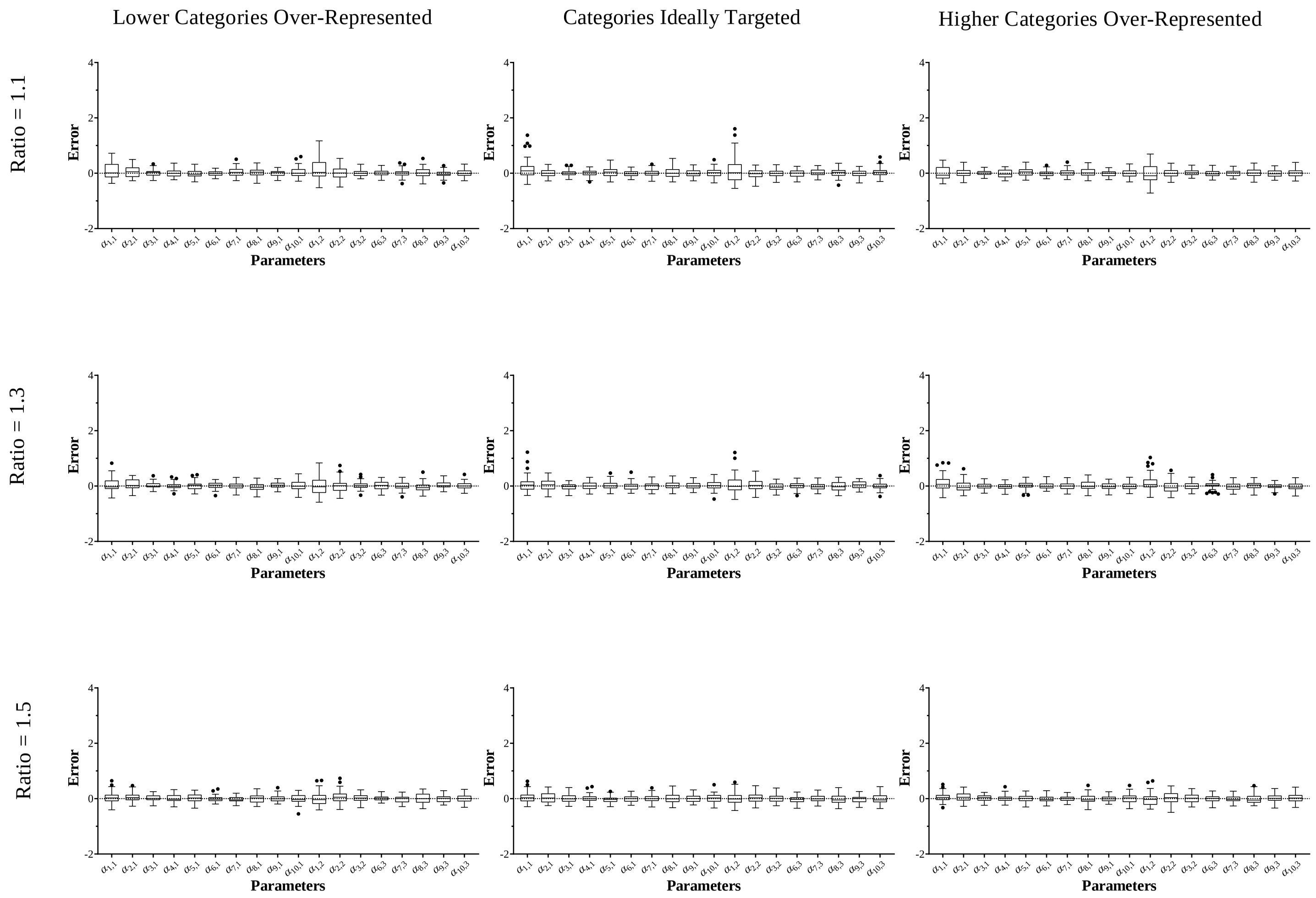


Figure 21

Figure 21. Boxplots summarizing the errors of the item discrimination estimates for $N = 1{,}000$ under full-information maximum likelihood (FIML), from Study 3. The plots in each row differ in Item 1's targetedness; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

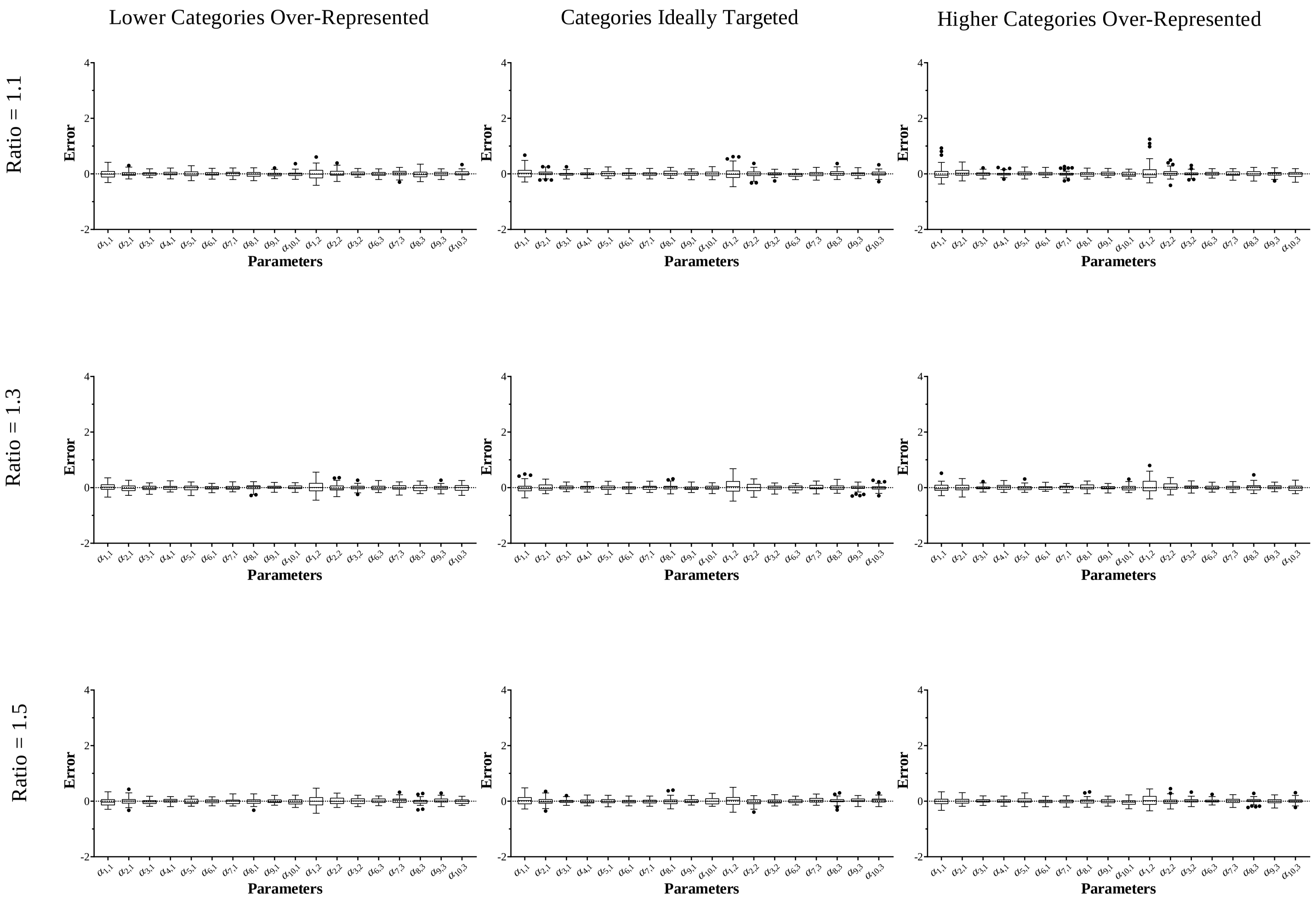


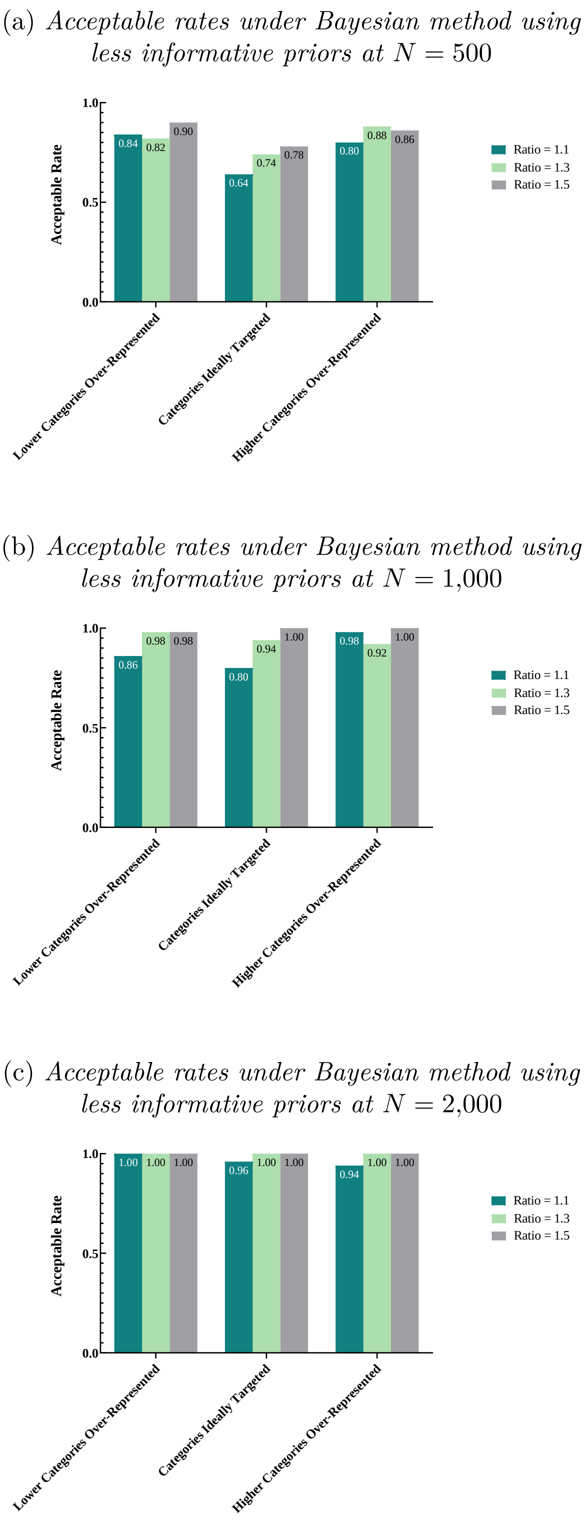
Figure 22

Figure 22. Boxplots summarizing the errors of the item discrimination estimates for $N = 2{,}000$ under full-information maximum likelihood (FIML), from Study 3. The plots in each row differ in Item 1's targetedness; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

further reveal how the targetedness of Item 1 affects the accuracy and stability of estimating parameters.

Parameter Recovery: Errors of the Parameter Estimates. For $N = 500$, the distribution of the errors related to the item discrimination estimates under LessInfo is summarized in Figure 24. When the ratio of Item 1's discriminations was 1.1, the discrepancies in the estimation accuracy between Item 1's discrimination parameters and that of the other items varied as Item 1's targetedness changed, as demonstrated in the plots in the first row of Figure 24). Specifically, when Item 1's lower categories had a greater representation in the data than its higher categories (shown in the first column of the figure), the range of the errors was 1.78 with an LUL of (−0.49, 1.29) for $\alpha_{1,1}$, whereas among the other items' discriminations on the primary dimension (i.e., α_{j1} where $j \neq 1$), the largest error range was 1.13 (LUL[−0.54, 0.59]) and the smallest was 0.73 (LUL[−0.36, 0.37]), which were for $\alpha_{5,1}$ and $\alpha_{3,1}$, respectively. Likewise, when Item 1's higher categories had a greater representation in the data than its lower categories (shown in the third column of the figure), the range of the errors for $\alpha_{1,1}$ was 1.91 with an LUL of (−0.51, 1.40), whereas the error ranges of the other items' discriminations on the primary dimension fell between 0.74 (LUL[−0.31, 0.42]) and 1.23 (LUL[−0.51, 0.72]). It is important to note that the range of the errors related to $\alpha_{1,1}$, as well as the largest and smallest error ranges across the other parameters on the primary dimension, were similar across the two targetedness conditions. However, at baseline (shown in the middle column of the figure), the range of the errors associated with $\alpha_{1,1}$ was larger, leading to greater discrepancy in the estimation accuracy for $\alpha_{1,1}$ and for the other discrimination parameters on the primary dimension. Particularly, when Item 1's categories were ideally targeted to the respondents, the error range for $\alpha_{1,1}$ was 2.82, with an LUL of (−0.59, 2.22), whereas the error ranges for the other items' discriminations on the primary dimension fell between 0.60 (LUL[−0.29, 0.31]) and 1.19 (LUL[−0.44, 0.75]).

When considering the discrimination parameters on the secondary dimension, similar



(a) *Acceptable rates under Bayesian method using less informative priors at $N = 500$*

(b) *Acceptable rates under Bayesian method using less informative priors at $N = 1{,}000$*

(c) *Acceptable rates under Bayesian method using less informative priors at $N = 2{,}000$*

Figure 23

Figure 23. Visualizations of the acceptable rates (ARs) under Bayesian method using less informative priors (LessInfo), from Study 3. The targetedness of Item 1 is represented along the horizontal axis; the proportion of runs in which the item discrimination estimates were acceptable are represented along the vertical axis. Ratios are represented in different colors.

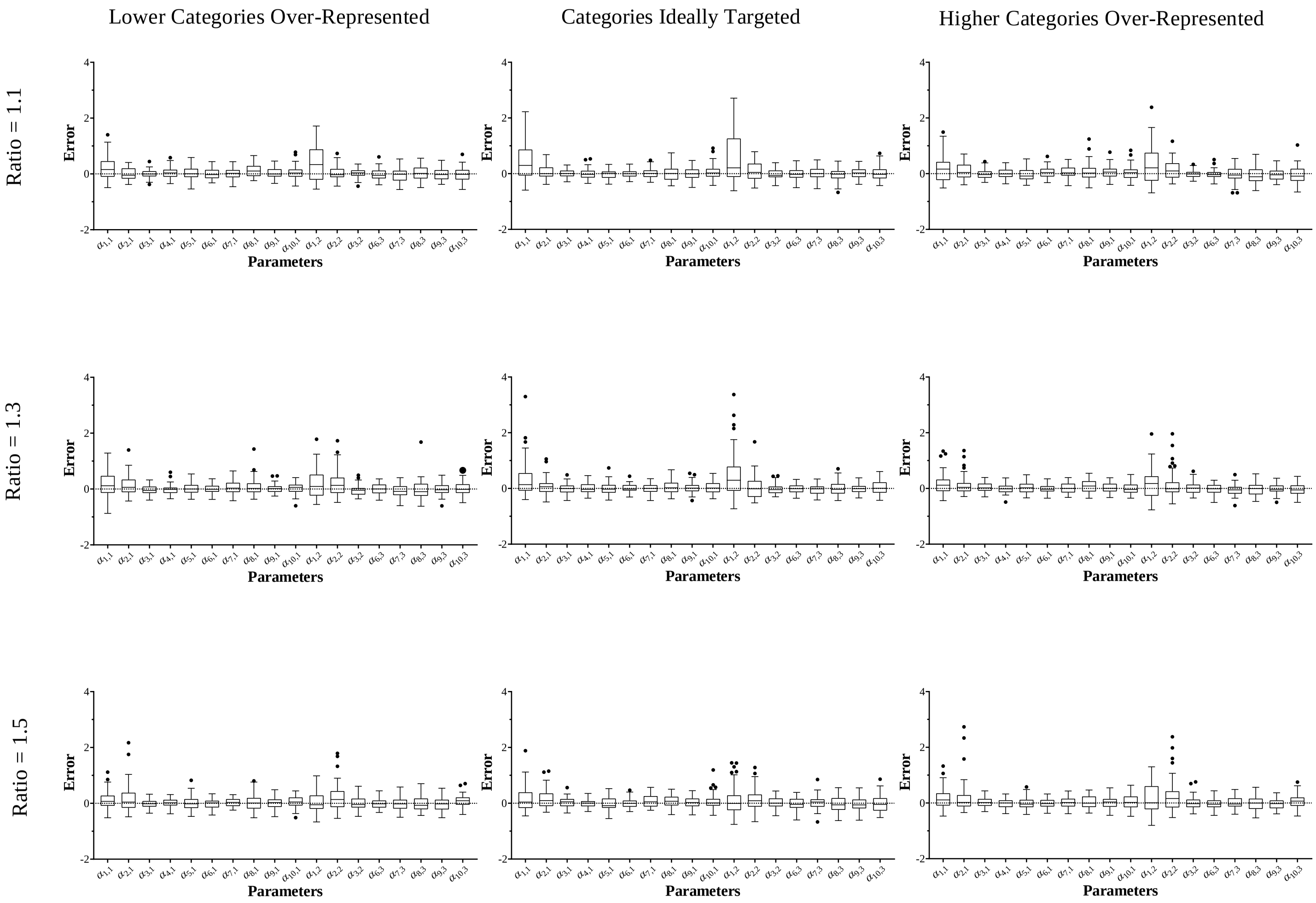


Figure 24

Figure 24. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under Bayesian method using less informative priors (LessInfo), from Study 3. The plots in each row differ in Item 1's targetedness; the plots in each column differ in similarity of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

patterns were observed. When Item 1's lower categories had a greater representation in the data than its higher categories (shown in the first column of the figure), the range of the errors for $\alpha_{1,2}$ was 2.26 (LUL[−0.55, 1.71], while the largest and smallest ranges across the other discrimination parameters on the secondary dimension were 1.21 (LUL[−0.56, 0.65]) for $\alpha_{10,3}$ and 0.70 (LUL[−0.35, 0.35]) for $\alpha_{3,2}$, respectively. When Item 1's higher categories had a greater representation in the data than its lower categories (shown in the third column of the figure), the range of the errors for $\alpha_{1,2}$ became 2.94 (LUL[−0.69, 2.25], while the largest and smallest ranges across the other discrimination parameters on the secondary dimension were 1.52 (LUL[−0.37, 1.15]) and 0.60 (LUL[−0.27, 0.32]), respectively. In contrast, under the ideal condition (shown in the middle column of the figure), the range of the errors increased to 3.32 (LUL[−0.61, 2.71]) for $\alpha_{1,2}$, with the largest and smallest ranges for the other discrimination parameters on the secondary dimension being 1.30 (LUL[−0.51, 0.79]) and 0.82 (LUL[−0.42, 0.39]), respectively. Overall, the results observed under ratio = 1.1 show that when Item 1 was off-targeted, the accuracy of Item 1's discrimination estimates improved, while the discrimination estimates for the other items remained relatively stable.

When the within-item discriminations for Item 1 became more differentiated, for example, a ratio of 1.3, similar trends were observed only for the secondary dimension (shown in the plots in the second row of Figure 24). Specifically, when Item 1's lower categories had a greater representation in the data than its higher categories, the range of the errors for $\alpha_{1,2}$ was 2.19 (LUL[−0.55, 1.64]), while the largest and smallest ranges for the other discrimination parameters related to the secondary dimensions were 1.71 (LUL[−0.48, 1.23]) for $\alpha_{2,2}$ and 0.73 (LUL[−0.35, 0.38]) for $\alpha_{3,2}$, respectively. Similarly, when Item 1's higher categories had a greater representation in the data than its lower categories, the range of the errors for $\alpha_{1,2}$ was 2.25 (LUL[−0.77, 1.48]), while the largest and smallest ranges for the other discrimination parameters related to the secondary dimensions were 1.30 (LUL[−0.55, 0.75]) and 0.79 (LUL[−0.43, 0.37]), respectively.

However, under the Ideal condition, the range of the errors related to $\alpha_{1,2}$ became 2.81 (LUL[-0.73, 2.08]), with the largest and smallest ranges for the other discrimination parameters related to the secondary dimensions being 1.65 (LUL[-0.52, 1.13]) and 0.68 (LUL[-0.35, 0.33]), respectively. It is critical to note that the range of the errors for $\alpha_{1,2}$ was larger under the Ideal condition than in the two problematic targetedness conditions, while the ranges of the errors for the other discrimination parameters related to the secondary dimensions remained relatively stable across all targetedness conditions.

As the ratio of Item 1's discriminations further increased to 1.5 (as shown in the plots in the third rows of Figure 24), the trends observed for ratio = 1.1 were no longer evident. Specifically, despite changes in Item 1's targetedness, the full ranges and LULs of the errors related to Item 1's parameters, as well as those related to the other items' parameters, remained stable.

The trends for $N = 500$ seem to indicate that under LessInfo, an item's targetedness might impact the estimation difficulties arising when the item's discriminations are similar in size. Specifically, an item being off-targeted may mitigate the estimation difficulties. However, the extent of the impact may depend on the degree of similarity of the item's discriminations, with the effect of the item's targetedness becoming more pronounced when the item's discriminations are more similar.

As the sample size increased to 1,000 (as depicted in Figure 25) and 2,000 (see Figure 26), similar patterns to those observed under $N = 500$ were found under each ratio condition, albeit with an overall enhancement in the estimation accuracy for all discrimination parameters, including those of Item 1. The findings suggest that under LessInfo, when an item's discriminations are relatively distinct, the item being off-targeted does not worsen the estimation difficulties. However, when an item's discriminations are highly similar, the item being off-targeted may lead to improved estimation accuracy for its discrimination parameters.

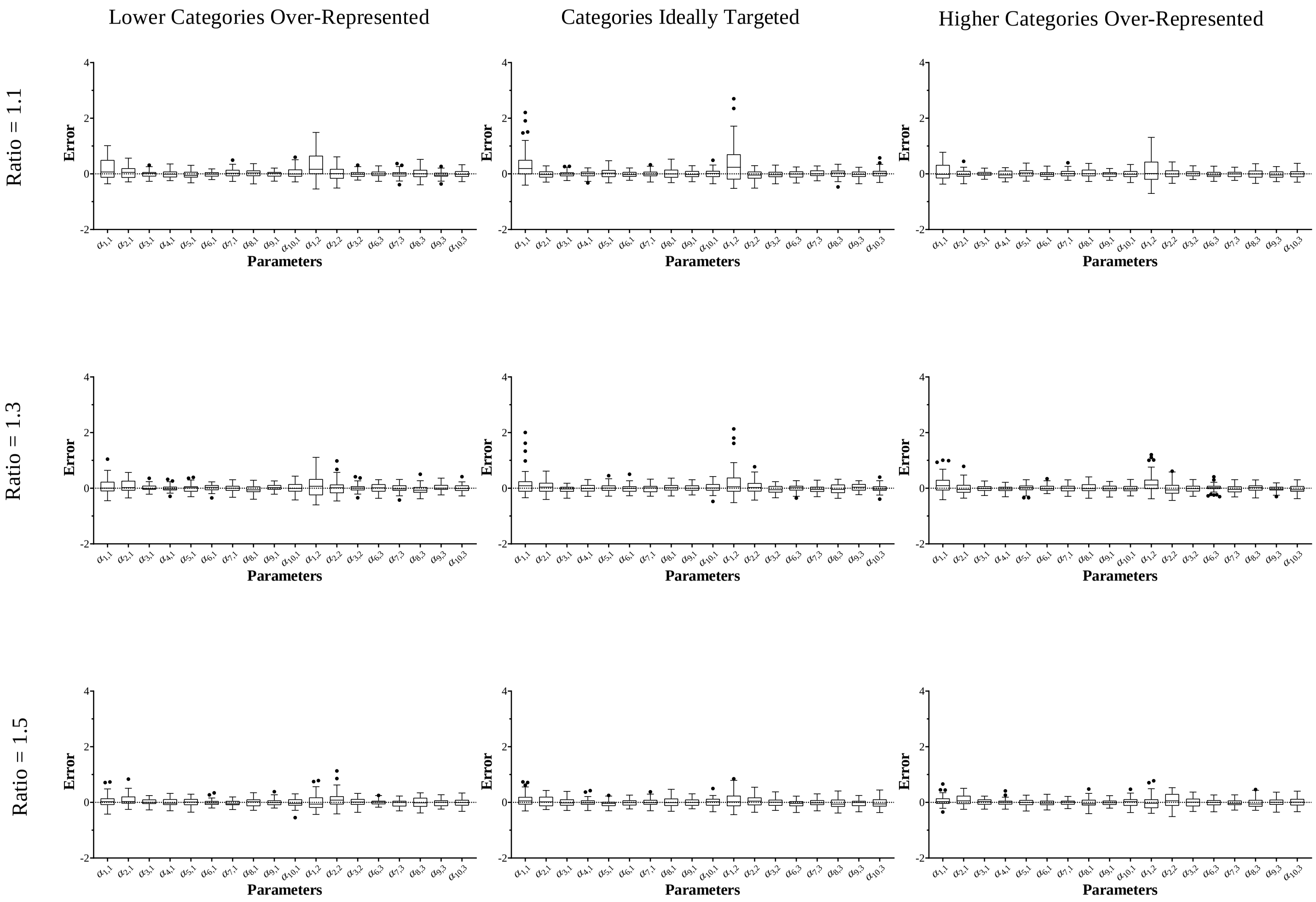


Figure 25

Figure 25. Boxplots summarizing the errors of the item discrimination estimates for $N = 1{,}000$ under Bayesian method using less informative priors (LessInfo), from Study 3. The plots in each row differ in Item 1's targetedness; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

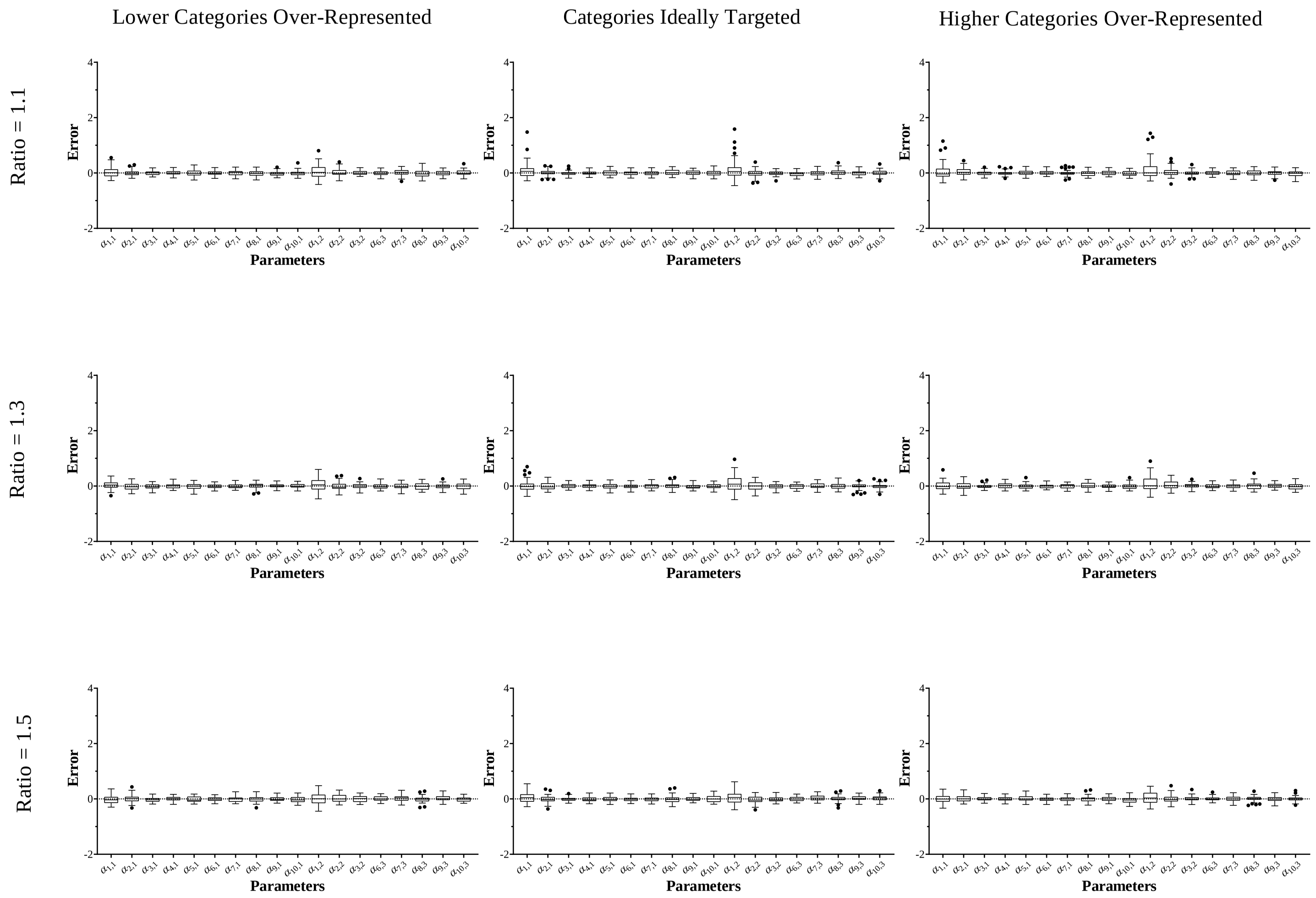


Figure 26

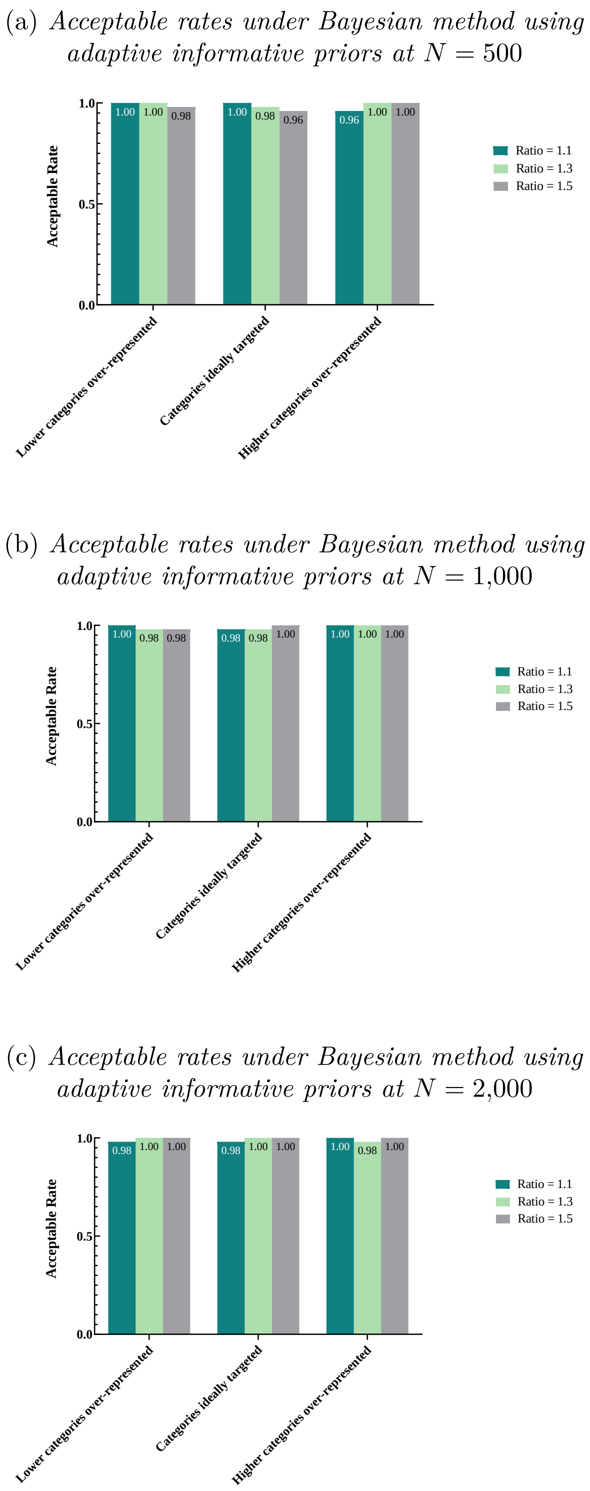
Figure 26. Boxplots summarizing the errors of the item discrimination estimates for $N = 2{,}000$ Bayesian method using less informative priors (LessInfo), from Study 3. The plots in each row differ in Item 1's targetedness; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

Bayesian Method Using Adaptive Informative Priors

Acceptable Rates. The ARs for the Bayesian method using adaptive informative priors (AdptInfo) are summarized in Figure 27. The patterns observed under AdptInfo were consistent to those seen under FIML in that the ARs did not vary evidently across the three targetedness conditions. Specifically, when $N = 500$ (as shown in Figure 27a), the average ARs across the ratios were almost identical under three item targetedness conditions, with it being at least .99. As the sample size increased to 1,000 and 2,000 (as depicted in Figures 27b and 27c), a similar trend was observed, indicating that, irrespective of the given ratio, the ARs remained largely unaffected by variations in item targetedness.

The findings regarding ARs under AdptInfo mirror those observed under FIML, indicating that item targetedness does not affect obtaining permissible results. The results of the parameter recovery reviewed next reveal how the targetedness of Item 1 affects the accuracy and stability of its estimation under AdptInfo.

Parameter Recovery: Errors of the Parameter Estimates. For $N = 500$, the distribution of the errors related to the item discrimination estimates under AdptInfo is summarized in Figure 28. For ratio = 1.1, the patterns regarding estimation accuracy observed between Item 1's discriminations and the other items' discriminations seemed consistent across the item targetedness conditions (as shown in the plots in the first row of Figure 28). Specifically, when Item 1's lower categories were more represented in the data than its higher categories, the range of the errors for $\alpha_{1,1}$ was 1.20 with an LUL of $(-0.59, 0.61)$, whereas among the other items' discriminations related to the primary dimension (i.e., α_{j1} where $j \neq 1$), the largest error range was 1.12 (LUL$[-0.52, 0.60]$) and the smallest was 0.71 (LUL$[-0.32, -.39]$), which were for $\alpha_{5,1}$ and $\alpha_{3,1}$, respectively. Similarly, when Item 1's higher categories were more represented than its lower categories, the range of the errors for $\alpha_{1,1}$ was 1.17 with an LUL of $(-0.52, 0.65)$, whereas the ranges of the errors for the other items' discriminations on the primary dimension fell between 0.74 (LUL$[-0.28, 0.46]$) and 1.17 (LUL$[-0.48, 0.69]$). Notice that the range of the errors related to $\alpha_{1,1}$, as


(a) *Acceptable rates under Bayesian method using adaptive informative priors at $N = 500$*

(b) *Acceptable rates under Bayesian method using adaptive informative priors at $N =$ 1,000*

(c) *Acceptable rates under Bayesian method using adaptive informative priors at $N =$ 2,000*

Figure 27

Figure 27. Visualizations of the acceptable rates (ARs) under Bayesian method using adaptive informative priors (AdptInfo), from Study 3. The targetedness of Item 1 is represented along the horizontal axis; the proportion of runs in which the item discrimination estimates were acceptable are represented along the vertical axis. Ratios are represented in different colors.

well as the largest and smallest error ranges across the other discriminations on the primary dimension, were highly similar between the two off-targetedness conditions. More importantly, comparable values were also observed at baseline (i.e., ideal condition), with a range of the errors for $\alpha_{1,1}$ being 1.11 (LUL[-0.64, 0.47]) and the largest and smallest error ranges for the other parameters related to the primary dimension being 0.62 (LUL[-0.26, 0.36]) and 1.13 (LUL[-0.42, 0.71]), respectively.

With respect to the discriminations related to the secondary dimensions, similar patterns were observed. Particularly, when Item 1's lower categories were more represented than its higher categories, the range of the errors for $\alpha_{1,2}$ was 1.30 (LUL[-0.60, 0.70], while the largest and smallest ranges for the other discrimination parameters related to the secondary dimensions were 1.10 (LUL[-0.53, 0.47]) for $\alpha_{10,3}$ and 0.65 (LUL[-0.29, 0.35]) for $\alpha_{3,2}$, respectively. When Item 1's higher categories were more represented than its lower categories, the range of the errors for $\alpha_{1,2}$ was 1.96 (LUL[-0.73, 1.22], while the largest and smallest ranges for the other discrimination parameters related to the secondary dimensions were 1.33 (LUL[-0.65, 0.67]) and 0.56 (LUL[-0.21, 0.35]), respectively. Similar to these two problematic targetedness conditions, when Item 1's categories were ideally targeted to the respondents, the range of the errors was 1.50 (LUL[-0.70, 0.80]) for $\alpha_{1,2}$, with the largest and smallest ranges for the other discrimination parameters related to the secondary dimensions being 1.03 (LUL[-0.43, 0.60]) and 0.71 (LUL[-0.31, 0.40]), respectively.

When the within-item discriminations for Item 1 became more distinct, as seen under ratios of 1.3 and 1.5 (shown in the plots in the second and third rows of Figure 28), the trends were comparable to those observed when the ratio = 1.1. Specifically, the full error ranges and LULs related to Item 1's parameters, as well as those related to other items' parameters, remained similar regardless of changes in Item 1's targetedness. These observations indicate that, under AdptInfo, the targetedness of an item might not worsen the estimation difficulties that arise when the item's discriminations are similar in size. In other words, the impact of within-item discriminations being similar on estimates is

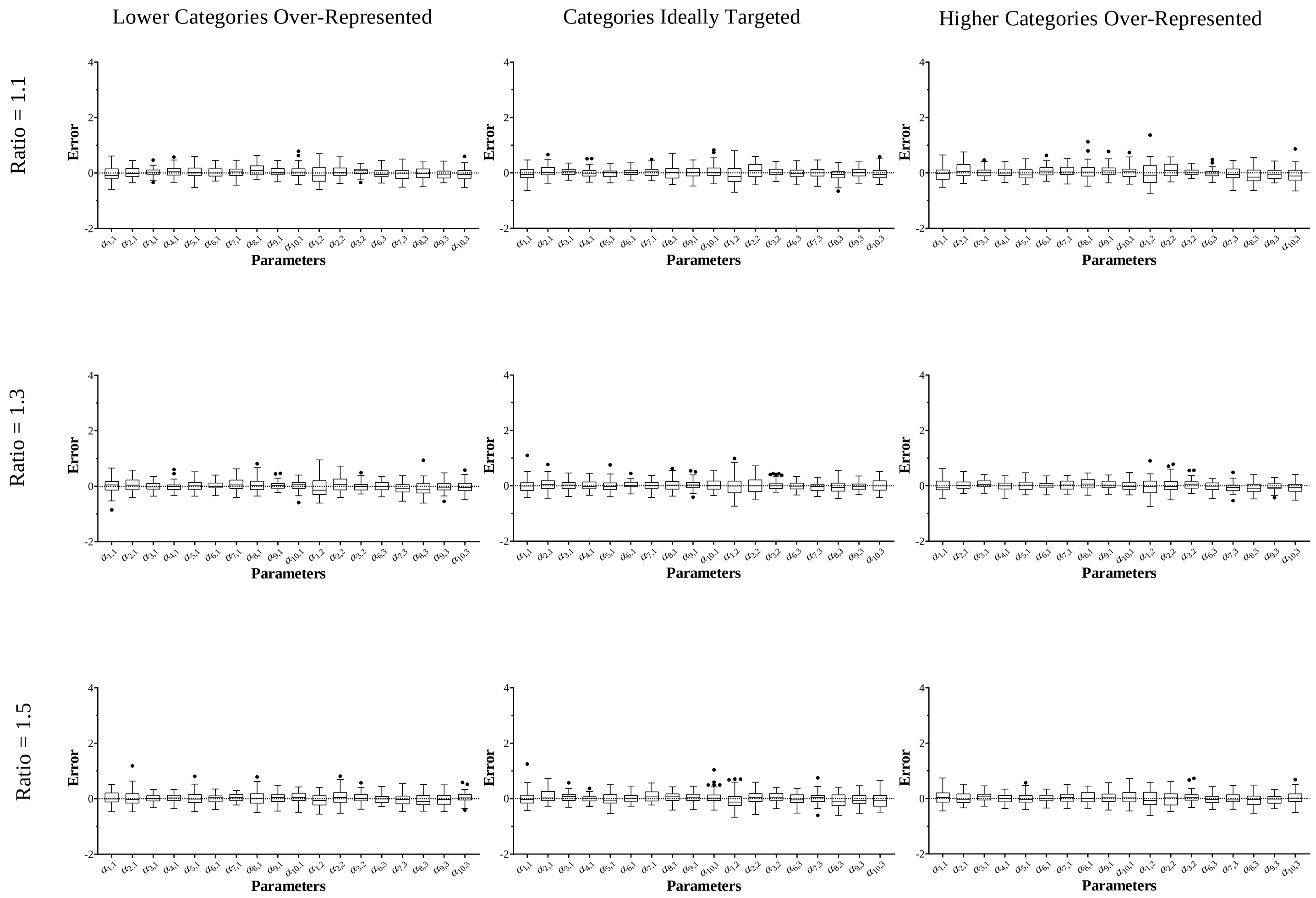


Figure 28

Figure 28. Boxplots summarizing the errors of the item discrimination estimates for $N = 500$ under Bayesian method using adaptive informative priors (AdptInfo), from Study 3. The plots in each row differ in Item 1's targetedness; the plots in each column differ in similarity of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

seemingly independent of item targetedness with AdptInfo.

As the sample size increased to 1,000 (as depicted in Figure 29) and 2,000 (see Figure 30), similar patterns to those observed under $N = 500$ were found under each ratio condition. However, there was an overall improvement in the accuracy of the estimates for all discrimination parameters, including those of Item 1. The findings suggest that, under AdptInfo, the item targetedness does not exacerbate the estimation challenges that arise when an item's discriminations are similar.

Comparison across the three estimation methods. The results reveal that, for FIML and AdptInfo, item targetedness does not influence the estimation challenges arising from within-item discriminations being similar in strength. Specifically, for a given sample size and ratio of within-item discriminations, the estimation of Item 1's discrimination parameters remained stable as item targetedness varied. In other words, the inconsistency concerning the estimation accuracy observed between Item 1 and the other items was not affected by item targetedness. However, it is important to note that the inconsistency in estimation accuracy between Item 1 and the remaining items was less pronounced under AdptInfo when compared with FIML.

In contrast, the results based on LessInfo reveal a slightly different pattern. Particularly, when an item's discriminations are relatively distinct, its targetedness does not play a role. In fact, when an item's discriminations are highly similar, the item being off-targeted could result in more accurate estimates for the discrimination parameters. Further investigation is needed to delve into this odd pattern.

Summary of Study 3

Study 3 demonstrated that an item's targetedness has minimal influence on the estimation difficulties that arise when the item's discriminations are similar in strength, especially when FIML or AdptInfo is used. Furthermore, in line with the previous two studies I have discussed, sample size is shown to influence the severity of the bias in the discrimination estimates. Specifically, for a given level of similarity of the within-item

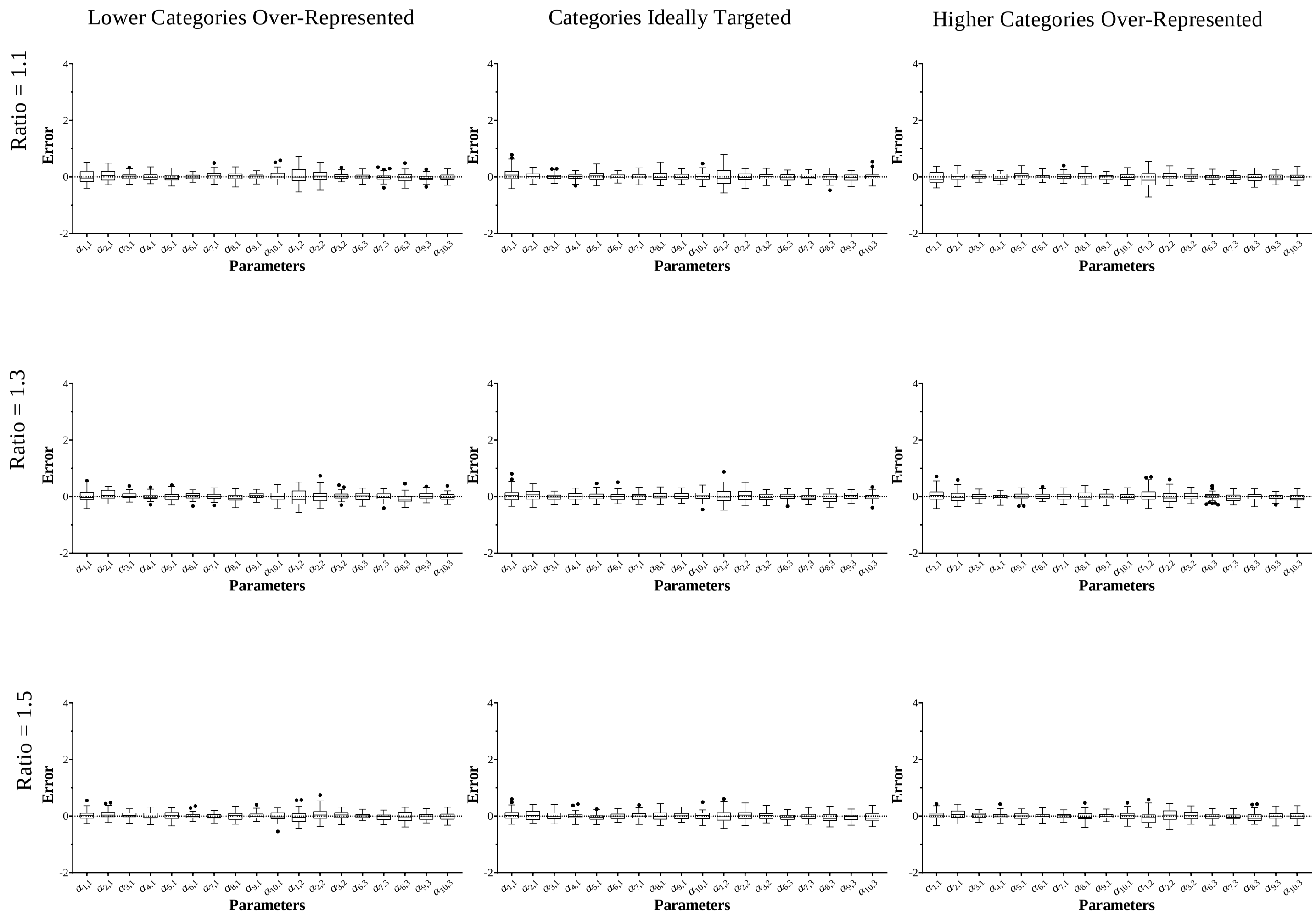


Figure 29

Figure 29. Boxplots summarizing the errors of the item discrimination estimates for $N = 1{,}000$ under Bayesian method using adaptive informative priors (AdptInfo), from Study 3. The plots in each row differ in Item 1's targetedness; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

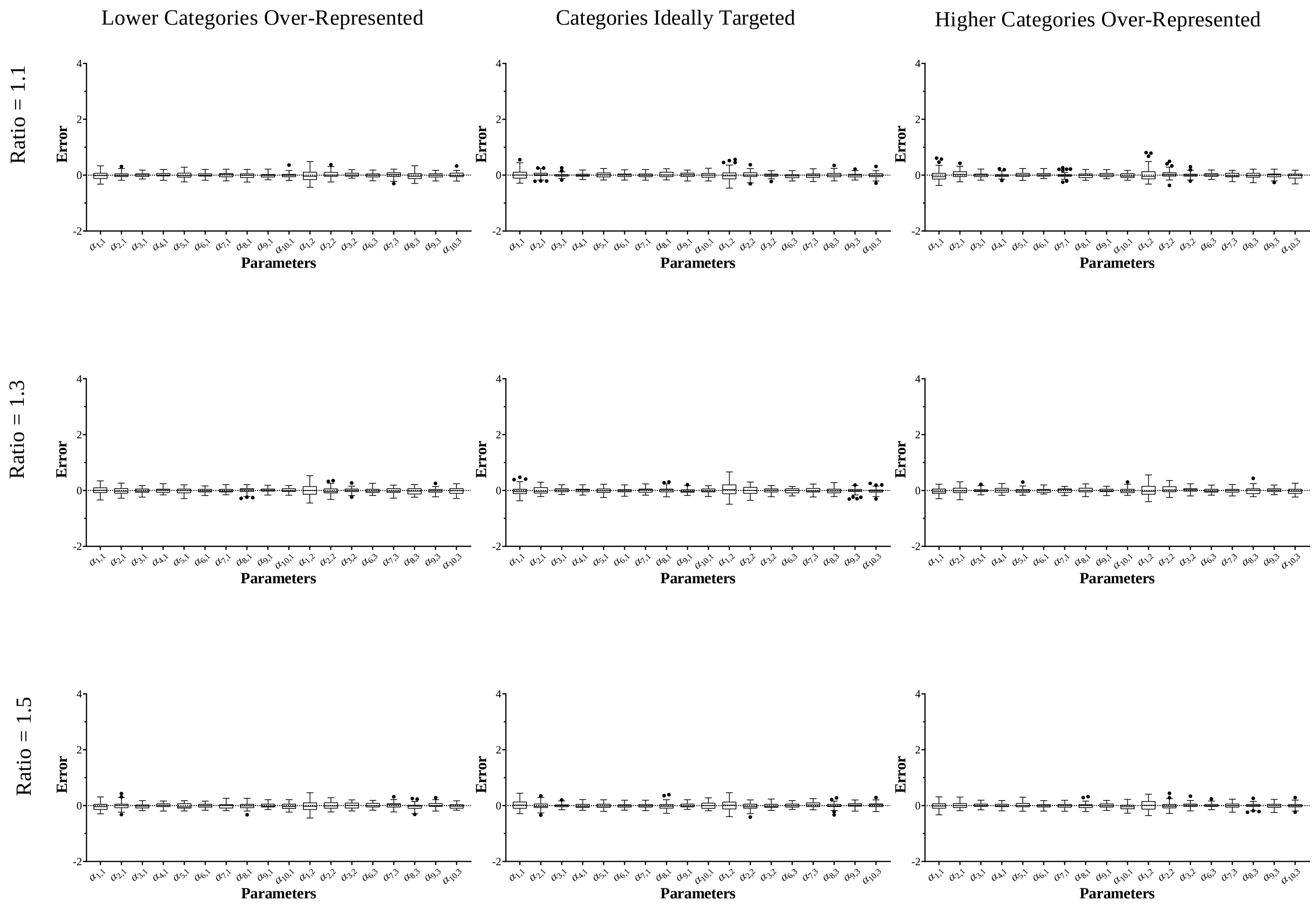


Figure 30

Figure 30. Boxplots summarizing the errors of the item discrimination estimates for $N = 2{,}000$ Bayesian method using adaptive informative priors (AdptInfo), from Study 3. The plots in each row differ in Item 1's targetedness; the plots in each column differ in ratio of Item 1's discriminations. In each plot, the horizontal axis represents the item discrimination parameters (where $\alpha_{j,d}$ is item j's discrimination on dimension d) and the vertical axis represents the errors.

discriminations and item targetedness, the bias decreases as the sample size increases. Study 3 further evaluated the performance of the different estimation methods for the bifactor model under more challenging situations. The results highlight that AdptInfo performs better than both FIML and LessInfo, as AdptInfo consistently yielded stable and more accurate estimates. In summary, the findings of Study 3 suggest that, when an item's discriminations are similar, the item's targetedness may not affect the estimation accuracy of the bifactor model in most scenarios.

CHAPTER 5

DISCUSSION

The bifactor item response theory (IRT) model is being used more frequently to determine the dimensionality of test data. Compared with some widely-used IRT models, such as the unidimensional model and the between-item-dimensionality model, the bifactor model allows researchers to assess the dependencies in the responses across all items, as well as the unique dependencies within the responses of subsets of items. This, in turn, enables researchers to better evaluate the degree to which a bifactor structure is represented in the data, thereby providing evidence of score validity.

To separate the different sources of dependencies influencing the responses, the bifactor model estimates two discrimination parameters for each item: one accounting for how much the general trait (or primary dimension) is represented in the item's responses, and the other for how much the extra (or secondary) dimension impacts the responses. Unfortunately, estimating two discrimination parameters per item can be challenging, as an empirical identification issue may arise during the estimation process. Specifically, the issue appears when an item's discriminations on the primary and secondary dimensions (or within-item discriminations) are similar in strength, leading to difficulties in estimating those discriminations. Currently, only limited evidence exists demonstrating that this issue occurs. How similar an item's discriminations need to be before the empirical identification issue arises is unknown. Also, whether the similarity of the within-item discriminations is moderated by factors like magnitude of the discriminations and item targetedness regarding when the empirical identification issue appears in bifactor modeling is unclear. Moreover, whether the empirical identification issue occurs similarly under different estimation methods has not been determined. By conducting three simulation studies in my dissertation, I provide insight about these questions.

Study 1 extended the investigation on the empirical identification issue demonstrated by Chen and Fujimoto (2022). This first study confirmed that an item's discriminations on the general and specific dimensions being similar in strength can lead to estimation challenges, and more importantly, it went beyond just this finding in that it provided additional evidence of this issue under varying sample sizes and degrees of within-item discrimination similarity, as well as explored the issue across different estimation methods. Study 1 revealed that the estimation method matters in the empirical identification issue of the bifactor model, a finding previously unexplored in literature, with Bayesian method using adaptive informative priors outperforming full-information maximum likelihood estimation and Bayesian method using less informative priors, particularly in situations where the sample size was small (e.g., $N \leq 1{,}000$) or when the item's discriminations were very similar (e.g., a ratio near 1).

Studies 2 and 3 also expanded on Chen and Fujimoto's (2022) work by showing the impact of magnitude of the discriminations and item targetedness on the empirical identification issue, respectively. Study 2 revealed an interaction effect between within-item discrimination similarity and magnitude of the discriminations. More specifically, when the magnitude of an item's discriminations was greater, the discriminations did not have to be as similar before the estimation problems appeared; when the magnitude of an item's discriminations was weaker, the discriminations had to be more similar before the estimation issue occurred. With regards to the performance of the different estimation methods, Bayesian method using adaptive informative priors, again, outperformed the other two methods consistently across all conditions. However, it is critical to note that this method underestimated the discrimination parameters for items having stronger discriminatory powers.

Study 3 further explored the impact of item targetedness on the within-item discrimination similarity before the empirical identification issue arises. The results did not show a clear pattern of the interaction between within-item discrimination similarity and item targetedness. Specifically, with full-information maximum likelihood estimation and Bayesian method using adaptive informative priors, item targetedness did not interact with within-item discrimination similarity, implying that item targetedness may not impact the similarity required for the estimation issue to emerge. However, under Bayesian method using less informative priors, when the item was off-targeted, its within-item discriminations had to be more similar to observe the empirical identification issue.

The three studies together, then, demonstrate that the empirical identification issue of the bifactor model due to the item's discriminations being similar is moderated by magnitude of the within-item discriminations. Additionally, these studies highlight the empirical nature of this identification issue that I focused on, showing that larger sample sizes can mitigate the estimation inaccuracies caused by within-item discriminations being similar and the discriminations being strong in magnitude. The reason is, as sample size increases, there is more information in the data, leading to these factors creating less of an estimation problem. The findings regarding the impact of item targetedness on the empirical identification issue were inconclusive, which was unexpected because Xia and Yang (2018) suggested that more extreme threshold values may lead to inaccurate parameter estimation. However, their focus was on all items being off targeted, whereas I focused on only one item, resulting in the overall targetedness of the items as a set not being sufficiently off-targeted. This might explain the lack of the interaction effect between item targetedness and similarity of the within-item discriminations on the estimation issue.

The studies I conducted also collectively demonstrated that the estimation method matters in the empirical identification issue of the bifactor model that I focused on. If we were to view FIML as a Bayesian model with noninformative priors on the item parameters (ignoring the difference between marginal and conditional likelihoods), then the only difference among the estimation methods explored would be the priors assigned to the item parameters. Thus, any observed differences in the parameter estimation across the three methods can be directly attributed to the differences in the priors assigned to the item parameters. In the conditions I examined, Bayesian estimation using adaptive informative priors was consistently observed to produce more accurate discrimination estimates than the other two estimation methods when the empirical identification issue occurred. However, as I noted earlier, this method underestimated an item's discrimination parameters when the item's discriminatory powers were strong in magnitude. This underestimation occurred because the hyperpriors applied in this method shifted the values likely to be supported by the priors to where the majority of the discrimination parameters were. However, underestimation may be desirable compared with overestimation. For example, in high-stakes testing, conservative estimates could be beneficial, as they can reduce the risk of overestimating the reliability of a test. Concerning Bayesian method using less informative priors, it generally underperformed when compared with the other two methods, which was unexpected considering the presence of some degree of information provided by the priors.

Regarding the most commonly used method in IRT modeling—full-information maximum likelihood estimation—this dissertation shows that it may not be an optimal estimation method because it may produce noticeably biased parameter estimates when an item's discriminations are similar in strength, large in magnitude, or both. The reason this method fails when the empirical identification issue occurs is that it only requires assigning a prior distribution to the latent trait dimensional positions, and no other information other than the response data is needed to estimate the other parameters. This could create problems in the estimation process when the within-item discriminations are similar in strength, as the information in the data alone may not be sufficient to differentiate the within-item discriminations.

Limitations and Future Directions

Although I provide insight into the factors that contribute to the estimation issue arising from within-item discrimination being similar, several limitations should be acknowledged. Firstly, I only examined a single bifactor structure composed of 10 items, a structure motivated by RSES data. I focused on one structure so that I could perform a more in-depth investigation (e.g., exploring factors that interact with the within-item discrimination similarity) than if I were to explore a range of bifactor structures. However, focusing on one structure may limit the generalizability of my findings. Nonetheless, it is reasonable to anticipate that the effect of the factors that I have shown to impact the empirical identification issue will persist—albeit to varying degrees—with other bifactor structures, as Stone and Zhu (2015) suggested that this empirical identification is a general issue. Future research, however, should confirm whether the number of items impacts the empirical identification issue because it is possible that including more items in a bifactor structure might mitigate the estimation challenges associated with item having similar within-item discriminations, as more items represent more information in the data.

Another limitation of this study is that I only manipulated a single item in order to establish clear patterns of how the factors of interest impact the empirical identification issue with the bifactor model. Future investigations might examine the empirical identification issue with more than one item presenting extreme parameter values and explore whether more than one such item could influence the estimation of nonproblematic items. A third limitation of this dissertation is that, within Bayesian estimation, I only examined two priors for the item discriminations. Of these two priors, the more informative one produced more accurate discrimination estimates when the empirical identification issue occurred, but it underestimated the parameters, particularly when the magnitude of an item's discriminations was strong. The specific impacts of the inaccuracies associated with these priors, however, remain unclear. Future studies should investigate how inferences are impacted by these inaccuracies. If a substantive impact is identified, then more accurate ways to estimating the bifactor model will be needed. For instance, a way that could potentially produce more accurate parameter estimates is a two-step approach, with the first step involving estimating the bifactor model using informative priors, followed by adjusting the prior for items with noticeably larger discrimination estimates.

A final limitation of this dissertation is that I only focused on the impact of the empirical identification issue on the item discrimination estimates. The broader impacts of this issue are unexplored. Future research should examine how the empirical identification issue affects the validity and reliability (e.g., categorical ω) of tests. In addition, given the equivalent transformation of factor loadings and IRT discrimination parameters, it is logical to anticipate that the empirical identification issue could also influence structural inferences within the context of structural equation modeling, which needs to be verified by further investigations.

Conclusions

Bifactor IRT modeling has grown considerably in fields like education and psychology, as it can verify a structure that is suitable to many theories. However, the practical application of the bifactor model presents certain challenges, one of which is an empirical identification issue that was the focus of this dissertation. This issue occurs when an item's discrimination parameters on the primary and secondary dimensions are similar in strength, resulting in potentially biased estimates for these parameters. Currently, only one study has demonstrated in a simple scenario that this problem exists. Whether other typical situations seen in practice like strength of the discriminations and item targetedness interact with the empirical identification issue and whether the estimation method matters have not been explored.

My work on this dissertation broadens our understanding about this empirical identification issue with the bifactor model. I provide new insight about the factors that influence this issue and offer guidance on the choice of estimation methods to mitigate the issue. My findings indicate that magnitude of the item discriminations impacts how different the within-item discriminations may have to be before the estimates are not biased. Furthermore, an adaptive informative prior within a Bayesian setting may be better to use than less informative priors within the same setting or FIML within the frequentist framework. However, I also reveal new concerns such as the underestimation, thereby highlighting the need for further exploration of alternative estimation methods and strategies for handling the empirical identification issue.

REFERENCES

Albert, J. H. (1992). Bayesian estimation of normal ogive item response curves using Gibbs sampling. *Journal of educational statistics*, *17*(3), 251–269. https://doi.org/10.3102/10769986017003251

Alessandri, G., Vecchione, M., Eisenberg, N., & Łaguna, M. (2015). On the factor structure of the Rosenberg (1965) General Self-Esteem Scale. *Psychological Assessment*, *27*(2), 621. https://doi.org/10.1037/pas0000073

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*.

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*(2), 171–187. https://doi.org/10.1093/pan/mpi010

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied psychological measurement*, *6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Bolt, D. M. (2019). Bifactor mirt as an appealing and related alternative to cdms in the presence of skill attribute continuity. In *Handbook of diagnostic classification models* (pp. 395–417). Springer. https://doi.org/10.1007/978-3-030-05584-4_19

Bornovalova, M. A., Choate, A. M., Fatimah, H., Petersen, K. J., & Wiernik, B. M. (2020). Appropriate use of bifactor analysis in psychopathology research: Appreciating benefits and limitations. *Biological Psychiatry*, *88*(1), 18–27. https://doi.org/10.1016/j.biopsych.2020.01.013

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 434–455. https://doi.org/10.2307/1390675

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Sage publications.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., et al. (2014). The p factor: One general psychopathology factor in the structure of psychiatric disorders? *Clinical psychological science*, *2*(2), 119–137. https://doi.org/10.1177/2167702613497473

Chen, W., & Fujimoto, K. A. (2022). An empirical identification issue of the bifactor item response theory model. *Applied Psychological Measurement*. https://doi.org/10.1177/01466216221108133

Coulacoglou, C., & Saklofske, D. H. (2017). *Psychometrics and psychological assessment: Principles and applications*. Academic Press.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. ERIC.

De Ayala, R. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*(2), 155–170. https://doi.org/10.1177/014662169401800205

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of educational measurement*, *43*(2), 145–168. https://doi.org/10.1111/j.1745-3984.2006.00010.x

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement*, *36*(2), 104–121. https://doi.org/10.1177/0146621612437403

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing*, *13*(4), 354–378. https://doi.org/10.1080/15305058.2013.799067

Donnellan, M. B., Ackerman, R. A., & Brecheen, C. (2016). Extending structural analyses of the Rosenberg Self-Esteem Scale to consider criterion-related validity: Can composite self-esteem scores be good enough? *Journal of Personality Assessment*, *98*(2), 169–177. https://doi.org/10.1080/00223891.2015.1058268

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*(1), 77–90. https://doi.org/10.1177/014662168901300108

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of irt graded response models: Limited versus full information methods. *Psychological methods*, *14*(3), 275. https://doi.org/10.1037/a0015825

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer.

Fujimoto, K. A. (2019). A more flexible Bayesian multilevel bifactor item response theory model. *Journal of Educational Measurement*. https://doi.org/10.1111/jedm.12249

Fujimoto, K. A., & Neugebauer, S. R. (2020). A general bayesian multidimensional item response theory model for small and large samples. *Educational and Psychological Measurement*, 1–30. https://doi.org/10.1177/0013164419891205

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., Bhaumik, D. K., Kupfer, D. J., Frank, E., Grochocinski, V. J., & Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*(1), 4–19. https://doi.org/10.1177/0146621606289485

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

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Zeitschrift für Psychologie/Journal of Psychology*, *216*(2), 89–101. https://doi.org/10.1027/0044-3409.216.2.89

Hendy, N. T., & Biderman, M. D. (2019). Using bifactor model of personality to predict academic performance and dishonesty. *The International Journal of Management Education*, *17*(2), 294–303. https://doi.org/10.1016/j.ijme.2019.05.003

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Hopko, D. R. (2003). Confirmatory factor analysis of the math anxiety rating scale–revised. *Educational and Psychological Measurement*, *63*(2), 336–351. https://doi.org/10.1177/0013164402251041

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in psychology*, *7*, 109. https://doi.org/10.3389/fpsyg.2016.00109

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. https://doi.org/10.1111/j.1745-3984.2011.00161.x

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo estimation. *Applied Psychological Measurement*, *36*(5), 399–419. https://doi.org/10.1177/0146621612446170

Kose, I. A., & Demirtasli, N. C. (2012). Comparison of unidimensional and multidimensional models based on item response theory in terms of both variables of test length and sample size. *Procedia-Social and Behavioral Sciences*, *46*, 135–140. https://doi.org/10.1016/j.sbspro.2012.05.082

Lee, W.-y., & Cho, S.-J. (2017). Detecting differential item discrimination (did) and the consequences of ignoring did in multilevel item response models. *Journal of Educational Measurement*, *54*(3), 364–393. https://doi.org/10.1111/jedm.12148

Linacre, J. M., et al. (2002). Optimizing rating scale category effectiveness. *Journal of applied measurement*, *3*(1), 85–106.

Liu, Y., & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. *Applied Psychological Measurement*, *36*(8), 670–688. https://doi.org/10.1177/0146621612458174

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). The bugs book. *A Practical Introduction to Bayesian Analysis, Chapman Hall, London.*

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

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the rosenberg self-esteem scale: Traits, ephemeral artifacts, and stable response styles. *Psychological assessment*, *22*(2), 366. https://doi.org/10.1037/a0019225

Maydeu-Olivares, A., & McArdle, J. J. (2005). *Contemporary psychometrics*. Psychology Press.

Michaelides, M. P., Koutsogiorgi, C., & Panayiotou, G. (2016). Method effects on an adaptation of the Rosenberg Self-Esteem Scale in Greek and the role of personality traits. *Journal of Personality Assessment*, *98*(2), 178–188. https://doi.org/10.1080/00223891.2015.1089248

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). The development of the general factor of psychopathology 'p factor'through childhood and adolescence. *Journal of abnormal child psychology*, *44*(8), 1573–1586. https://doi.org/10.1007/s10802-016-0132-1

Muthén, L. K., & Muthen, B. O. (2017). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Muthén, L., & Muthén, B. (2016). Mplus. *The comprehensive modelling program for applied researchers: user's guide*, *5*.

Paek, I., Cui, M., Öztürk Gübeş, N., & Yang, Y. (2018). Estimation of an IRT model by Mplus for dichotomously scored responses under different estimation methods.

Educational and Psychological Measurement, *78*(4), 569–588. https://doi.org/10.1177/0013164417715738

Rauch, D. P., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, *52*(4), 354.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79–112). Springer.

Reise, S. P., Cook, K. F., & Moore, T. M. (2014). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In *Handbook of item response theory modeling* (pp. 31–58). Routledge.

Reise, S. P., & Haviland, M. G. (2005). Item response theory and the measurement of clinical change. *Journal of personality assessment*, *84*(3), 228–238. https://doi.org/10.1207/s15327752jpa8403_02

Reise, S. P., Kim, D. S., Mansolf, M., & Widaman, K. F. (2016). Is the bifactor model a better model or is it just better at modeling implausible responses? application of iteratively reweighted least squares to the Rosenberg Self-Esteem Scale. *Multivariate Behavioral Research*, *51*(6), 818–838. https://doi.org/10.1080/00273171.2016.1243461

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, *92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477

Reiser, M., & VandenBerg, M. (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology*, *47*(1), 85–107. https://doi.org/10.1111/j.2044-8317.1994.tb01026.x

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological methods*, *21*(2), 137. https://doi.org/10.1037/met0000045

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton,NJ: Princeton Pniversity Press.

Salerno, L., Ingoglia, S., & Coco, G. L. (2017). Competing factor structures of the Rosenberg Self-Esteem Scale (RSES) and its measurement invariance across clinical and non-clinical samples. *Personality and Individual Differences*, *113*, 13–19. https://doi.org/10.1016/j.paid.2017.02.063

Samejima, F. (1997). Graded response model. In *Handbook of modern item response theory* (pp. 85–100). Springer.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of cronbach's alpha. *Psychometrika*, *74*, 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sorensen, T., & Vasishth, S. (2015). Bayesian linear mixed models using stan: A tutorial for psychologists, linguists, and cognitive scientists. *arXiv preprint arXiv:1506.06201*. https://doi.org/10.48550/arXiv.1506.06201

Stone, C. A., & Zhu, X. (2015). *Bayesian analysis of item response theory models using SAS*. Sas Institute.

Team, S. D., et al. (2016). Rstan: The r interface to stan. *R package version*, *2*(1), 522.

Toland, M. D., Sulis, I., Giambona, F., Porcu, M., & Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of school psychology*, *60*, 41–63. https://doi.org/10.1016/j.jsp.2016.11.001

Wang, W.-C., Chen, H.-F., & Jin, K.-Y. (2015). Item response theory models for wording effects in mixed-format scales. *Educational and Psychological Measurement*, *75*(1), 157–178. https://doi.org/10.1177/0013164414528209

Xia, Y., & Yang, Y. (2018). The influence of number of categories and threshold values on fit indices in structural equation modeling with ordered categorical data.

Multivariate Behavioral Research, *53*(5), 731–755. https://doi.org/10.1080/00273171.2018.1480346

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*(2), 125–145. https://doi.org/10.1177/014662168400800201

Yeo, Z. Z., & Suárez, L. (2022). Validation of the mental health continuum-short form: The bifactor model of emotional, social, and psychological well-being. *Plos one*, *17*(5), e0268232. https://doi.org/10.1371/journal.pone.0268232

VITA

Wenya Chen earned a Bachelor of Arts in English from Hubei University of Education in Wuhan, China, in 2012. Continuing her education at Indiana University Bloomington, she obtained a Master of Science in International and Comparative Education with a minor in Quantitative Research Methods in 2014. Wenya attended Loyola University Chicago for her doctoral training, where she studied Research Methodology and received her Ph.D. in 2024.

During her time at Loyola, Wenya served as a Graduate Research Assistant and Teaching Assistant, and performed independent research. Her publication has appeared in *Applied Psychological Measurement*, entitled “An Empirical Identification Issue of the Bifactor IRT Model”. Also, her research findings were presented at various national and international conferences, including the National Council on Measurement in Education, the Psychometric Society, and the International Studies Association.

Currently, Wenya is a Senior Statistician at Stanley Manne Children’s Research Institute at Lurie Children’s Hospital of Chicago.