EVALUATION OF STRUCTURE COMPLEXITY MAGNITUDE, DEGREE OF CROSS-LOADING ON SECONDARY DIMENSION AND MODEL SPECIFICATION ON MIRT PARAMETER ESTIMATION

By

MOSTAFA HOSSEINZADEH

Bachelor of Science in Biosystems Engineering
Shahid Bahonar University of Kerman
Kerman, Iran
2007

Master of Science in Industrial and System Engineering
Islamic Azad University
Arak, Iran
2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2021

EVALUATION OF STRUCTURE COMPLEXITY MAGNITUDE, DEGREE OF CROSS-LOADING ON SECONDARY DIMENSION AND MODEL SPECIFICATION ON MIRT PARAMETER ESTIMATION

Dissertation Approved:

Ki Cole, Ph.D.

Dissertation Adviser

Jam Khojasteh, Ph.D.

Mwarumba Mwavita, Ph.D.

Isaac Washburn, Ph.D.

ACKNOWLEDGEMENTS

To the memory of my loving father, always in my mind, forever in my heart

To my beloved mother

"They say there is a doorway from heart to heart, but what is the use of a door when there are no walls?" (Molana Rumi)

First and foremost, I would like to praise and thank God, the almighty. I express my deepest gratitude to my dissertation adviser, Dr. Ki Cole, for her valuable insights, tremendous guidance, encouragement and support throughout this Journey. I would like to extend my sincere appreciation to Dr. Jam Khojasteh, Dr. Mwarumba Mwavita and Dr. Isaac Washburn for their valuable and constructive insights that helped me to improve the quality of this work. I express my heartfelt gratitude to my family and all of my teachers for their unconditional and endless love and support. Special thanks to all of my friends and colleagues.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: MOSTAFA HOSSEINZADEH

Date of Degree: DECEMBER, 2021

Title of Study: EVALUATION OF STRUCTURE COMPLEXITY MAGNITUDE, DEGREE OF CROSS-LOADING ON SECONDARY DIMENSION AND MODEL SPECIFICATION ON MIRT PARAMETER ESTIMATION

Major Field: EDUCATIONAL PSYCHOLOGY

Abstract: In real-world situations, multidimensional data may appear on large-scale tests or attitudinal surveys. A simple structure, multidimensional model may be used to evaluate the items, ignoring the cross-loading of some items on the secondary dimension. The purpose of this study was to investigate the influence of structure complexity magnitude of the data incorporating the degree of cross-loading on secondary dimensions and model specification, especially when the model was misspecified as a simple structure, ignoring the cross-loading, while the data are truly complex on item parameter recovery in MIRT models. In order to address the research question a simulation study that replicated this scenario was designed in order to manipulate the variables that could potentially influence the precision of item parameter estimation in the MIRT models. Item parameters were estimated using marginal maximum likelihood (MML), utilizing the expectation-maximization (EM) algorithms. A compensatory 2PL-MIRT model with two dimensions and dichotomous item response type (Reckase, 1985) was used to simulate and calibrate the data for each combination of conditions across 500 replications. The result of this study indicated that ignoring complex structure of the multidimensional data incorporating the degree of cross-loading and model specification severely impact item discrimination estimations resulting biased and inaccurate item discrimination parameters. When the complex structure of the data was misspecified, whether the data were correlated or uncorrelated, item discrimination parameters were adversely affected. As the complexity magnitude incorporating the degree of cross-loading increased, the error and bias estimates of item discrimination worsened. Furthermore, the results of this study indicated that if the data are correlated and the correlation is not specified nor are the item cross-loadings, item discrimination estimates, specifically for the truly cross-loaded items had extremely poor error and bias estimates.

TABLE OF CONTENTS

Chapter Page

I. INTRODUCTION..1

Item Response Theory (IRT)...1

Unidimensional IRT Models (UIRT) 2

Multidimensional IRT Models (MIRT) 2

Compensatory and Non-compensatory MIRT Models................................3

Structure Complexity of MIRT Models..4

Statement of the Problem 6

Purpose of the Study 8

Research Question 8

II. REVIEW OF LITERATURE 9

Item Response Data 9

Multidimensional Item Response Theory (MIRT) Models 9

Assumption of MIRT Models 9

Dichotomous and Polytomous MIRT Models...11

Compensatory and Non-compensatory MIRT Models..............................11

Simple Structure MIRT Model..13

Complex Structure MIRT Model 14

Variables with Potential Influence on Item Parameter Estimation in MIRT Models..16

Conditions of the Current Study...20

Sample Size (Number of Examinees) 21

Correlation between Latent Abilities...21

Structure Complexity Magnitude 22

Degree of Cross-Loading on Secondary Dimension 23

Model Specification...24

Chapter Page

Summary 24

III. METHODOLOGY 26

Simulation Study Design 27

Manipulated Variables 27

Item Parameter Specifications 30

Analysis 32

Instrument 33

Evaluation Criteria 33

IV. RESULTS 35

Section I: Item Parameter Recovery Results in Terms of Average RMSE 37

Correct Specified Models 37

Misspecified Models 38

Effect of Sample Size on Item Parameter Recovery in Terms of Average RMSE 45

Correct Specified Models 45

Misspecified Models 45

Effect of Correlation on Item Parameter Recovery in Terms of RMSE 46

Correct Specified Models 46

Misspecified Models 48

Effect of Structure Complexity Magnitude on Item Parameter Recovery in Terms of RMSE 49

Correct Specified Models 49

Misspecified Models 50

Effect of Degree of Cross-Loading on Item Parameter Recovery in Terms of RMSE 51

Correct Specified Models 51

Misspecified Models 52

Section II: Item Parameter Recovery Results in Terms of Average Bias 61

Correct Specified Models 61

Misspecified Models 62

Effect of Sample Size on Item Parameter Recovery in Terms of Average Bias 68

Chapter Page

Correct Specified Models..68

Misspecified Models..68

Effect of Correlation between Abilities on Item Parameter Recovery in Terms of Average Bias 69

Correct Specified Models..69

Misspecified Models..70

Effect of Structure Complexity Magnitude on Item Parameter Recovery in Terms of Average Bias 71

Correct Specified Models..71

Misspecified Models..72

Effect of Degree of Cross-Loading on Item Parameter Recovery in Terms of Average Bias 73

Correct Specified Models..73

Misspecified Models..74

Section III: Item Parameter Recovery Results in Terms of Average Standard Error of Estimate (SEE) 84

Correct Specified Models..84

Misspecified Models..85

Effect of Sample Size on Item Parameter Recovery in Terms of SEE......................91

Correct Specified Models..91

Misspecified Models..91

Effect of Correlation between Abilities on Item Parameter Recovery in Terms of SEE 92

Correct Specified Models..92

Misspecified Models..93

Effect of Structure Complexity Magnitude on Item Parameter Recovery in Terms of SEE 93

Correct Specified Models..93

Misspecified Models..94

Effect of the Degree of Cross-Loading on Item Parameter Recovery in Terms of SEE 95

Chapter Page

Correct Specified Models 95

Misspecified Models 96

V. DISCUSSION 105

Summary of Results 107

Effect of Structure Complexity Magnitude (RMSE, Bias and SEE) 107

Effect of the Degree of Cross-Loading (RMSE, Bias and SEE) 108

Effect of Sample Size (RMSE, Bias and SEE) 109

Effect of Correlation (RMSE, Bias and SEE) 110

Item Location 110

Summary of Results Compared to the Literature 111

Implications and Suggestions 111

Limitations and Future Studies 114

REFERENCES 116

LIST OF TABLES

Table Page

3. 1. Item Parameters for 2D Compensatory MIRT Model for 10 Items per Dimension for Three Types of Complexity Structures Including Low Degree of Cross-Loading on Secondary Discrimination. 31

4. 1. Average RMSE of the primary (cross-loaded and non-cross loaded items) and secondary discrimination on the first and second dimensions when the models were correctly specified. 41

4. 2. Average RMSE of the truly primary (cross-loaded and non-cross loaded items) on the first and second dimensions when the models were misspecified. 42

4. 3. Average RMSEs of the item location parameter (d) for the correct and misspecified models. 43

4. 4. Average bias of the primary (cross-loaded and non-cross loaded items) and secondary discrimination on the first and second dimensions when the models were correctly specified. 64

4. 5. Average bias of the truly primary (cross-loaded and non-cross loaded items) on the first and second dimensions when the models were misspecified. 65

4. 6. Average bias of the item location parameter (d) for the correct and misspecified models. 66

4. 7. Average SEE of the primary (cross-loaded and non-cross loaded items) and secondary discrimination on the first and second dimensions when the models were correctly specified. 87

4. 8. Average SEE of the truly primary (cross-loaded and non-cross loaded items) on the first and second dimensions when the models were misspecified. 88

4. 9. Average SEE of the item location parameter (d) for the correct and misspecified models. 89

LIST OF FIGURES

Figure Page

2. 1. A simple structure MIRT model with two underlying abilities ($\theta 1$ and $\theta 2$) and correlation ρ. 14

2. 2. A complex structure MIRT model with two underlying abilities ($\theta 1$ and $\theta 2$) and correlation ρ. 15

4. 1. Average RMSEs for item discrimination parameters when models were specified correctly with a low degree of cross-loading 53

4. 2. Average RMSEs for item discrimination parameters when models were specified correctly with a medium degree of cross-loading 54

4. 3. Average RMSEs for item discrimination parameters when models were specified correctly with a high degree of cross-loading 55

4. 4. Average RMSEs for item discrimination parameters when models were misspecified with a low degree of cross-loading 56

4. 5. Average RMSEs for item discrimination parameters when models were misspecified with a medium degree of cross-loading 57

4. 6. Average RMSEs for item discrimination parameters when models were misspecified with a high degree of cross-loading 58

4. 7. Average RMSEs for item location parameter when models were specified correctly. 59

4. 8. Average RMSEs for item location parameter when models were misspecified. 60

4. 9. Average bias for item discrimination parameters when models were specified correctly with a low degree of cross-loading 76

4. 10. Average bias for item discrimination parameters when models were specified correctly with a medium degree of cross-loading 77

Figure Page

4. 11. Average bias for item discrimination parameters when models were specified correctly with a high degree of cross-loading 78

4. 12. Average bias for item discrimination parameters when models were misspecified with a low degree of cross-loading..79

4. 13. Average bias for item discrimination parameters when models were misspecified with a medium degree of cross-loading..80

4. 14. Average bias for item discrimination parameters when models were misspecified with a high degree of cross-loading...81

4. 15. Average bias for item location parameter when models were specified correctly. .82

4. 16. Average bias for item location parameter when models were misspecified. 83

4. 17. Average SEE for item discrimination parameters when models were specified correctly with a low degree of cross-loading 97

4. 18. Average SEE for item discrimination parameters when models were specified correctly with a medium degree of cross-loading 98

4. 19. Average SEE for item discrimination parameters when models were specified correctly with a high degree of cross-loading 99

4. 20. Average SEE for item discrimination parameters when models were misspecified with a low degree of cross-loading..100

4. 21. Average SEE for item discrimination parameters when models were misspecified with a Medium degree of cross-loading 101

4. 22. Average SEE for item discrimination parameters when models were misspecified with a high degree of cross-loading...102

4. 23. Average SEE for item location parameter when models were specified correctly. 103

4. 24. Average SEE for item location parameter when models were misspecified. 104

CHAPTER I

INTRODUCTION

Item Response Theory (IRT)

Item Response Theory (IRT) refers to a family of mathematical and statistical models that explain the relationship between examinees' latent ability, and the responses to items manifesting that latent ability. The ultimate goal of IRT is to measure the location of examinees in terms of the underlying latent ability that is hypothesized to be measured utilizing a test, survey or an instrument (Reckase, 2009). Generally, based on examinees' item response patterns, IRT analyzes the data and estimates the examinees' ability and item parameters such as measures of item location, item discrimination, etc. (Reckase, 2009). Mathematically, IRT relates an examinee's ability level (θ) to the probability of endorsing or responding correctly to an item (Lord, 1980). Technically, IRT estimates unique item parameters such as item location, item discrimination, and item guessing parameter, and then estimates examinees' latent ability based on their item response patterns. IRT models can be classified into two categories of dichotomous and polytomous. Dichotomous IRT models can be used for response data to items with two possible response categories such as "Yes/No" or "True/False", etc. Polytomous IRT models, on the other hand, can be used for response data to items having more than two response categories such as five-level Likert scale (e.g., "Strongly agree", "Agree", "Neutral", "Disagree", "Strongly disagree") (Reckase, 2009).

Unidimensional IRT Models (UIRT)

Unidimensional IRT models (UIRT) measure only a single ability that the item response data may represent. One of the underlying assumptions of unidimensional IRT is that all of the items in an instrument, survey, or a test measure a single ability (unidimensionality). Generally, in UIRT models a single ability is required for examinees in order to correctly respond to a set of items. In addition, it is assumed that examines respond to each test item as an independent event (Reckase, 2009). For example, the reading section (sub-test) of the Test of English as a Foreign Language (TOEFL iBT®) is hypothesized to measure the ability of the English as a Second Language (ESL) learner’s English reading skills (TOEFL iBT, 2020).

Multidimensional IRT Models (MIRT)

In practice and real-world situations, the assumption of unidimensionality is less likely to hold. Therefore, multidimensional IRT (MIRT) models have been utilized to model and estimate multiple latent abilities of the examinees at the same time (Ackerman, 1996; Ansley & Forsyth, 1985; Reckase, 1985). Dimensionality is related to the number of underlying latent abilities assessed by an exam, survey or set of items in an instrument. In real world situations, it is common that multiple latent abilities are present in a single educational test or psychological survey that leads to a potentially multidimensional structure of item response data (Svetina & Levy, 2016).

For instance, in the Quantitative Reasoning sub-test of GRE® exam especially for long word problems, the sub-test may be hypothesized to measure the single ability of examinees in terms of quantitative reasoning skills (GRE, 2018). However, in reality if the examinee has limited vocabulary inventory and reading skills the examinee may not successfully respond to an item that is supposed to measure quantitative reasoning skills even if the examinee is competent in this area. Thus, the data may be multidimensional, requiring vocabulary and reading skills in addition to quantitative reasoning.

There are many ways of describing MIRT models according to the assumed relationship among latent abilities and dimensional structure of a set of items in a test measuring those abilities. For example, in a mathematic test, some items may be hypothesized to measure algebra skills, and some other items may be hypothesized to measure geometry skills (two sub-tests in which each measures one ability). In addition, it is likely that in a mathematics test some items may be hypothesized to measure algebra skills, but also require some geometry skills in order for an examinee to respond to that particular item correctly. This means a set of items measuring two abilities at the same time or the items might be hypothesized to measure one ability but in reality, it requires two abilities for an examinee to respond to the set of items successfully (Reckase, 2009).

Compensatory and Non-compensatory MIRT Models

MIRT models can also be categorized into two types: compensatory and non-compensatory. A compensatory MIRT model measures multiple abilities in such a way that a low ability on one dimension to be compensated for by a high ability on another dimension. Mathematically, this is modeled by summing the probabilities of a correct response across dimensions (Fox, Klein, Entink & Avetisyan, 2014; Reckase, 1985). An example of an application of a compensatory MIRT model could be a TOEFL test that measures multiple abilities such as speaking, listening, reading and writing skills in ESL learners (TOEFL iBT, 2020). If the examinee is strong on writing ability then the strength on the writing ability can compensate for the speaking weaknesses on the speaking sub-test.

On the other hand, the non-compensatory MIRT model assumes a high ability on one dimension does not compensate for a lower ability on the other dimension. Mathematically, a non-compensatory model uses multiplication to calculate the probability of a correct response as opposed to summing the probabilities of a correct response across dimensions in a compensatory MIRT models (Ackerman, 1992; Sympson, 1978). An example of a non-compensatory MIRT model could be the MCAT test where it measures the examinees' abilities in the areas of Biological and Biochemical Foundations of Living Systems; Chemical and Physical Foundations of Biological Systems; Psychological, Social, and Biological Foundations of Behavior; and Critical Analysis and Reasoning Skills (MCAT®, 2020). If an examinee demonstrates strength in the area of Biological Foundations it might not compensate for the weaknesses in the area of Physical Foundations (Biology knowledge versus Physics knowledge of the examinee).

Technically, compensatory and non-compensatory MIRT models analyze the response data patterns and how the examinees utilize their multiple abilities to respond to a set of items. Compensatory and non-compensatory MIRT models differ based on the theoretical foundation whether multiple abilities work together or the fact that abilities are independent of one another. Generally, non-compensatory MIRT model is more complex in terms of parameter estimation and interpretation (Ackerman, 1992; Fox et al., 2014; Reckase, 1985; Sympson, 1978).

Structure Complexity of MIRT Models

When data are multidimensional, it may be that the complete set is made of multiple unidimensional data sets or that, in a set of items, each individual item measure multiple abilities. A simple structure concept was initially introduced in the area of factor analysis (Thurstone, 1947). In the area of psychometry and measurement a simple structure refers to the fact that within a set of items, subsets of items highly load on, or measure, one of the multiple dimensions and has no loading on the other dimensions (Finch, 2006). Technically, a simple structured test is a multidimensional test that is composed of multiple unidimensional sub-tests in which subsets of items measure only a single ability (Finch, 2011; Svetina & Levy, 2016). Structure of a set of items for a population of examinees is simple if items depend on only one underlying latent ability (McDonald, 1999). For instance, in the TOEFL test each sub-test is hypothesized to measure one ability of the examinee. The, reading sub-test is hypothesized to measure the examinees' ability in the area of reading skills in English language. The listening sub-test is supposed to measure the examinees' ability in the area of listening skills in English language (TOEFL iBT, 2020).

A complex structure, on the other hand, consists of unique sets of items that are associated with more than one dimension (Finch, 2011; Svetina & Levy, 2016). Technically, the structure of a set of items for a population of examinees is complex if the items depend, or load, on multiple underlying latent abilities (McDonald, 1999). As a result, a complex structure refers to a condition where at least one item requires an examinee to demonstrate more than one underlying latent ability in order to respond to the item successfully (McDonald, 1999; Reckase, 2009). For example, in a mathematics test it is likely some items may be hypothesized to measure algebra skills, but also require some geometry or trigonometry skills in order for an examinee to respond to that particular item correctly (an individual item measuring multiple abilities at the same time).

Understanding the structure complexity of the data is imperative in order to utilize the correct MIRT model and to make appropriate and accurate inferences about the data especially when appropriate interpretation of examinees' score across tests is considered. Parameter estimation such as item location and item discrimination may be affected if a unidimensional model is applied to multidimensional data (Finch, 2011; Svetina & Levy, 2016). In addition, item parameter estimates such as item location, item discrimination and examinees' ability estimation and their interpretation may also be affected if a simple structure model is applied to the data with a complex structure, or vice versa (Svetina & Levy, 2016).

It is worth mentioning that sometimes a complex structure can be approximated to a simple structure where each item depends predominantly and strongly on one primary underlying latent ability and relatively weakly on other secondary latent abilities (Hulin et al., 1983; Strachan et al., 2020; Svetina & Levy, 2016). In this study, this phenomenon is referred to as the "degree of cross-loading" where it defines how strongly underlying latent abilities in complex structure MIRT models are related to the items. The degree of cross-loading is an indication of how strongly primary and secondary dimensions are associated with the item. Are these dimensions being measured equally strongly by the item? It should be considered if the item measures a dominant dimension and potentially one or more weaker dimensions or the item measures both dimensions equally strongly. Additionally, in this study, “structure complexity magnitude” is referred to as the percent of the items in a test or sub-test that demonstrate a complex structure. If 5 of the 10 items in a test or subtest display a complex structure, the structure complexity magnitude would be 50%. If 10 of 10 items have a complex structure, this would be 100%.

Furthermore, in this study, model specification, or the degree to which the true structure of the data meets the structure fit to the data, will be one of the variables that will be evaluated in terms of item parameter estimation precision. Misspecification refers to the situation where the true dimensionality of the data does not equal the dimensionality of the model fit to the data. One of the situations that misspecification can occur and is the interest of this study is when fitting a simple structure model to data that have a true complex structure MIRT model. Investigation of model specification is important due to the fact that sensitivity to model specification can either result in biases and inaccuracy in item parameter estimation and interpretation or the estimation procedure is robust to the model misspecification.

Statement of the Problem

One of the assumptions of standard item response theory (IRT) models is that the underlying latent ability being measured is unidimensional in nature (Finch, 2011). However, there are a great number of surveys, instruments and tests that measure multiple latent abilities, which leads to a potentially multidimensional structure of item response data. Even when data are multidimensional, structure complexity (simple structure or complex structure) of the data should be considered in order to ensure the precision of item parameter or examinees’ ability score estimation and appropriate interpretation of examinees’ score and item characteristics. Therefore, understanding the true structure of the data is imperative in order to make appropriate inferences about the underlying latent abilities (Svetina & Levy, 2016; Sevetina et al., 2017; Zhang 2007; Zhang, 2012). In addition, understanding the effects of model specification, i.e., when the structure of the model applied is not that of the true structure of the data, is imperative to ensure the precision of item parameter estimation and appropriate interpretation of examinees' score and item characteristics when dealing with multidimensional data. As a result, assessing dimensionality is beneficial and necessary prior to applying item response theory (IRT) models in social sciences (Finch 2010; Finch, 2011; Strachan et al., 2020; Svetina & Levy, 2016).

In this study, cross-loading refers to the items in a test, survey, or an instrument that are associated with multiple abilities at the same time or those items that require an examinee to demonstrate knowledge on multiple underlying abilities at the same time whether it is hypothesized that those items measure all underlying latent abilities equally strongly or not. Thus, the degree of cross-loading is an indication of how strongly primary and secondary dimensions are associated with the items. In addition, in this study, structure complexity magnitude refers to the number or percentage of the items in a test, survey or an instrument that exhibit cross loading. Previous research studies discussed dimensionality assessment and its performance precision with various structure complexity levels (Finch & Habing, 2005; Svetina, 2013, Svetina & Levy, 2016). Some other researches have taken into account multidimensionality and structure complexity in order to evaluate item parameter estimation precision (Finch, 2011, Svetina et al., 2017; Zhang, 2012).

In practice and in real-world situations, it is very likely that when the items in a test exhibit a complex structure with a strong loading on one ability but a small loading on the other ability (small degree of cross-loading), the data are treated as having a simple structure, ignoring the small cross-loading of some items. For example, let's consider a mathematics test with 20 items. The assumption is that the 10 algebra items measure only algebra knowledge and the 10 geometry items measure only geometry knowledge. However, in the reality some of the items might primarily measure ability of the examinee in algebras skills, but also require for an examinee to have some geometry knowledge in order to respond to the item successfully (cross-loading). Similarly, for those items that primarily test examinees ability related to geometry may require some algebra skills for an examinee in order to respond to the item correctly.

Previous studies have investigated the effects of structure complexity, the correlation between the underlying latent abilities, sample size, distribution of examinees on dimensionality assessment and item parameter recovery on complex structure MIRT models (Finch, 2011; Finch & Habing, 2005; Svetina, 2013; Svetina & Levy, 2016; Svetina et al., 2017; Zhang 2012). However, previous studies did not consider the degree of cross-loading on secondary dimension (how strongly each ability is being measured with an item) and model specification (misspecified simple structure model when the data are truly complex) and their effects on item parameter estimation precision. In addition, previous studies did not collectively discuss the effects of structure complexity magnitude of the data incorporating the degree of cross-loading on secondary dimension, and model specification on item parameter estimation in complex structure MIRT models.

Purpose of the Study

The purpose of this study is to investigate the impact of structure complexity magnitude of the data incorporating the degree of cross-loading on secondary dimension, and model specification (misspecified simple structure model when data are truly complex) on item parameter estimation in Multidimensional Item Response Theory (MIRT).

Research Question

What are the effects of structure complexity magnitude of the data incorporating the degree of cross-loading on secondary dimension, and model specification on item parameter estimation in Multidimensional Item Response Theory (MIRT)?

CHAPTER II

REVIEW OF LITERATURE

Item Response Data

Item response data refers to data collected from responses to sets of items on an instrument, e.g., test, survey, etc.). Dichotomous data refers to the items with two possible response categories such as “Yes/No” or “True/False”, etc. On the other hand, polytomous data refers to the items having more than two response categories such as five-level Likert scale (“Strongly agree”, “Agree”, “Neutral”, “Disagree”, “Strongly disagree”) (Reckase, 2009). This study primarily utilizes dichotomous data.

Multidimensional Item Response Theory (MIRT) Models

As it was stated in the first chapter, one of the assumptions of IRT is unidimensionality. Unidimensional item response theory models measure one underlying latent ability of the examinees. However, in practice and real-world situations this assumption is less likely to hold within educational tests and psychological instruments (Ackerman, 1992; McDonalds, 1999; Reckase, 1985; Sympson, 1978). As a result, multidimensional item response theory (MIRT) was introduced to address the complexity of educational tests and psychological instruments in terms of measuring multiple underlying latent abilities concurrently (Reckase & McKinley, 1982; Reckase, 1985; Reckase, 1997; Sympson, 1978).

Assumption of MIRT Models

Mathematically speaking MIRT models explain the relationship between an examinee's underlying ability, or location along an ability scale, and the probability of responding to an item correctly. This relationship is also associated with item parameters such as item location and item discrimination. Item parameter values such as item location and item discrimination can be estimated from the examinees' response data. The development of IRT models are based on a number of assumptions (Reckase, 2009).

In addition to the assumed number of dimensions in a dataset, the assumption of local independence in IRT models refers to the fact that probability of responding to an item successfully by an examinee with a certain level of ability is independent of the probability of responding correctly to the other items in a test (Lord, 1980). Moreover, it is assumed that the location of the examinee, i.e., an examinee's underlying ability, does not change or is independent of the set of items within the assessment. However, this assumption may not hold in practice and real world as examinees may learn from the interaction with the other items or even cheating may result in some learning or changing the location of the examinee (Reckase, 2009).

The assumption of monotonicity in MIRT models refers to a situation that the probability of responding to an item correctly by an examinee increases as the locations of examinees increase on any of the dimensions. Reckase (2009) noted that "the relationship between locations in the multidimensional space and the probabilities of correct response to a test item can be represented as a continuous mathematical function." This means that for every ability location there is one and only one value of probability of correct response associated with it and that probabilities are defined for every location in the multidimensional space and there are no discontinuities. This assumption is important for the mathematical forms of models that can be considered for representing the interaction between persons and test items." (Reckase, 2009)

MIRT models can be categorized in various ways. In what follows, MIRT models will be elaborated based on response category type (dichotomous or polytomous), MIRT models based on the way they explain how examinees utilize their multiple abilities to respond to an item (compensatory or non-compensatory), and MIRT models based on structure complexity (simple or complex structure).

Dichotomous and Polytomous MIRT Models

MIRT models are mathematical expressions considering multiple abilities that are required for an examinee in order to respond to an item successfully (Reckase, 2009). MIRT models can be categorized in terms of response types of dichotomous or a polytomous response type (Reckase, 1985; Sympson, 1978). Dichotomous IRT models refers to the items with two possible response categories such as “Yes/No” or “True/False”. Polytomous IRT models, on the other hand, can be used for items having more than two response categories such as five-level Likert scale (“Strongly agree”, “Agree”, “Neutral”, “Disagree”, “Strongly disagree”) (Reckase, 2009).

Compensatory and Non-compensatory MIRT Models

MIRT models can also be categorized into two types of compensatory and non-compensatory. Technically, compensatory and non-compensatory MIRT models analyze the response data patterns and how the examinees utilize their multiple abilities to respond to an item. Compensatory and non-compensatory MIRT models differ based on the theoretical foundation whether multiple abilities work together or that abilities are independent of one another. Generally, a non-compensatory MIRT model is more complex in terms of parameter estimation and interpretation (Ackerman, 1987a; Ackerman, 1992; Reckase, 1985; Sympson, 1978).

A compensatory MIRT model measures multiple abilities in such a way that a low ability on one dimension to be compensated for by a high ability on another dimension. Mathematically, this is modeled by summing the probabilities of a correct response across dimensions (Ackerman, 1987a; Ackerman, 1992; Reckase, 1985). An example of an application of a compensatory MIRT model could be a TOEFL test that measures multiple abilities such as speaking, listening, reading and writing skills in ESL learners (TOEFL iBT®, 2020). If the examinee is strong on writing ability then the strength on the writing ability can compensate for the speaking weaknesses on the speaking sub-test.

On the other hand, the non-compensatory MIRT model assumes a high ability on one dimension does not compensate for a lower ability. Mathematically, a non-compensatory model uses multiplication to calculate the probability of a correct response as opposed to summing the probabilities of a correct response across dimensions in a compensatory MIRT models (Ackerman, 1992; Sympson, 1978). An example of a non-compensatory MIRT could be MCAT test where it measures the examinees' abilities in the area of Biological and Biochemical Foundations of Living Systems; Chemical and Physical Foundations of Biological Systems; Psychological, Social, and Biological Foundations of Behavior; and Critical Analysis and Reasoning Skills (MCAT®, 2020). If an examinee demonstrate strength in the area of Biological Foundations it might not compensate for the weaknesses in the area of Physical Foundations (Biology knowledge versus Physics knowledge of the examinee).

The focus of this study is a compensatory dichotomous two-parameter logistic (2PL) MIRT model. Reckase (1985) introduced Equation 1 that expresses a multidimensional compensatory 2PL item response theory. The probability of examinee *j* responding correctly to item *i*, $(U_{ij} = 1)$ is a function of a vector of the discrimination parameters for item *i* $(\boldsymbol{a}_i')$ across m dimensions, a vector of person *j*'s ability on the m dimensions $(\boldsymbol{\theta}_j)$ and the multidimensional location of item *i* (d_i);

$$P\left(U_{ij} = 1 \,\middle|\boldsymbol{a}_i', \boldsymbol{\theta}_j, d_i\right) = \frac{e^{\boldsymbol{a}_i'\theta_j + d_i}}{1 + e^{\boldsymbol{a}_i'\boldsymbol{\theta}_j + d_i}} \qquad (1)$$

where $P\left(U_{ij} = 1\right)$ is the probability of a correct response (1) to item *i* by examinee *j* in a dimensional ability space. $\boldsymbol{\theta}_j$ is a $m \times 1$ vector of person *j*'s ability on the *m* dimensions, $\boldsymbol{a}_i'$ is a

$m \times 1$ vector of the discrimination parameters for item i across m dimensions, and d_i is measure of the multidimensional location of item i.

Lastly, MIRT models can be categorized in terms of dimensional structure of the items on an educational test or a psychological instrument. Multidimensionality can be categorized as between item or within item multidimensionality. Between item multidimensionality is associated with a test consisting of several unidimensional parts; that is referred to as simple structure. Within item multidimensionality, on the other hand, is associated with an item in a test that measures more than one underlying latent ability; that is referred to as complex structure MIRT model (Adam, Wilson & Wang,1997). When data are multidimensional, it may be that the complete set is made of multiple unidimensional data sets (between multidimensionality or simple structure) or that each individual item measure multiple abilities (within multidimensionality or complex structure). The simple structure concept was initially introduced in the area of factor analysis (Thurstone, 1947). In the area of psychometry and measurement, a simple structure refers to the fact that a few items load primarily and highly on an underlying latent ability and low on the other underlying latent abilities suggesting an association between the underlying latent ability and the item (Finch, 2006).

Simple Structure MIRT Model

A simple structured test is a multidimensional test that is composed of multiple unidimensional sub-tests in which each item measures a single ability (Finch, 2011, Svetina & Levy, 2016). Structure of a set of items for a population of examinees is simple if some sets of items depend on only one underlying latent ability and other sets of items depend on a different single latent ability (McDonald, 1999; Zhang & Stout, 1999). Technically, in a simple structure MIRT model the vector of discrimination parameters for each item on the m dimensions has only one non-zero value, meaning the item discriminates on a single dimension. Typically, there is a correlation between the abilities in simple structure MIRT models (Finch, 2011). Generally, simple structure MIRT models were introduced to minimize the number of factors needed to explain each variable where additional complexities associated with MIRT models can be decreased if the simple structure model fits the data reasonably (McDonald, 1999; Reckase, 2009).

For instance, in a TOEFL test each sub-test is hypothesized to measure one ability of the examinee (TOEFL iBT®, 2020). For example, the reading sub-test is hypothesized to measure the examinees' ability in the area of reading skills in English language. The listening sub-test is intended to measure the examinees' ability in the area of listening skills in English language (TOEFL iBT®, 2020). Figure 2.1 demonstrates a simple structure MIRT model where two underlying abilities (θ_1 and θ_2) with a correlation of ρ are measured by 10 items, five for each ability.

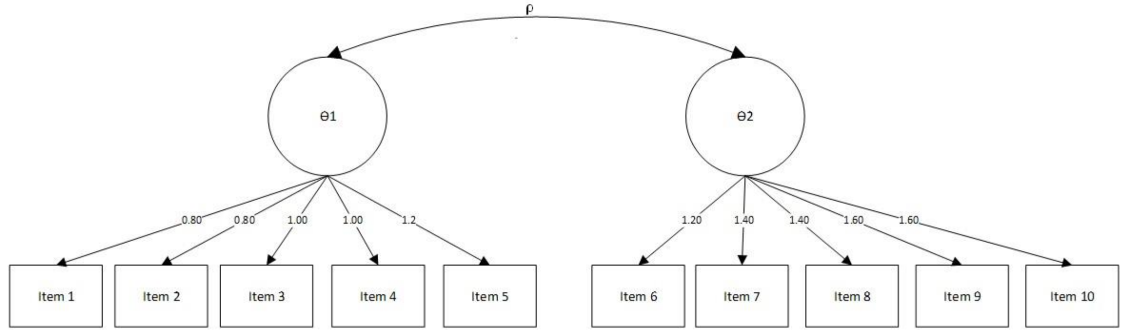

Figure 2. 1. A simple structure MIRT model with two underlying abilities (θ_1 and θ_2) and correlation ρ.

Complex Structure MIRT Model

MIRT models that are associated with within item multidimensionality in which an item in a test measures more than one underlying latent ability. Within item dimensionality is referred to as complex structure (Wilson & Wang, 1997). In MIRT models, structure complexity has to do with allowing the discrimination parameter on each dimension to vary and indicates the degree to which the item measures that dimension. Technically, in a complex structure MIRT model each item loads on multiple abilities (items exhibit cross loadings on multiple dimensions) (Finch, 2011; Mc Donald, 1999; Svetina & Levy, 2016).

Educational tests and psychological instruments are more likely to represent a complex structure MIRT model where more than one underlying latent abilities are being measured simultaneously (Finch, 2011; Reckase, 2009). For example, in a mathematics test it is likely some items may be hypothesized to measure algebra skills, but also require some geometry or trigonometry skills in order for an examinee to respond to that particular item correctly (an individual item measuring multiple abilities at the same time). Figure 2.2 illustrates a complex structure MIRT model where two underlying abilities (θ_1 and θ_2) with a correlation of ρ are measured by 10 items. As shown in Figure 2.2, item 4 and item 7 load on both abilities of θ_1 and θ_2, meaning some amount of ability on both dimensions is required to answer these items correctly. When an item is associated with multiple abilities it is referred to as "cross-loading". For example, item 4 and item 7 in figure 2 demonstrate cross-loading in which they are associated with both abilities of θ_1 and θ_2.

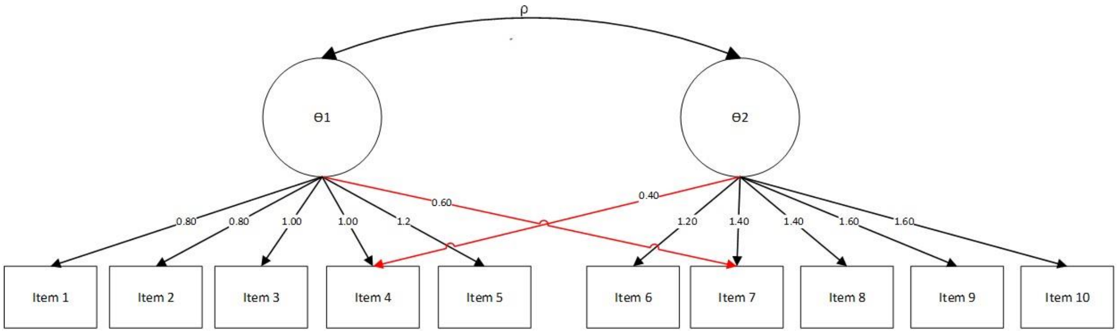

Figure 2. 2. A complex structure MIRT model with two underlying abilities (θ_1 and θ_2) and correlation ρ.

Understanding the structure complexity of the data is imperative in order to utilize the correct MIRT model and to make appropriate and accurate inferences about the data, especially when appropriate interpretation of examinees' score across tests is considered. Parameter estimation such as item location and item discrimination may be affected if a unidimensional model is applied to multidimensional data (Finch, 2011, Matlock & Turner, 2016; Svetina & Levy, 2016). In addition, item parameter estimates such as item location, item discrimination and examinees' ability estimation and interpretation may also be affected if a simple structure model is applied to the data with a complex structure, or vice versa (Svetina & Levy, 2016).

It is worth mentioning that sometimes a complex structure can be approximated to a simple structure where each item depends predominantly and strongly on one underlying latent ability while the items may depend on the other underlying latent ability relatively weakly. That means a test may not measure all dimensions equally (Hulin et al., 1983; Strachan et al., 2020; Svetina & Levy, 2016).

Variables with Potential Influence on Item Parameter Estimation in MIRT Models

There are a number of variables that could potentially influence item parameter estimation, or recovery, and must be taken into account to evaluate the precision of item parameter estimation when applying a MIRT model. These variables include but are not limited to sample size, correlation level between the latent abilities, estimation method, number of items, distribution of the latent abilities, structure complexity of the data, etc. Previous studies in the area of item recovery applying MIRT models have frequently investigated variables such as sample size, correlation level between the abilities, distribution of the latent abilities and the number of items (Bolt & Lall, 2003; Finch, 2010, 2011; Svetina et al., 2017; Zhang, 2012).

Bolt and Lall (2003) investigated item parameter estimation precision of multidimensional compensatory and non-compensatory item response models. In this study, the authors performed a simulation study in order to evaluate parameter recovery for the multidimensional two-parameter logistic model (M2PL) and the multidimensional latent ability model (MLTM) under various conditions. Manipulated variables were including sample size at two levels (1,000, 3,000), number of items at two levels (25, 50), and three levels of correlation between abilities (.0, .3, and .6). In addition, Latent ability parameters were generated from a multivariate normal distribution, with each latent ability dimension having a mean of 0 and variance of 1. Item parameter estimation results were evaluated in terms of the root mean square error (RMSE) for each parameter. The results suggested that for M2PL model parameter recovery appeared to be reasonably good using the MCMC algorithm in terms of RMSEs. However, it was consistently inferior to NOHARM method. For the MLTM, on the other hand, sample size, number of items and latent ability correlation had noticeable effects on parameter recovery. Relative to the M2PL, less precise item parameter estimations were obtained for the MLTM model. Further, the authors noted that MCMC method for M2PL model was more successful and accurate than MLTM method especially as the correlation between latent abilities increased (Bolt & Lall, 2003).

Finch (2010) investigated the accuracy of item parameter estimates in the area of MIRT model context. In this study, the author examined the two MIRT estimation methods of unweighted least squares (ULS) and robust weighted least squares (RWLS), and the unidimensional estimation approach under a variety of conditions such as sample sizes, test lengths, intertrait correlations, pseudo-guessing, and latent ability distribution using a simulation study and software packages such as NOHARM, MPlus, and BILOGMG. The results were evaluated based on overall accuracy, bias, and standard error of item parameter estimates. Results indicated that regardless of the distribution of the latent ability, MPlus bias was much higher in the 3PL than the 2PL case. Results also indicated that estimates provided by both methods were influenced by the distribution of the latent abilities where, in skewed cases, larger standard errors were calculated for NOHARM and MPlus estimates of item location and discrimination. In addition, both techniques exhibited greater location bias in the skewed condition. Moreover, in this study, the author, noted that higher correlation values in the skewed conditions demonstrated a greater bias in ULS discrimination and location estimates. However, the standard errors of the unidimensional estimates were not greatly influenced by any of the manipulated variables in this study.

Finch (2011), investigated the accuracy of item location and discrimination parameter estimation using NOHARM in the multidimensional Item Response Theory (MIRT) models when some items exhibit a complex structure. In this study, the author performed a simulation study in order to evaluate item parameter recovery for MIRT models. In this study, manipulated variables were including the number of examinees at four levels (250, 500, 1,000, and 2,000), four level of correlation between latent abilities (.0, .3, .5, or .8), distribution of latent abilities at two level of normal and non-normal, and three level of structure complexity (simple, semi-complex and complex). In this study, the results indicated that MIRT model parameter estimates including both item location and item discrimination exhibited lower levels of bias for items that did not exhibit simple structure when two latent abilities were present compared to a unidimensional estimation approach. In addition, the results in this study indicated that item discrimination parameters were consistently underestimated when the latent abilities were non-normal. Further, the author noted that both bias and standard error increased when item response data did not conform simple structure (Finch, 2011). This study did not investigate the complexity magnitude in a more detailed approach and more of a general definition of semi-complex and complex structure were introduced compared to a detailed complexity magnitude. In addition, this study did not investigate the effect of complex structure incorporating the effect of the degree of cross-loading on secondary dimension (how strongly each ability is being measured with an item) and model specification (misspecified simple structure model when the data are truly complex) and their effects on item parameter estimation precision.

Zhang (2012) conducted a simulation study in order to compare the unidimensional and multidimensional approaches with the marginal maximum likelihood method (MMLE) when a test or an assessment instrument is composed of several unidimensional subtests or it exhibit a simple structure. In this study, a simulation study was utilized in order to evaluate item parameter estimation under various conditions such as sample size at six levels (500, 1,000, 2,000, 3,000, 4,000, and 5,000), the number of items at three levels of 30, 46, or 62, and three levels of correlation between abilities (.0, .5, and .8). RMSEs and average of RMSEs (ARMSE) were used to evaluate the item parameter estimations. The author in this study noted that the unidimensional and multidimensional approaches are equivalent in parameter estimation if the joint maximum likelihood method was used as a method of estimation. However, estimation results of these two approaches differ if the marginal maximum likelihood method (MMLE) is applied. In this study, the simulation results indicated that item parameter estimation utilizing a multidimensional approach was more precise than item parameter estimation utilizing a unidimensional approach when the number of items in a test or an instrument was small. In addition, the author, investigated the effect of structure complexity on the item parameter estimation. Results of this study indicated that the correlation coefficients between abilities were overestimated when a set of response data did not have a simple structure but was specified as a simple structure (Zhang, 2012). This study did not discuss a detailed complexity magnitude and its effect on the precision of item parameter estimation and was only performed on comparing a simple structure and a mixed structure and their effect on item parameter estimation performance. In addition, this study did not address the degree of cross-loading on secondary dimension and its effect on the precision of item parameter estimation. In addition, this study did not discuss model specification and its effect on item parameter estimation in MIRT models. As it was stated in the previous chapter sometimes a complex structure can be approximated to a simple structure where each item depends predominantly and strongly on one primary underlying latent ability and relatively weakly on other secondary latent abilities (Hulin et al., 1983; Strachan et al., 2020; Svetina & Levy, 2016). The degree of cross-loading is an indication of how strongly primary and secondary dimensions are associated with the item and should be considered if the item measures a dominant dimension and potentially one or more weaker dimensions or the item measures both dimensions equally strongly.

Svetina et al. (2017) utilizing a simulation study investigated the effects of complex structures and the distribution of examinees' latent ability on item parameter recovery in dichotomous compensatory MIRT models. In this study, manipulated variables were including two levels of model type, three levels of correlation between dimensions (.0, .4, or .7), three levels of distribution of the latent variables (normal-skewed, skewed-skewed or normal-normal), five levels of complexity (simple, balanced 20%, balanced 40%, imbalanced 20% and imbalanced 40%). The fully crossed design yielded 180 conditions with 500 replications each. Evaluation criteria in this study were bias and root mean square error (RMSE). The authors in this study noted that when latent abilities were skewed, item parameter recovery was generally adversely impacted. In addition, the presence of complexity contributed to decreases in the precision of parameter recovery, particularly for discrimination parameters along one dimension when at least one latent ability was generated as skewed (Svetina et al., 2017). This study did not address the effect of degree of cross-loading on secondary dimension when item response data represent a complex structure. Sometimes a complex structure can be approximated to a simple structure depending on the degree of cross-loading where each item depends predominantly and strongly on one primary underlying latent ability and relatively weakly on other secondary latent abilities. Therefore, model specification incorporating the degree of cross-loading could affect the precision of item parameter estimation. This study did not investigate the effect of model specification incorporating the degree of cross-loading on the precision of item parameter estimation.

Conditions of the Current Study

Components of the aforementioned studies investigated the effects of sample size, model type (2PL or 3PL), correlation between latent abilities, the distribution of examinees' ability, structure complexity magnitude of the data, but none investigated , structure complexity magnitude incorporating the degree of cross-loading on secondary dimension (low, medium, high), and model specification. Though there exists a wide collection of variables that may affect item parameter estimations in MIRT models, the following were chosen as the focus of this study:

sample size, correlation between latent abilities, structure complexity magnitude of the data, the degree of cross-loading on secondary dimension, and model specification.

Sample Size (Number of Examinees)

One of the variables that may influence the item parameter estimation in the MIRT models is the number of examinees or the sample size. Previous studies have investigated the effect of a broad range of sample size from a small (500) to very large sample size (5,000) on the item parameter estimation in the MIRT models (Bolt & Lall, 2003; Finch, 2010, 2011; Jing et al., 2016; Zhang, 2012). Bolt and Lall (2003) explored the effect of sample size at two levels of 1,000 and 3,000 on item parameter estimation in compensatory and non-compensatory MIRT models. Finch (2010) and Finch (2011) simulated the number of examinees at four levels of 250, 500, 1,000 and 2,000 in order to investigate the item parameter precision under the influence of various sample sizes. Zhang (2012) evaluated item parameter estimation under various conditions and six levels of sample sizes (500, 1,000, 2,000, 3,000, 4,000, and 5,000). In this study, in order to stay with a manageable number of conditions three level of sample sizes of 500, 1,000 and 2,000 examinees will be selected in order to investigate the effect of sample size on the item parameter estimation in MIRT models.

Correlation between Latent Abilities

One of the variables that might affect the precision of item parameter estimation in MIRT models is the correlation between latent abilities. Previous research studies (Bolt & Lall, 2003; Finch, 2010, 2011; Svetina et al., 2017; Zhang, 2012) have investigated correlation between latent abilities and its effects on item parameter estimation in MIRT models. Those studies have utilized a wide range of correlation values from relatively small values of correlations (.0 and .3) to larger values of correlation (.8 and .9) to investigate the influence of the correlation between latent abilities on item parameter estimation (Bolt & Lall, 2003; Finch, 2010, 2011; Svetina et al., 2017; Zhang, 2012). Bolt and Lall (2003) utilized three levels of correlation between abilities (.0, .3, and .6) in order to investigate item parameter estimation precision (Bolt and Lall, 2003). Using a simulation study Finch (2010) and Finch (2011) investigated the effect of correlation between latent abilities on item parameter estimation in 2PL-MIRT model. The two latent abilities were simulated to be correlated at .0, .3, .5, or .8. The results indicated that that correlated latent abilities had an influence on item parameter estimation (Finch, 2010, 2011). Zhang (2012) utilized three levels of correlation between abilities (.0, .5, and .8) in order to investigate item parameter estimation precision. Svetina et al. (2017) simulated the latent abilities to be correlated at three levels of .0, .4, or .7. In the current study, in order to investigate the effect of correlation between latent abilities on the precision of item parameter estimation in MIRT model the latent abilities were simulated to be correlated at three levels of .0, .6 or .9.

Structure Complexity Magnitude

In this study, “structure complexity magnitude” is referred to the percent of the items in a test or sub-test that demonstrate a complex structure. For instance, if 5 of 10 items display a complex structure, the structure complexity would be 50%. If 10 of 10 items have a complex structure, this would be 100%. Finch (2011) introduced one of the first examinations of parameter estimation in the non-simple structure case. Finch (2011) investigated item parameter estimation precision in MIRT models under two latent abilities with an equal number of items per dimension in two levels of complex and semi-complex. Unfortunately, Finch (2011) did not indicate the exact complexity magnitude in each condition. Finch (2011) noted that “parameter estimates obtained using the MIRT model exhibited lower levels of bias in both discrimination and location parameter estimates for items that do not exhibit simple structure when two latent abilities are present than did unidimensional estimation” (Finch, 2011).

In another study, Zhang (2012) investigated the impact of the violation of the simple structure assumption (structure complexity) on item parameter estimation precision in MIRT models in two levels of simple structure and mixed structure. The results in this research indicated that when a set of response data does not have a simple structure but is specified as such the models will be incorrectly estimated and the correlation coefficients between abilities will be overestimated (Zhang 2012). Svetina et al. (2017) also considered a simple structure and two levels of complexity magnitude (20% and 40 % of items represent a complex structure) and its effect on item parameter estimation in MIRT models. The authors in this study noted that the recovery of item discrimination in imbalanced complexity conditions was generally poorer for complex items in comparison to their simple item counterparts. However, item guessing parameter recovery was better for complex item compared to the simple structured items (Svetina et al., 2017). In order to further investigate the effect of complexity magnitude on item parameter estimation in MIRT models, in the current study, three levels of complexity magnitudes (when 10%, 30% or 50% of items represent a complex structure) incorporating a low, medium, or high degree of cross-loading on secondary dimension will be considered.

Degree of Cross-Loading on Secondary Dimension

As it was stated in the previous chapter, in MIRT models sometimes a complex structure can be approximated to a simple structure where each item depends predominantly and strongly on one primary underlying latent ability and relatively weakly on other secondary latent abilities (Hulin et al., 1983; Strachan et al., 2020; Svetina & Levy, 2016). In this study, this phenomenon is referred to as the “degree cross-loading” where it defines how strongly underlying latent abilities in complex structure MIRT models are related to the items. The degree of cross-loading is an indication of how strongly primary and secondary dimensions are associated with the item. Are these dimensions being measured equally strongly by the item? It should be considered if the item measures a dominant dimension and potentially one or more weaker dimensions or the item measures both dimensions equally strongly. In this study, three levels of low, medium or high degree of cross-loading on secondary dimension will be considered.

Model Specification

One of the variables that can affect the item parameter estimation precision is model specification prior to MIRT estimation procedures (Chen & Jiao, 2013; Strachan et al., 2020). As it was mentioned before, In the MIRT models misspecification refers to the situation where the true dimensionality of the data does not equal the dimensionality of the model fit to the data. One of the situations that misspecification can occur and is the interest of this study is when fitting a simple structure model to data that have a true complex structure MIRT model. Investigation of model specification is important due to the fact that sensitivity to model specification can either result in biases and inaccuracy in item parameter estimation and interpretation or the estimation procedure is robust to the model misspecification (Strachan et al., 2020).

Summary

The purpose of this study was to investigate the impact of structure complexity magnitude of the data, the degree of cross-loading on secondary dimension, and model specification (misspecified simple structure model when data are truly complex) on item parameter estimation in MIRT model. When the item response data are multidimensional, structure complexity (simple structure or complex structure) and complexity magnitude of the data should be considered in order to ensure the precision of item parameter estimation and appropriate interpretation of item characteristics. In addition, understanding the effects of model specification, i.e., when the structure of the model applied is not that of the true structure of the data, is imperative to ensure the precision of item parameter estimation and appropriate interpretation item characteristics when dealing with multidimensional data. In practice and in real-world situations, it is very likely that when the items in a test exhibit a complex structure with a strong loading on one ability but a small loading on the other ability (small degree of cross-loading), the data are treated as having a simple structure, ignoring the small cross-loading of some items.

As it was mentioned in the previous sections, past studies have investigated the effects of structure complexity, the correlation between the underlying latent abilities, sample size, distribution of examinees on dimensionality assessment and item parameter recovery on complex structure MIRT models (Finch, 2011; Svetina & Levy, 2016; Svetina et al., 2017; Zhang 2012). However, previous studies did not consider the degree of cross-loading on secondary dimension (how strongly each ability is being measured with an item) and model specification (misspecified simple structure model when the data are truly complex) and their effects on item parameter estimation precision. In addition, previous studies did not collectively discuss the effects of structure complexity magnitude of the data incorporating the degree of cross-loading on secondary dimension and model specification on item parameter estimation in complex structure MIRT models.

CHAPTER III

METHODLOGY

This study was designed to investigate the impact of structure complexity magnitude of the data, the degree of cross-loading on secondary dimension, and model specification (misspecified simple structure vs correct specified model when data are truly complex) on item parameter estimation in Multidimensional Item Response Theory (MIRT). In this chapter, model specifications and simulation factors, item parameter specifications, structure complexity features, data generation procedure and evaluation criteria of the simulation results are presented.

Simulation Study Design

A simulation study was designed to address the research question regarding the variables influencing the precision of item parameter estimation in the MIRT models. These variables are including structure complexity magnitude of the data, the degree of cross-loading on secondary dimension, and model specification (misspecified simple structure vs correct specified model when data are truly complex). In order to simulate the data, a compensatory MIRT model with two dimensions and dichotomous items response type (Reckase, 1985) was considered for the data generation procedure. Item specifications of the 20 items reported in Svetina (2016) were used to simulate item response data pattern in each condition using geometry and codes written in R Studio (RStudio Team, 2018) and the "mirt" package (Chalmers, 2012).

Manipulated Variables

The following variables were manipulated in order to address the research question regarding the item parameter estimation precision. The discrimination of items on the primary dimension (a_p) was fixed for all conditions. At different levels of structure complexity magnitude and degree of cross-loading, the discrimination on the secondary dimension (a_s) was adjusted. For items with no cross-loading, $a_s = 0$. For those with some magnitude or level of cross-loading, the discrimination on the secondary dimension was less than the discrimination on the primary dimension, i.e., $a_s < a_p$.

Three levels of structure complexity magnitude. Finch (2011) introduced one of the first examinations of MIRT parameter estimation in the non-simple structure case with an equal number of items per dimension in two levels of complex and semi-complex. Unfortunately, Finch (2011) did not indicate the exact complexity magnitude in each condition. Svetina et al. (2017) considered a simple structure and two levels of complexity magnitude (when 20% and 40 % of items represent a complex structure) and its effect on item parameter estimation in MIRT models. In another study, Svetina and Levy (2016) considered three levels of structure complexity magnitude. Three levels of complexity magnitudes in this study are including when 10% (one complex item of 10 items on each dimension), 30% (three complex items of 10 items on each dimension) or 50% (five complex items of 10 items on each dimension) of items represent a complex structure.

Three levels of cross-loading discrimination. In the MIRT models, sometimes a complex structure can be approximated to a simple structure where each item depends predominantly and strongly on one primary underlying latent ability and relatively weakly on other secondary latent abilities (Hulin et al., 1983; Strachan et al., 2020; Svetina & Levy, 2016). The degree of cross-loading is an indication of how strongly primary and secondary dimensions are associated with the item. In this study, three levels of low, medium or high degree of cross-loading on secondary dimension will be considered. Exact values from the Svetina and Levy (2016) for the high degree of cross-loading were used, where the discrimination on the secondary dimension of items ranged from 0.80 to 1.40. A modified medium degree of cross-loading, where the degree of association (as measured by discrimination) ranged from 0.40 to 0.70. A modified low degree of cross-loading was specified where the discrimination on the secondary dimension of items ranged from 0.20 to 0.35.

Three levels of correlation between dimensions. Previous studies investigated the effect of correlation between the dimensions within a variety of simulated correlation levels ranging from .0 to .95 (Bolt & Lall, 2003; Finch, 2010, 2011; Svetina et al., 2017; Zhang, 2012). Bolt and Lall (2003) utilized three levels of correlation between abilities (.0, .3, and .6) in order to investigate the accuracy of item parameter estimation. Using a simulation study Finch (2010) and Finch (2011) investigated the effect of correlation between latent abilities on item parameter estimation M2PL model. The two latent abilities were simulated to be correlated at .0, .3, .5, or .8. Zhang (2012) utilized three levels of correlation between abilities (.0, .5, and .8) in order to investigate the precision of item parameter estimation. Svetina et al. (2017) utilized a simulation study in order to investigate the accuracy of item parameter estimation where the correlations between abilities were set to .0, .4, or .7. In the current study, the data were simulated considering when the correlation between the dimensions was set to .0, .6, or .9.

Three levels of sample size. Previous studies have investigated the effect of a broad range of sample size from a small (500) to very large sample size (5,000) on the item parameter estimation in the MIRT models (Bolt & Lall, 2003; Finch, 2010, 2011; Zhang, 2012). Bolt and Lall (2003) explored the effect of sample size at two levels of 1,000 and 3,000 on item parameter estimation in compensatory and no compensatory MIRT models. Finch (2010) and Finch (2011) simulated the number of examinees at four levels of 250, 500, 1,000 and 2,000 in order to investigate the item parameter precision under the influence of various sample sizes. Zhang (2012) evaluated item parameter estimation under various conditions and six levels of sample sizes (500, 1,000, 2,000, 3,000, 4,000, and 5,000). In this study, in order to stay with a manageable number of conditions three levels of sample sizes of 500, 1,000 and 2,000 examinees were selected in order to investigate the effect of sample size on the item parameter estimation in MIRT models.

Two levels of model specifications. Sometimes a complex structure can be approximated to a simple structure where each item depends predominantly and strongly on one primary underlying latent ability and relatively weakly on other secondary latent abilities (Hulin et al., 1983; Strachan et al., 2020; Svetina & Levy, 2016). In practice and in real-world situations, it is very likely that when the items in a test exhibit a complex structure with a strong loading on one ability but a small loading on the other ability (small degree of cross-loading), the data are treated as having a simple structure, ignoring the small cross-loading of some items. One of the variables that can affect the accuracy of item parameter estimation is model specification prior to MIRT estimation procedures (Chen & Jiao, 2013; Strachan et al., 2020). As it was mentioned before, in the MIRT models misspecification refers to the situation where the true dimensionality of the data does not equal the dimensionality of the model fit to the data. One of the situations that misspecification can occur and is the interest of this study is when fitting a simple structure model to data that have a true complex structure MIRT model considering a high, medium or low degree of cross-loading. Investigation of model specification is important due to the fact that sensitivity to model specification can either result in biases and inaccuracy in item parameter estimation and interpretation or the estimation procedure is robust to the model misspecification (Strachan et al., 2020). In order to investigate the effect of model specification on item parameter estimation, two levels of model specification including the misspecified simple structure model when data are truly complex and the correct model specification were considered in this study.

Item Parameter Specifications

Table 3.1 reports the item parameter specifications for two dimensional compensatory MIRT model for 10 primary items per dimension for three different types of structure complexity magnitude from Svetina and Levy (2016) and three levels of degree of cross-loading, the first of which was taken from Svetina and Levy (2016) and additional medium and low degree of cross-loading on secondary discrimination values.

Table 3. 1. Item Parameters for 2D Compensatory MIRT Model for 10 Items per Dimension for Three Types of Complexity Structures Including Low Degree of Cross-Loading on Secondary Discrimination.

		10% Complex structure						30% Complex structure						50% Complex structure					
		High		Medium		Low		High		Medium		Low		High		Medium		Low	
Item	d	a_1	a_2	a_1	a_2	a_1	a_2	a_1	a_2	a_1	a_2	a_1	a_2	a_1	a_2	a_1	a_2	a_1	a_2
1	-1.5	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-
2	-0.75	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-
3	0	1.00	0.80	1.00	0.40	1.00	0.20	1.00	0.80	1.00	0.40	1.00	0.20	1.00	0.80	1.00	0.40	1.00	0.20
4	0.75	1.00	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00	-
5	1.5	1.20	-	1.20	-	1.20	-	1.20	-	1.20	-	1.20	-	1.20	1.00	1.20	0.50	1.20	0.25
6	-1.5	1.20	-	1.20	-	1.20	-	1.20	-	1.20	-	1.20	-	1.20	1.00	1.20	0.50	1.20	0.25
7	-0.75	1.40	-	1.40	-	1.40	-	1.40	1.20	1.40	0.60	1.40	0.30	1.40	1.20	1.40	0.60	1.40	0.30
8	0	1.40	-	1.40	-	1.40	-	1.40	-	1.40	-	1.40	-	1.40	-	1.40	-	1.40	-
9	0.75	1.60	-	1.60	-	1.60	-	1.60	1.40	1.60	0.70	1.60	0.35	1.60	1.40	1.60	0.70	1.60	0.35
10	1.5	1.60	-	1.60	-	1.60	-	1.60	-	1.60	-	1.60	-	1.60	-	1.60	-	1.60	-
11	1.5	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80
12	0.75	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80	-	0.80
13	0	0.80	1.00	0.40	1.00	0.20	1.00	0.80	1.00	0.40	1.00	0.20	1.00	0.80	1.00	0.40	1.00	0.20	1.00
14	-0.75	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00	-	1.00
15	-1.5	-	1.20	-	1.20	-	1.20	-	1.20	-	1.20	-	1.20	1.00	1.20	0.50	1.20	0.25	1.20
16	1.5	-	1.20	-	1.20	-	1.20	-	1.20	-	1.20	-	1.20	1.00	1.20	0.50	1.20	0.25	1.20
17	0.75	-	1.40	-	1.40	-	1.40	1.20	1.40	0.60	1.40	0.30	1.40	1.20	1.40	0.60	1.40	0.30	1.40
18	0	-	1.40	-	1.40	-	1.40	-	1.40	-	1.40	-	1.40	-	1.40	-	1.40	-	1.40
19	-0.75	-	1.60	-	1.60	-	1.60	1.40	1.60	0.70	1.60	0.35	1.60	1.40	1.60	0.70	1.60	0.35	1.60
20	-1.5	-	1.60	-	1.60	-	1.60	-	1.60	-	1.60	-	1.60	-	1.60	-	1.60	-	1.60
M	0	1.16	1.16	1.13	1.13	1.11	1.11	1.18	1.18	1.05	1.05	0.99	0.99	1.16	1.16	0.98	0.98	0.89	0.89
SD	1.09	0.31	0.31	0.37	0.37	0.41	0.41	0.29	0.29	0.38	0.38	0.48	0.48	0.27	0.27	0.41	0.41	0.51	0.51

For example, considering the 30% complex structure item specifications introduced in Table 3.1, if we suspect that the 20 items in the Svetina and Levy (2016) study are a mathematics test. The assumption is that the 10 algebra items measure only algebra knowledge and the 10 geometry items measure only geometry knowledge. However, in reality items 3, 7, and 9 might primarily measure algebra, but also require some geometry skills; items 13, 17, and 19 primarily test geometry, but require some algebra skills.

In this study, the effects of misspecified simple structure and ignoring the need of the secondary dimension of complex data on item parameter estimates by manipulating the complexity magnitude (when 10%, 30%, or 50% of the items represent a complex structure) and the degree of cross-loading on secondary dimension (low, medium or high degree of cross-loading on secondary dimension) on item parameter estimation in MIRT models was investigated. As reported in Table 3.1, the high degree of cross-loading on secondary dimension (0.80 to 1.40) are the exact values from the Syetina & Levy (2016). In this study, we also investigated the effect of the misspecified model (simple structure) in the presence of medium (0.40 to 0.70) and low (0.20 to 0.35) degree of cross-loadings on the secondary dimension.

Analysis

The described manipulated variables influencing the precision of item parameter estimation in this study led to 162 simulated item response data sets. Each condition combination was replicated 500 times. A compensatory 2PL-MIRT model with two dimensions and dichotomous items response type was used to simulate and calibrate the data for every replication for each condition combination, with a standard bivariate normal θ distribution with three levels of correlation. Parameters were estimated using marginal maximum likelihood (MML), utilizing the expectation-maximization (EM) algorithms. According to Chalmers (2012), the EM algorithm is considered generally as an effective estimation method with few dimensions (Chalmers, 2012).

Instrument

R Studio (RStudio Team, 2018) was used for both data generation and item parameter estimation and analyses. The “mirt” package (Chalmers, 2012) was used in order to generate the simulated item response patterns for a sample size of 500, 1,000 and 2,000 examinees based on the designated distribution of the examinees on latent abilities within each combination of conditions. In addition, item parameters of the correct specified or misspecified MIRT models were estimated using “mirt” package (Chalmers, 2012).

Evaluation Criteria

For each combination of conditions, 500 replications were simulated and item parameter estimates were averaged resulting in mean (M) and standard deviation (SD) of the estimated parameters across replications. In order to investigate the effect of the complexity magnitude of the data, the degree of cross-loading on secondary dimension, and model specification on item parameter estimation, estimated item parameters were compared to true item parameter specifications. Root mean square of error (RMSE), bias (B), and standard error of estimate (SEE) were calculated to evaluate the performance and accuracy of item parameter estimation within each combination of conditions across 500 replications.

The RMSE can be defined as;

$$RMSE_j = \sqrt{\frac{1}{r}\sum_{r=1}^{r}(\hat{v}_{jr} - v_j)^2} \ , \tag{3.1}$$

where r is the number of replications, $\hat{v}_{jr}$ is an estimate for parameter j at replication r, and v_j is a true value of parameter j.

The Bias (B) can be defined as;

$$B_j = \frac{1}{r}\sum_{r=1}^{r} \hat{v}_{jr} - v_j \, , \tag{3.2}$$

where r is the number of replications, $\hat{v}_{jr}$ is an estimate for parameter j at replication r, and v_j is a true value of parameter j. In the case of MIRT, the term v_j can be the discrimination or location parameter for the jth item. Technically, bias is an indicator of whether the parameters were overestimated or underestimated and how much and in what direction the estimated parameters differ from the true value of parameter.

The SEE can be defined as follows;

$$SEE_j = \sqrt{\frac{1}{r}\sum_{r=1}^{r}(\hat{v}_{jr} - \overline{\hat{v}_J})^2} \tag{3.3}$$

Where r denotes the number of replications, $\overline{\hat{v}_J}$ is an average of $\hat{v}_{jr}$ across the r replications.

CHAPTER IV

RESULTS

The purpose of this study was to investigate the impact of structure complexity magnitude of the data, the degree of cross-loading on secondary dimension, model specification (misspecified simple structure model when data are truly complex), sample size and correlation between abilities on item parameter estimation in MIRT model. Simulated item response patterns for sample size of 500, 1,000 and 2,000 examinees were generated and item parameters were estimated within each combination of conditions. Estimated item parameters were compared with the true item parameter specifications in order to investigate the influence of the manipulated variables on item parameter recovery in each combination of conditions.

The described manipulated variables influencing the precision of item parameter estimation in this study led to 162 simulated item response data sets. Each condition combination was replicated 500 times. A compensatory 2PL-MIRT model with two dimensions and dichotomous item response type with a standard bivariate normal θ distribution with three levels of correlation between abilities was utilized to simulate and calibrate the data for every replication for each combination of conditions. Parameters were estimated using marginal maximum likelihood (MML), utilizing the expectation-maximization (EM) algorithms. Within each combination of conditions across the 500 replicated datasets of item response data, the estimated item discriminations on each dimension ($\widehat{a_1}$ and $\widehat{a_2}$) and item location ($\hat{d}$) were calculated.

In this chapter, results are reported in three main sections. Within each section, the results are shown considering the structure complexity magnitude of the data (when 10%, 30% or 50% of items represent a complex structure) incorporating the degree of cross-loading (low, medium and high), sample size (500, 1,000 and 2,000), and model specification (correct specified model compared to misspecified model) at three levels of correlation (.0, .6, and .9). Section I elaborates on the effect of the studied variables (sample size, correlation between the abilities, model specification, structure complexity magnitude and the degree of cross-loading) on the item parameter estimation in terms of average RMSE of the item parameter recovery (true item parameters compared with the estimated item parameters). Section II illustrates the average bias of the item parameter recovery results considering the effect of studied variable across the combinations of conditions. Section III demonstrates the standard error of estimate of the item parameter estimation across the conditions.

Section I: Item Parameter Recovery Results in Terms of Average RMSE

Correct Specified Models

In order to investigate the effect of the manipulated variables on the precision of item parameter estimation the RMSEs were calculated comparing the true item parameter specification with the estimated item parameters across 500 replications. The RMSEs were averaged for the estimated item discrimination parameters (a_1 and a_2) and the estimated item location parameter (d) within each combination of conditions for three sets of item cross-loading. For the item discrimination parameters (a_1 and a_2) in correct specified models, the RMSEs for the primary cross-loaded item discriminations on first and second dimensions (a_{1pcl} and a_{2pcl}) were averaged based on those items that had a primary item discrimination and were cross-loaded on both dimensions. Likewise, the RMSEs for the primary non-cross loaded item discriminations on first and second dimensions (a_{1pncl} and a_{2pncl}) were averaged based on the items that had a primary item discrimination but were not cross-loaded on both dimensions. For the secondary item discriminations (a_{1s} and a_{2s}) the RMSEs were averaged based on the items that had a secondary item discrimination.

Primary item discrimination parameters. Table 4.1 reports the average RMSEs of the estimated item discrimination parameters when the models were correctly specified considering three levels of structure complexity magnitude incorporating low, medium or high degree of cross-loading. As shown in table 4.1, the average RMSEs for the primary cross-loaded item discrimination parameters on the first dimension (a_{1pcl}) ranged from 0.078 to 0.334. The average RMSEs for the primary non-cross loaded item discrimination parameters on the first dimension (a_{1pncl}) ranged from 0.092 to 0.221 across all combination of conditions. On the other hand, the average RMSEs for the primary cross-loaded item discrimination parameters on the second dimension (a_{2pcl}) ranged from 0.075 to 0.340. The average RMSEs for the primary non-cross loaded item discrimination parameters on the second dimension (a_{2pncl}) ranged from 0.091 to

0.216 across all combination of conditions. Item discrimination estimates on the primary dimension tended to have larger RMSEs when the item was cross-loaded than when it was non-cross loaded.

Secondary item discrimination parameters. The average RMSEs for the secondary item discrimination parameter for the first dimension (a_{1s}) ranged from 0.062 to 0.617. On the other hand, the average RMSEs for the secondary item discrimination parameter on the second dimension (a_{2s}) ranged from 0.061 to 0.619 across all combination of conditions. It was interesting to observe that secondary item discrimination parameters on first and second dimensions had a very similar patterns and values in terms of average RMSEs. However, compared to the RMSEs of corresponding items on the primary dimension (a_{1pcl} and a_{2pcl}), the RMSEs were larger on the secondary dimension than on the primary dimension.

Item location parameter (d). Table 4.3 reports the average RMSEs for the item location parameter (d) for both correct specified and misspecified models across all combination of conditions. The average RMSEs for item location parameter when the model was correctly specified ranged from 0.067 to 0.169. The lowest average RMSE for item location parameter was associated with the condition when the correlation between the abilities was .0 and the sample size was 2,000. The highest average RMSE for item location parameter was associated with the condition when the correlation between the abilities was .9 and the sample size was 500.

Misspecified Models

The RMSEs were averaged for the estimated item discrimination parameters (a_1 and a_2) and the estimated item location parameter (d) within each combination of conditions for three sets of item cross-loading. For the item discrimination parameters (a_1 and a_2) in misspecified models, the RMSEs for the *truly* primary cross-loaded item discriminations on first and second dimensions (a_{1pcl} and a_{2pcl}) were averaged based on those items that had a primary item discrimination and were supposed to be specified as cross-loaded items on both dimensions (these are the items that were misspecified in the model). Likewise, the RMSEs for the *truly* primary non-cross loaded item discriminations on first and second dimensions (a_{1pncl} and a_{2pncl}) were averaged based on the items that had a primary item discrimination but were not truly cross-loaded on both dimensions. It should be noted that there was no secondary item discrimination defined for the first and second dimension in the misspecified models (a_{1s} and a_{2s}).

Primary item discrimination parameters. Table 4.2 reports the average RMSEs of the estimated item discrimination parameters when the models were misspecified considering three levels of structure complexity magnitude incorporating low, medium or high degree of cross-loading. As shown in table 4.2, the average RMSEs for the truly primary cross-loaded item discrimination parameters on the first dimension (a_{1pcl}) ranged from 0.080 to 1.119. The average RMSEs for the truly primary non-cross loaded item discrimination parameters on the first dimension (a_{1pncl}) ranged from 0.083 to 0.262 across all combinations of conditions. On the other hand, the average RMSEs for the truly primary cross-loaded item discrimination parameters on the second dimension (a_{2pcl}) ranged from 0.075 to 1.144. The average RMSEs for the truly primary non-cross loaded item discrimination parameters on the second dimension (a_{2pncl}) ranged from 0.084 to 0.269 across all combinations of conditions. Item discrimination estimates on the truly primary dimension tended to have much larger RMSEs when the item was supposed to be specified as cross-loaded than when it was non-cross loaded. In addition, it should be noted that very similar patterns and values in terms of average RMSEs were observed for the truly primary item discrimination parameters on the first and second dimensions.

Item location parameter (d). Table 4.3 reports the average RMSEs for the item location parameter (d) for both correct specified and misspecified models across all combinations of conditions. The average RMSEs for item location parameter when the models were misspecified ranged from 0.067 to 0.165. The lowest average RMSE for item location parameter was associated with the condition when the correlation between the abilities was .0 and the sample size was 2,000. The highest average RMSE for item location parameter was associated with the condition when the correlation between the abilities was .9 and the sample size was 500.

Table 4. 1. Average RMSE of the primary (cross-loaded and non-cross loaded items) and secondary discrimination on the first and second dimensions when the models were correctly specified.

			LOW						MED						HGH					
			a_1			a_2			a_1			a_2			a_1			a_2		
Complexity	Correlation	N	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}
10%	.0	500	0.156	0.199	0.124	0.163	0.200	0.122	0.151	0.197	0.129	0.165	0.199	0.123	0.168	0.198	0.155	0.170	0.194	0.151
		1000	0.111	0.135	0.087	0.111	0.138	0.09	0.114	0.139	0.091	0.114	0.135	0.091	0.115	0.140	0.109	0.114	0.138	0.112
		2000	0.081	0.095	0.062	0.075	0.096	0.061	0.078	0.097	0.065	0.079	0.096	0.064	0.083	0.095	0.077	0.084	0.098	0.077
	.6	500	0.174	0.202	0.24	0.180	0.198	0.24	0.188	0.195	0.246	0.182	0.198	0.231	0.225	0.198	0.235	0.207	0.196	0.24
		1000	0.128	0.137	0.221	0.125	0.135	0.221	0.120	0.138	0.211	0.123	0.143	0.213	0.152	0.134	0.192	0.159	0.137	0.19
		2000	0.094	0.100	0.207	0.091	0.098	0.207	0.086	0.094	0.197	0.083	0.096	0.19	0.125	0.096	0.164	0.125	0.097	0.168
	.9	500	0.256	0.203	0.419	0.248	0.199	0.423	0.214	0.198	0.392	0.222	0.200	0.392	0.244	0.197	0.313	0.231	0.199	0.29
		1000	0.224	0.138	0.414	0.240	0.138	0.395	0.164	0.138	0.36	0.163	0.136	0.355	0.162	0.138	0.272	0.175	0.137	0.264
		2000	0.216	0.098	0.393	0.214	0.095	0.389	0.142	0.096	0.337	0.138	0.098	0.341	0.123	0.096	0.239	0.130	0.096	0.242
30%	.0	500	0.211	0.189	0.14	0.207	0.196	0.142	0.204	0.194	0.149	0.202	0.196	0.15	0.210	0.195	0.194	0.213	0.195	0.189
		1000	0.137	0.130	0.103	0.141	0.133	0.101	0.143	0.131	0.108	0.140	0.133	0.103	0.145	0.137	0.134	0.147	0.131	0.135
		2000	0.098	0.096	0.069	0.097	0.092	0.068	0.101	0.095	0.074	0.099	0.095	0.074	0.098	0.093	0.093	0.101	0.095	0.092
	.6	500	0.215	0.195	0.345	0.216	0.194	0.339	0.225	0.201	0.334	0.224	0.194	0.327	0.293	0.191	0.335	0.300	0.198	0.328
		1000	0.158	0.136	0.319	0.159	0.136	0.314	0.147	0.137	0.305	0.152	0.135	0.299	0.226	0.135	0.277	0.228	0.136	0.272
		2000	0.118	0.096	0.31	0.119	0.097	0.304	0.104	0.097	0.284	0.110	0.094	0.281	0.186	0.094	0.242	0.191	0.096	0.246
	.9	500	0.308	0.201	0.596	0.318	0.206	0.593	0.253	0.203	0.543	0.241	0.199	0.524	0.315	0.196	0.422	0.325	0.199	0.422
		1000	0.284	0.139	0.579	0.292	0.141	0.581	0.191	0.142	0.503	0.187	0.135	0.505	0.246	0.135	0.357	0.266	0.138	0.386
		2000	0.259	0.098	0.568	0.266	0.098	0.562	0.150	0.096	0.49	0.154	0.098	0.484	0.210	0.098	0.341	0.201	0.096	0.336
50%	.0	500	0.204	0.188	0.147	0.204	0.196	0.151	0.206	0.190	0.156	0.209	0.190	0.158	0.207	0.199	0.189	0.208	0.200	0.194
		1000	0.142	0.132	0.104	0.144	0.131	0.103	0.145	0.137	0.11	0.145	0.135	0.11	0.142	0.138	0.131	0.143	0.140	0.132
		2000	0.101	0.093	0.073	0.103	0.091	0.073	0.100	0.092	0.078	0.100	0.091	0.077	0.100	0.097	0.095	0.101	0.092	0.092
	.6	500	0.213	0.203	0.354	0.219	0.202	0.359	0.223	0.206	0.355	0.223	0.199	0.355	0.305	0.213	0.343	0.300	0.204	0.347
		1000	0.153	0.142	0.335	0.155	0.141	0.328	0.154	0.139	0.315	0.152	0.143	0.322	0.244	0.146	0.285	0.234	0.148	0.295
		2000	0.117	0.098	0.324	0.116	0.098	0.323	0.102	0.101	0.305	0.108	0.103	0.304	0.193	0.104	0.262	0.193	0.108	0.262

	.9	500	0.304	0.216	0.617	0.308	0.216	0.619	0.241	0.213	0.56	0.244	0.215	0.554	0.334	0.221	0.441	0.340	0.210	0.442
		1000	0.266	0.147	0.593	0.267	0.146	0.594	0.180	0.152	0.528	0.183	0.149	0.519	0.261	0.147	0.397	0.259	0.155	0.386
		2000	0.247	0.108	0.58	0.246	0.108	0.579	0.142	0.108	0.509	0.143	0.109	0.506	0.222	0.110	0.367	0.223	0.112	0.365

Table 4. 2. Average RMSE of the truly primary (cross-loaded and non-cross loaded items) on the first and second dimensions when the models were misspecified.

			LOW						MED						HGH					
			a_1			a_2			a_1			a_2			a_1			a_2		
Complexity	Correlation	N	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}
10%	.0	500	0.153	0.199	—	0.161	0.200	—	0.148	0.198	—	0.162	0.199	—	0.184	0.200	—	0.181	0.197	—
		1000	0.110	0.135	—	0.109	0.138	—	0.111	0.140	—	0.110	0.136	—	0.145	0.141	—	0.147	0.140	—
		2000	0.080	0.095	—	0.075	0.097	—	0.083	0.097	—	0.082	0.097	—	0.135	0.096	—	0.133	0.099	—
	.6	500	0.215	0.201	—	0.217	0.198	—	0.294	0.194	—	0.304	0.197	—	0.448	0.197	—	0.426	0.196	—
		1000	0.170	0.137	—	0.167	0.135	—	0.248	0.138	—	0.251	0.142	—	0.392	0.134	—	0.395	0.136	—
		2000	0.139	0.099	—	0.144	0.098	—	0.238	0.094	—	0.242	0.095	—	0.385	0.096	—	0.387	0.097	—
	.9	500	0.263	0.200	—	0.267	0.196	—	0.420	0.196	—	0.417	0.196	—	0.746	0.194	—	0.740	0.195	—
		1000	0.220	0.137	—	0.215	0.137	—	0.387	0.136	—	0.392	0.134	—	0.711	0.135	—	0.711	0.135	—
		2000	0.190	0.097	—	0.201	0.094	—	0.364	0.095	—	0.370	0.097	—	0.691	0.094	—	0.697	0.094	—
30%	.0	500	0.209	0.189	—	0.204	0.194	—	0.209	0.189	—	0.207	0.191	—	0.261	0.202	—	0.261	0.198	—
		1000	0.137	0.129	—	0.141	0.132	—	0.148	0.130	—	0.143	0.131	—	0.196	0.154	—	0.199	0.149	—
		2000	0.098	0.095	—	0.098	0.091	—	0.109	0.095	—	0.107	0.095	—	0.153	0.126	—	0.156	0.123	—
	.6	500	0.297	0.186	—	0.294	0.184	—	0.449	0.183	—	0.460	0.182	—	0.839	0.178	—	0.837	0.181	—
		1000	0.241	0.131	—	0.237	0.130	—	0.395	0.127	—	0.402	0.127	—	0.789	0.130	—	0.791	0.128	—
		2000	0.214	0.092	—	0.209	0.093	—	0.368	0.091	—	0.377	0.088	—	0.765	0.099	—	0.770	0.101	—
	.9	500	0.379	0.186	—	0.369	0.188	—	0.615	0.184	—	0.613	0.181	—	1.119	0.174	—	1.144	0.176	—
		1000	0.318	0.129	—	0.307	0.130	—	0.557	0.128	—	0.564	0.124	—	1.092	0.120	—	1.093	0.122	—

		2000	0.289	0.091	—	0.288	0.090	—	0.541	0.087	—	0.540	0.089	—	1.061	0.087	—	1.053	0.086	—
50%	.0	500	0.206	0.184	—	0.206	0.191	—	0.228	0.182	—	0.231	0.185	—	0.410	0.262	—	0.407	0.269	—
		1000	0.145	0.130	—	0.147	0.129	—	0.171	0.136	—	0.171	0.136	—	0.367	0.237	—	0.363	0.242	—
		2000	0.104	0.092	—	0.106	0.091	—	0.128	0.101	—	0.127	0.105	—	0.338	0.226	—	0.339	0.223	—
	.6	500	0.290	0.185	—	0.294	0.181	—	0.448	0.173	—	0.454	0.170	—	0.887	0.183	—	0.886	0.180	—
		1000	0.233	0.128	—	0.241	0.126	—	0.414	0.122	—	0.407	0.122	—	0.850	0.146	—	0.836	0.146	—
		2000	0.208	0.088	—	0.209	0.088	—	0.383	0.088	—	0.388	0.091	—	0.816	0.123	—	0.816	0.122	—
	.9	500	0.344	0.175	—	0.348	0.181	—	0.584	0.171	—	0.590	0.170	—	1.098	0.167	—	1.103	0.167	—
		1000	0.301	0.124	—	0.302	0.121	—	0.537	0.125	—	0.542	0.119	—	1.054	0.114	—	1.053	0.120	—
		2000	0.272	0.086	—	0.274	0.086	—	0.519	0.083	—	0.519	0.084	—	1.033	0.084	—	1.034	0.085	—

Table 4. 3. Average RMSEs of the item location parameter (*d*) for the correct and misspecified models.

			Correct Specified Models			Misspecified Models		
Complexity	Correlation	N	LOW	MED	HGH	LOW	MED	HGH
10%	.0	500	0.139	0.140	0.139	0.139	0.140	0.138
		1000	0.096	0.097	0.099	0.096	0.096	0.098
		2000	0.069	0.069	0.067	0.069	0.069	0.067
	.6	500	0.138	0.140	0.142	0.137	0.140	0.140
		1000	0.097	0.098	0.097	0.096	0.097	0.096
		2000	0.069	0.068	0.069	0.069	0.067	0.069
	.9	500	0.138	0.143	0.140	0.137	0.143	0.139
		1000	0.098	0.099	0.099	0.098	0.099	0.098
		2000	0.070	0.069	0.070	0.069	0.069	0.070
30%	.0	500	0.144	0.141	0.147	0.143	0.138	0.144
		1000	0.097	0.098	0.102	0.097	0.096	0.106
		2000	0.068	0.069	0.070	0.067	0.069	0.079

	.6	500	0.142	0.143	0.151	0.141	0.140	0.145
		1000	0.100	0.101	0.108	0.099	0.099	0.105
		2000	0.069	0.072	0.073	0.068	0.071	0.074
	.9	500	0.144	0.145	0.155	0.142	0.142	0.151
		1000	0.100	0.104	0.108	0.099	0.102	0.105
		2000	0.069	0.074	0.077	0.068	0.072	0.075
50%	.0	500	0.140	0.141	0.147	0.139	0.137	0.145
		1000	0.099	0.101	0.103	0.098	0.099	0.108
		2000	0.069	0.071	0.072	0.069	0.070	0.082
	.6	500	0.145	0.149	0.163	0.143	0.146	0.157
		1000	0.101	0.103	0.112	0.100	0.101	0.110
		2000	0.071	0.073	0.079	0.070	0.071	0.079
	.9	500	0.148	0.154	0.169	0.145	0.149	0.165
		1000	0.104	0.107	0.117	0.103	0.104	0.115
		2000	0.072	0.075	0.081	0.070	0.073	0.079

Effect of Sample Size on Item Parameter Recovery in Terms of Average RMSE

Correct Specified Models

Item discrimination parameters. Figures 4.1 to 4.3 show the average RMSEs for item discrimination parameters when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figures 4.1 to 4.3, all of the item discrimination parameters including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}\ a_{2pcl}, a_{2pncl}, a_{2s}$) had a consistent decreasing trend in terms of average RMSE as the sample size increased from 500 to 2,000 across all combinations of conditions. The lowest average RMSE of item discrimination for correct specified models (0.061) was associated with the sample size of 2,000 and the highest average RMSE of item discrimination for correct specified models (0.619) was associated with the sample size of 500.

Item location parameter (d). Figure 4.7 shows the average RMSEs for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figure 4.7, item location parameter had a consistent decreasing trend in terms of average RMSE as the sample size increased from 500 to 2,000 across all combinations of conditions. The lowest average RMSE of item location for correct specified models (0.67) was associated with the sample size of 2,000 and the highest average RMSE was associated with the sample size of 500 (0.169).

Misspecified Models

In misspecified models, the misspecification specifically occurs on the cross-loaded items. In misspecified models, it is assumed that none of the items are cross-loaded. Therefore, for the low, medium, and high degree of cross loading conditions, the cross loading is ignored on those items. Investigating these situations will provide a better understanding to real-world situations when a few or a lot of items within multidimensional data are cross-loaded, but the cross loading is unaccounted for.

Item discrimination parameters. Figures 4.4 to 4.6 show the average RMSEs for item discrimination parameters when models were misspecified with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figures 4.4 to 4.6, all of the primary item discrimination parameters on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{2pcl}, a_{2pncl}$) had a consistent decreasing trend in terms of average RMSE as the sample size increased from 500 to 2,000 across all combinations of conditions. The lowest average RMSE of item discrimination for misspecified models (0.75) was associated with the sample size of 2,000 and the highest average RMSE of item discrimination for misspecified models (1.144) was associated with the sample size of 500.

Item location parameter (d). Figure 4.8 shows the average RMSEs for item location parameter (d) when models were misspecified with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figure 4.7, item location parameter had a consistent decreasing trend in terms of average RMSE as the sample size increased from 500 to 2,000 across all combinations of conditions. The lowest average RMSE (value) was associated with the sample size of 2,000 and the highest average RMSE The lowest average RMSE of item location for misspecified models (0.067) was associated with the sample size of 2,000 and the highest average RMSE of item location for misspecified models (0.165) was associated with the sample size of 500.

Effect of Correlation on Item Parameter Recovery in Terms of RMSE

Correct Specified Models

Item discrimination parameters. Figures 4.1 to 4.3 show the average RMSE trends considering the effect of the correlation between abilities on the precision of item discrimination parameters including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}$ $a_{2pcl}, a_{2pncl}, a_{2s}$). When the sample size and complexity magnitude were held constant for each section while the correlation between abilities varied across the conditions, the average RMSEs increased consistently for the primary and secondary item discriminations on both dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}\ a_{2pcl}, a_{2pncl}, a_{2s}$) as the correlation increased from .0 to .9 across combination of conditions.

It was interesting that this increasing trend was more obvious for the secondary item parameter discriminations on both dimensions (a_{1s} and a_{2s}) compared to the primary item discrimination parameters ($a_{1pcl}, a_{1pncl}, a_{2pcl}, a_{2pncl}$). When the true correlation was zero, estimated item discrimination on the secondary dimension was estimated with the smallest RMSE, and item discrimination on the primary dimension (both cross-loading items and non cross-loading items) was slightly larger. On the primary item discrimination parameters, the RMSE values for the cross-loaded items (a_{1pcl}, a_{2pcl}) tended to increase at a higher rate as correlation increased than for the RMSE of the non-cross loaded items ($a_{1pncl},\ a_{2pncl}$).

Correlation had very little to no effect on the RMSE of estimated item discrimination for the primary non-cross loaded items; as correlation increased, the RMSE of the estimated item discrimination for primary cross-loaded items tended to decrease (a slight increase in RMSE when correlation increased from .0 to .6 and a larger increase when RMSE increased from .6 to .9). As correlation increased, the RMSE of the estimated item discrimination on the secondary dimension increased substantially. This may be due to the lack of freely estimated correlation in the model specification when calibrating the simulated data; all models assumed a correlation of .0.

Item location parameter (d). Figure 4.7 shows the average RMSEs for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figure 4.7, item location parameter had a consistent constant trend in terms of average RMSE as the correlation between abilities increased from .0 to .9 across all combinations of conditions.

Misspecified Models

Item discrimination parameters. Figures 4.4 to 4.6 show the average RMSE trends considering the effect of the correlation between abilities on the precision of item discrimination parameters on first and second dimensions ($a_{1pcl}, a_{1pncl}, \ a_{2pcl}, a_{2pncl}$). When the sample size and complexity magnitude were held constant for each section while the correlation between abilities varied across the conditions, the average RMSEs of the item discrimination parameter for the truly cross-loaded item (i.e., the model did not account for these cross loadings) on first and second dimensions ($a_{1pcl}, \ a_{2pcl}$) increased consistently as the correlation increased from .0 to .9 across combinations of conditions. This increase was much greater when the complexity magnitude was highest (i.e., 50%) and the degree of cross-loading was highest. On the other hand, the average RMSEs of item discrimination parameters for the truly non-cross loaded items ($a_{1pncl}, \ a_{2pncl}$) decreased slightly as the correlation increased from .0 to .9 across combinations of conditions. This decreasing trend was more obvious when the complexity magnitude was highest at 50% and the degree of cross-loading was highest. The RMSE values for the truly cross-loaded items (i.e., those that were misspecified, a_{1pcl}, a_{2pcl}) were similar to the RMSE values on the non-cross loaded items ($a_{1pncl}, \ a_{2pncl}$) when correlation was zero, but increased to be greater than the discrimination on the non-cross loaded items as the correlation increased from .0 to .9 across combination of conditions.

Item location parameter (d). Figure 4.8 shows the average RMSEs for item location parameter (d) when models were misspecified with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figure 4.8, item location parameter had a constant trend in terms of average RMSE as the correlation between abilities increased from .0 to .9 across all combinations of conditions.

Effect of Structure Complexity Magnitude on Item Parameter Recovery in Terms of RMSE

Correct Specified Models

Item discrimination parameters. When holding the correlation and sample size constant, figures 4.1 to 4.3 represent the average RMSE trends considering the effect of the three levels of complexity magnitude on the precision of item discrimination parameter estimates including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}\ a_{2pcl}, a_{2pncl}, a_{2s}$). As shown in figures 4.1 to 4.3, the average RMSEs increased consistently for the primary and secondary item discriminations on both dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}\ a_{2pcl}, a_{2pncl}, a_{2s}$) as the structure complexity magnitude increased from 10% to 30% to 50%. This increase was much greater for the secondary items parameter discriminations on both dimensions (a_{1s} and a_{2s}) compared to the primary item discrimination parameters ($a_{1pcl}, a_{1pncl}, a_{2pcl}, a_{2pncl}$) especially when the degree of cross-loading was low or medium.

The RMSE values of the primary item discrimination parameters for the cross-loaded items (a_{1pcl}, a_{2pcl}) were greater than the values on the non-cross loaded items ($a_{1pncl},\ a_{2pncl}$) as the complexity magnitude increased from 10% to 50% across combinations of conditions when correlation was greater than zero. The RMSE values for the discrimination of the non-cross loaded items on both dimensions ($a_{1pncl},\ a_{2pncl}$) had a constant trend with slightly increasing pattern as the complexity magnitude increased from 10% to 50%.

Item location parameter (d). Figure 4.7 shows the average RMSEs for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. Figure 4.7 shows as the complexity magnitude increased from 10% to 50%, the item location parameter had a constant trend with negligible changes in terms of average RMSE across all combinations of conditions.

Misspecified Models

Item discrimination parameters. Figures 4.4 to 4.6 show the average RMSE trends considering the effect of structure complexity magnitude incorporating the degree of cross loading for primary item discrimination parameters ($a_{1pcl}, a_{1pncl}, a_{2pcl}\ \ a_{2pncl}$) when the models were misspecified. When the sample size and the correlation between abilities were held constant and the complexity magnitude varied across the conditions, the average RMSEs for the truly cross-loaded item discrimination parameters on first and second dimensions (i.e., those that were misspecified, $a_{1pcl},\ a_{2pcl}$) increased consistently across combinations of conditions as the complexity magnitude increased from 10% to 50% for the high degrees of cross-loading. When there was a low or medium degree of cross-loading, RMSE of the truly cross-loaded item discrimination increased when complexity increased from 10% to 30%, but had little to no change when complexity increased from 30% to 50%. On the other hand, the average RMSEs for the item discrimination parameters for non-cross loaded items (i.e., no misspecification, a_{1pncl}, a_{2pncl}) decreased slightly as the complexity magnitude decreased from 10% to 50% across combinations of conditions. When correlation was zero, the RMSE of estimated item discrimination was slightly higher for non-cross loaded items ($a_{1pncl},\ a_{2pncl}$) than for truly cross-loaded (a_{1pcl}, a_{2pcl}); however, as correlation increased, the RMSE of the truly cross-loaded item discrimination was higher than for the non-cross loaded items.

Item location parameter (d). Figure 4.8 shows the average RMSEs for item location parameter (d) when models were misspecified across combinations of conditions. Figure 4.8 shows as the complexity magnitude increased from 10% to 50% item location parameter had a constant trend in terms of average RMSE across all combinations of conditions.

Effect of Degree of Cross-Loading on Item Parameter Recovery in Terms of RMSE

Correct Specified Models

Item discrimination parameters. By holding the complexity magnitude level, correlation and sample size constant, figures 4.1 to 4.3 show the average RMSE trends considering the effect of the degree of cross-loading on the precision of item discrimination parameters including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}$ $a_{2pcl}, a_{2pncl}, a_{2s}$). As shown in figures 4.1 to 4.3, the changes in the degree of cross loading from low to medium to high had very little effect on the average RMSEs of item discrimination on non-cross loaded items on both dimensions ($a_{1pncl},\ a_{2pncl}$). On the other hand, the average RMSEs of the item discrimination on cross-loaded items on both dimensions ($a_{1pcl},\ a_{2pcl}$) decreased slightly as the degree of cross-loading increased from low to medium, and remained constant as the degree of cross-loading increased to high.

The RMSEs for the item discrimination parameters were larger for the cross-loaded items compared to non-cross loaded items only when correlation was .9 for the low and medium degrees of cross loading, and only when correlation was .6 or higher for the high degree of cross loading. For the secondary item discrimination parameters (a_{1s} and a_{2s}) the average RMSEs was not affected by changes in the dgree of cross-loading when correlation was .0 and .6; when correlation was .9, the RMSE of the secondary item discrimination parameters had a decreasing trend on both dimensions.

Item location parameter (d). Figure 4.7 shows the average RMSEs for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combination of conditions. As shown in figure 4.7, the average RMSEs for item location parameter had a constant trend with a very slight increase as the degree of cross-loading shifted from low to high.

Misspecified Models

Item discrimination parameters. Figures 4.4 to 4.6 show the average RMSE trends considering the effect of the degree of cross-loading on the precision of item discrimination parameters on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{2pcl}, a_{2pncl}$) when the models were misspecified. By holding the complexity magnitude level, correlation and sample size constant, the average RMSEs for the item discrimination on non-cross loaded items on both dimensions ($a_{1pncl},\ a_{2pncl}$) remained constant for low, medium and high degrees of cross-loading. On the other hand, the average RMSEs for the item discrimination on cross-loaded items on both dimensions (i.e., those that were misspecified, $a_{1pcl},\ a_{2pcl}$) had a large increasing pattern across the combinations of conditions as the degree of cross-loading increased from low to high.

Item location parameter (d). Figure 4.8 shows the average RMSEs for item location parameter (d) when models were misspecified with a low, medium or high degree of cross-loading across combination of conditions. As shown in figure 4.8, the average RMSEs for item location parameter had a constant trend with a very slight increase as the degree of cross-loading shifted from low to high.

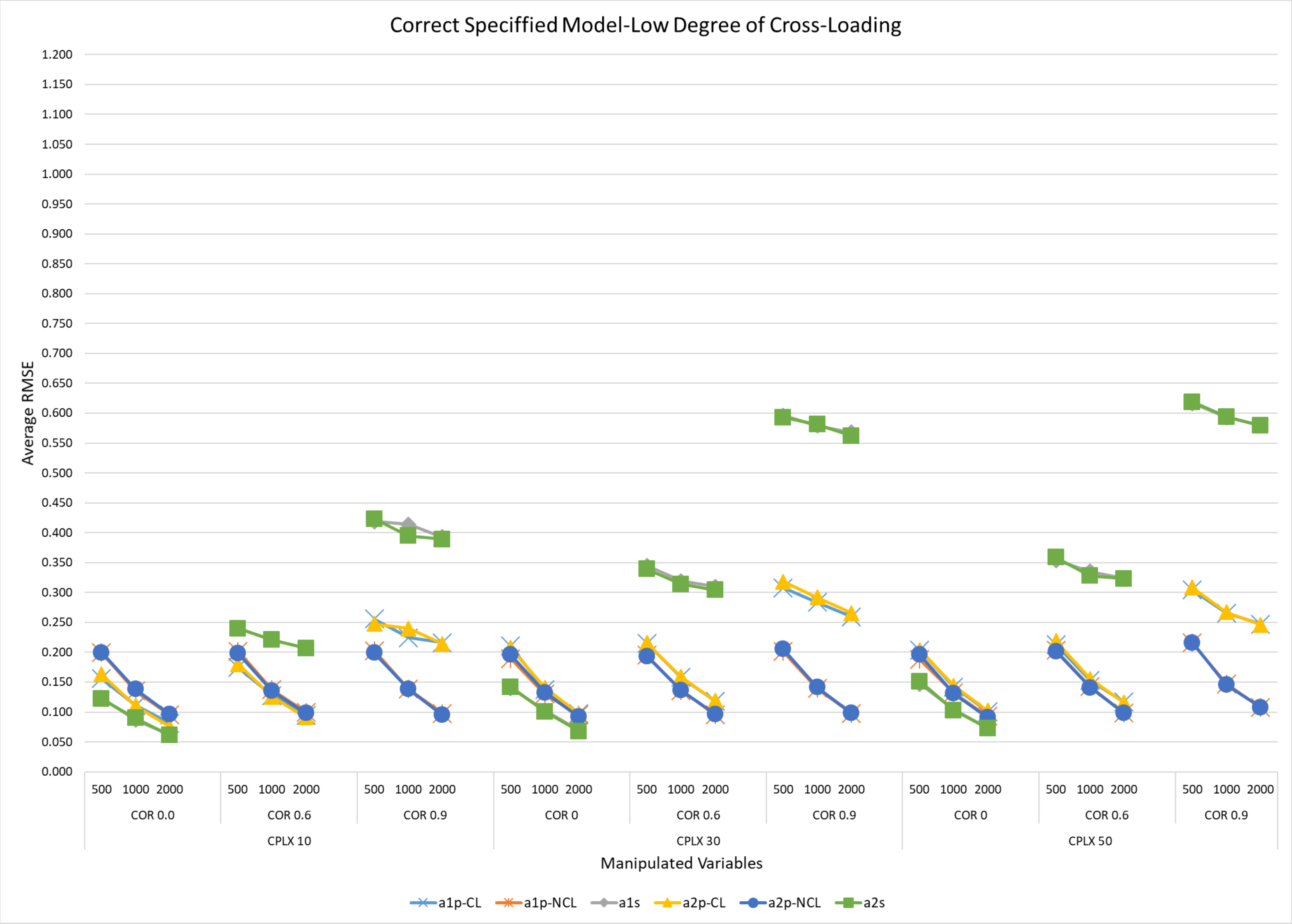

Figure 4. 1. Average RMSEs for item discrimination parameters when models were specified correctly with a low degree of cross-loading

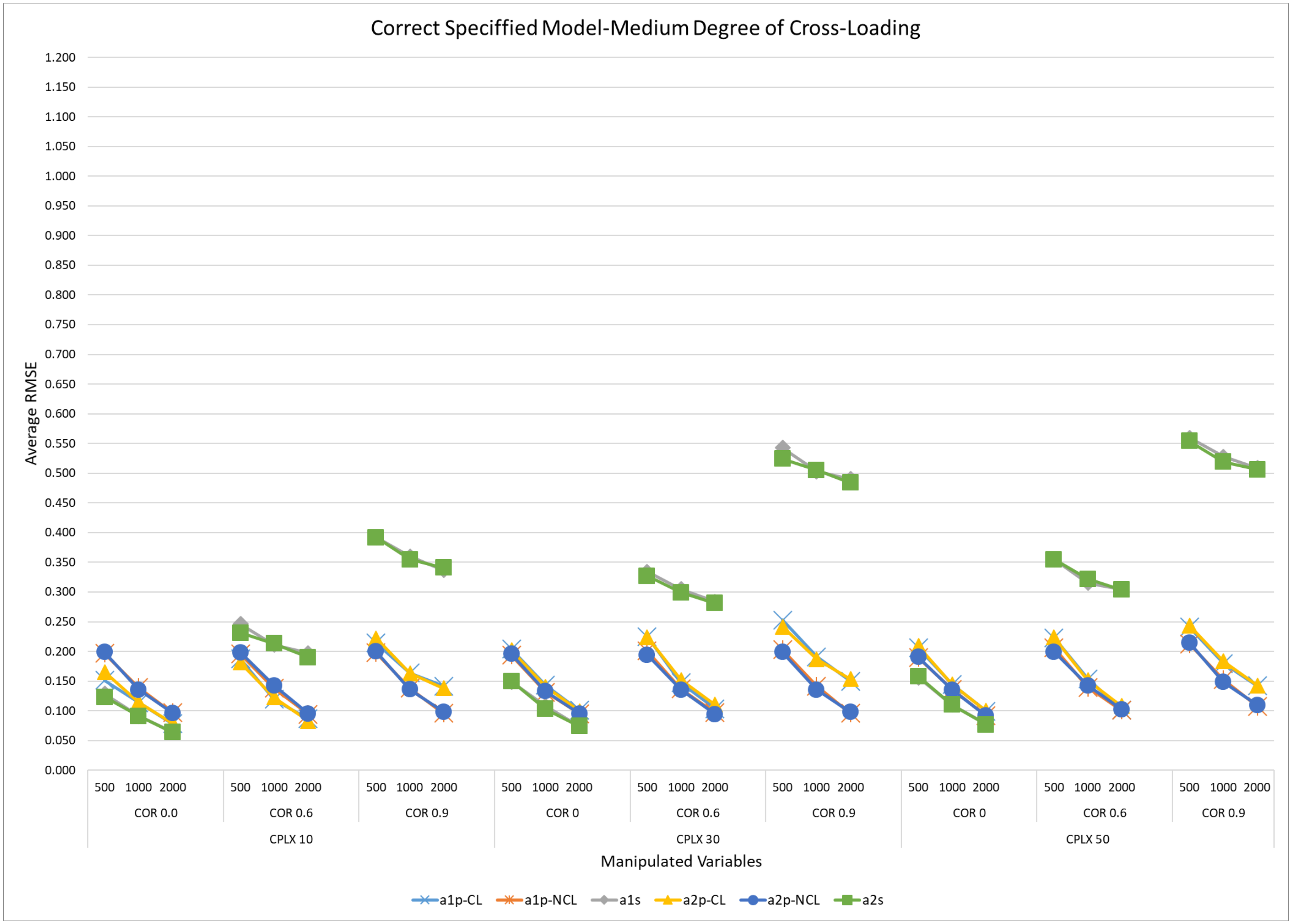

Figure 4. 2. Average RMSEs for item discrimination parameters when models were specified correctly with a medium degree of cross-loading

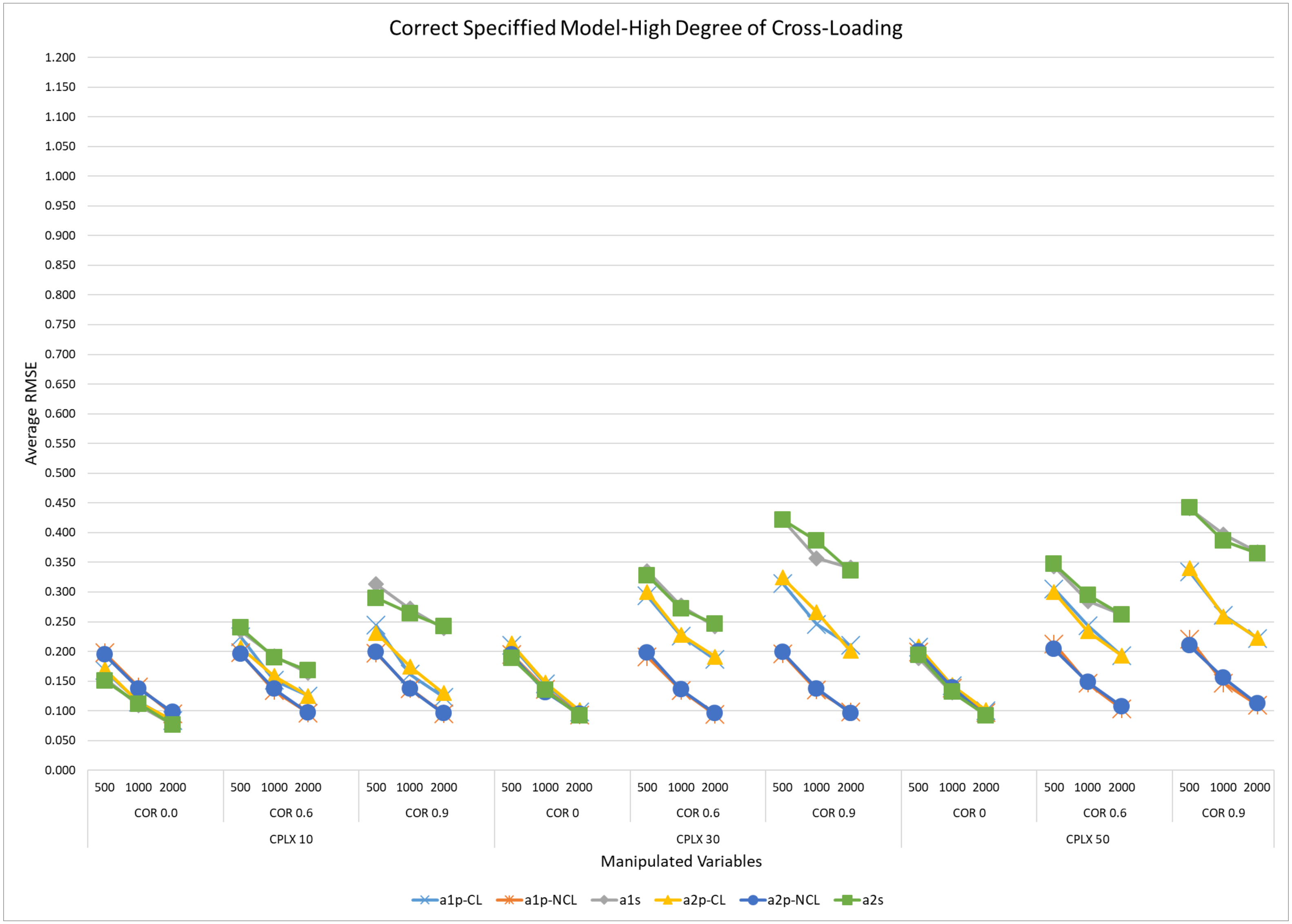

Figure 4. 3. Average RMSEs for item discrimination parameters when models were specified correctly with a high degree of cross-loading

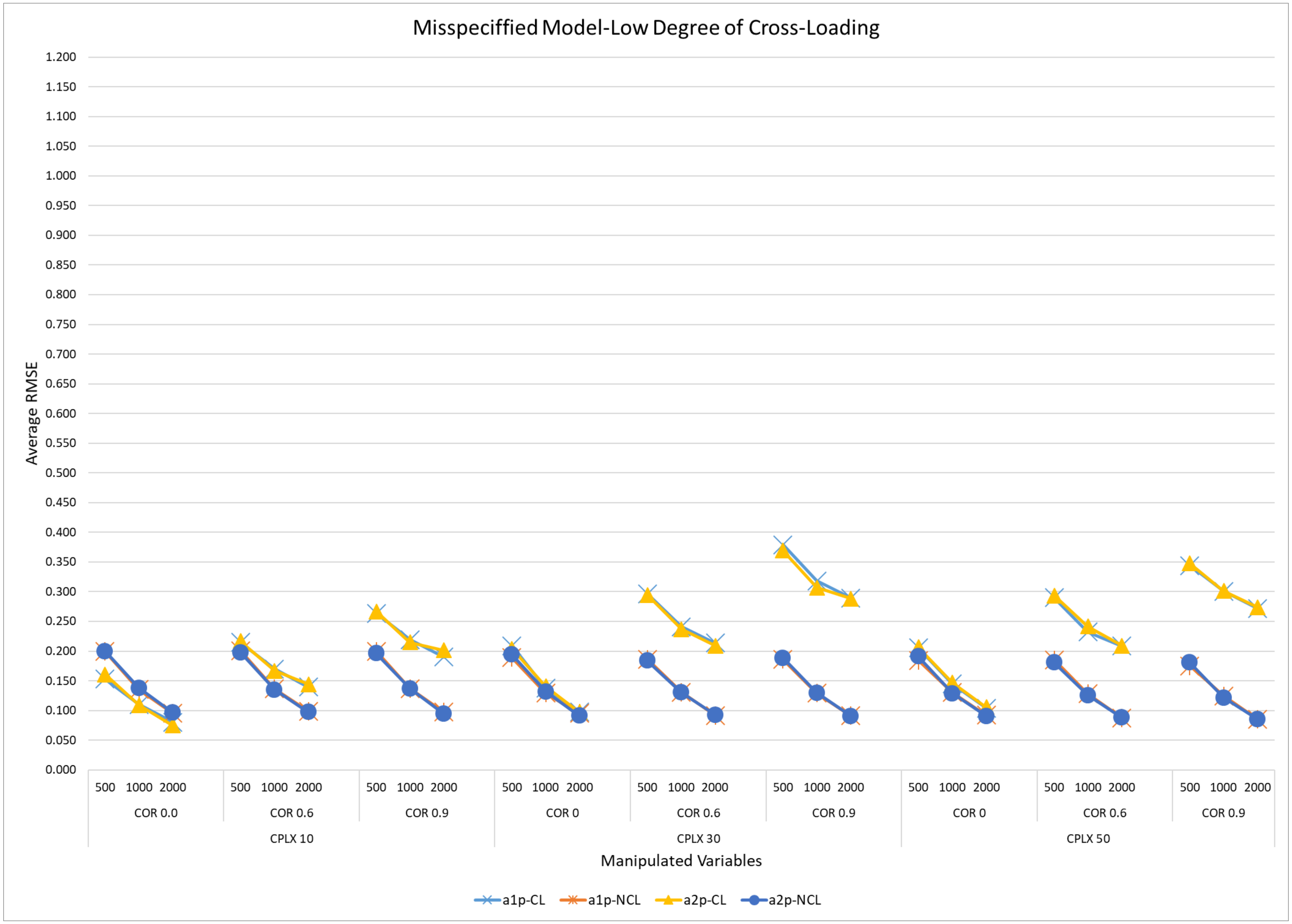

Figure 4. 4. Average RMSEs for item discrimination parameters when models were misspecified with a low degree of cross-loading

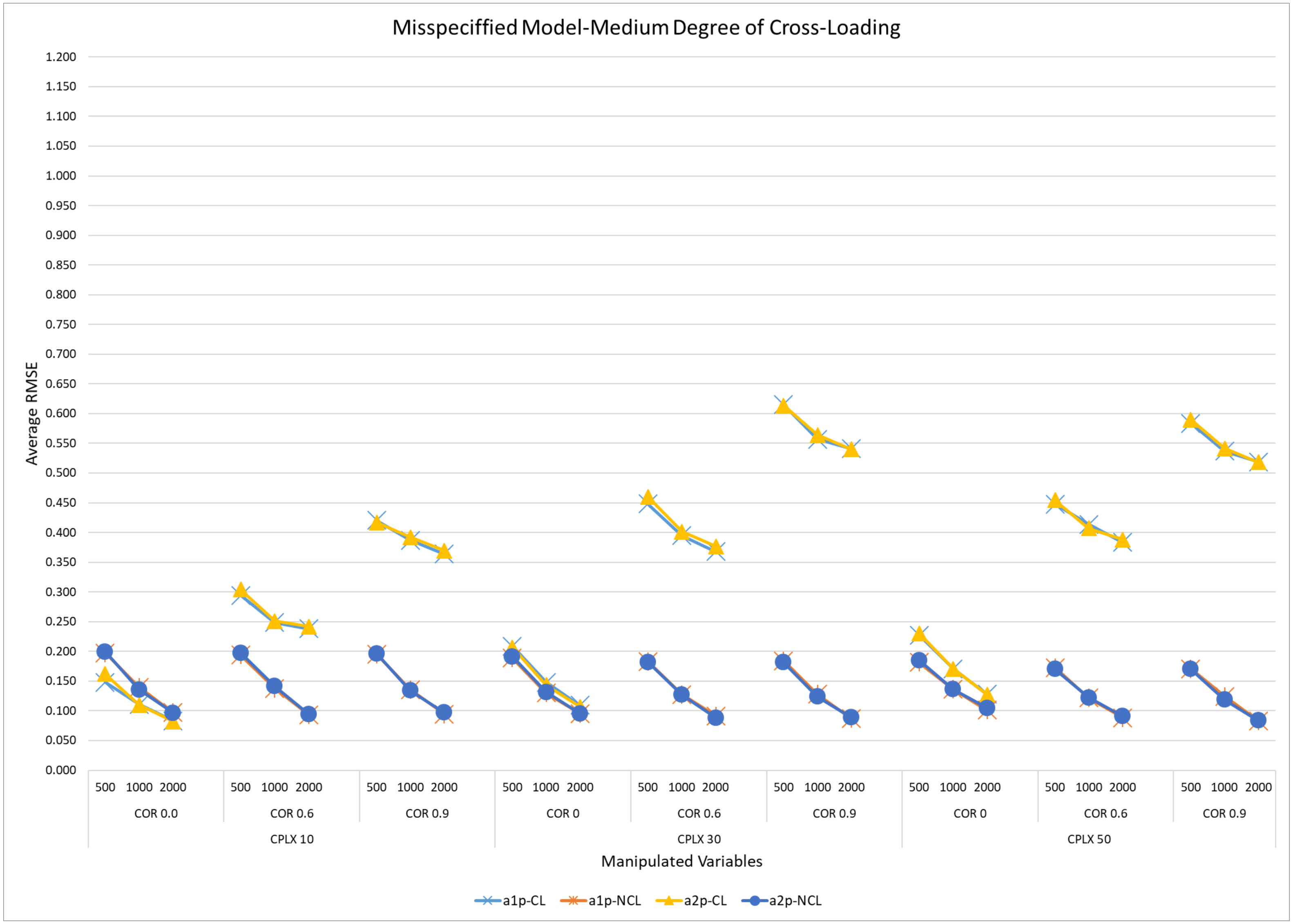

Figure 4. 5. Average RMSEs for item discrimination parameters when models were misspecified with a medium degree of cross-loading

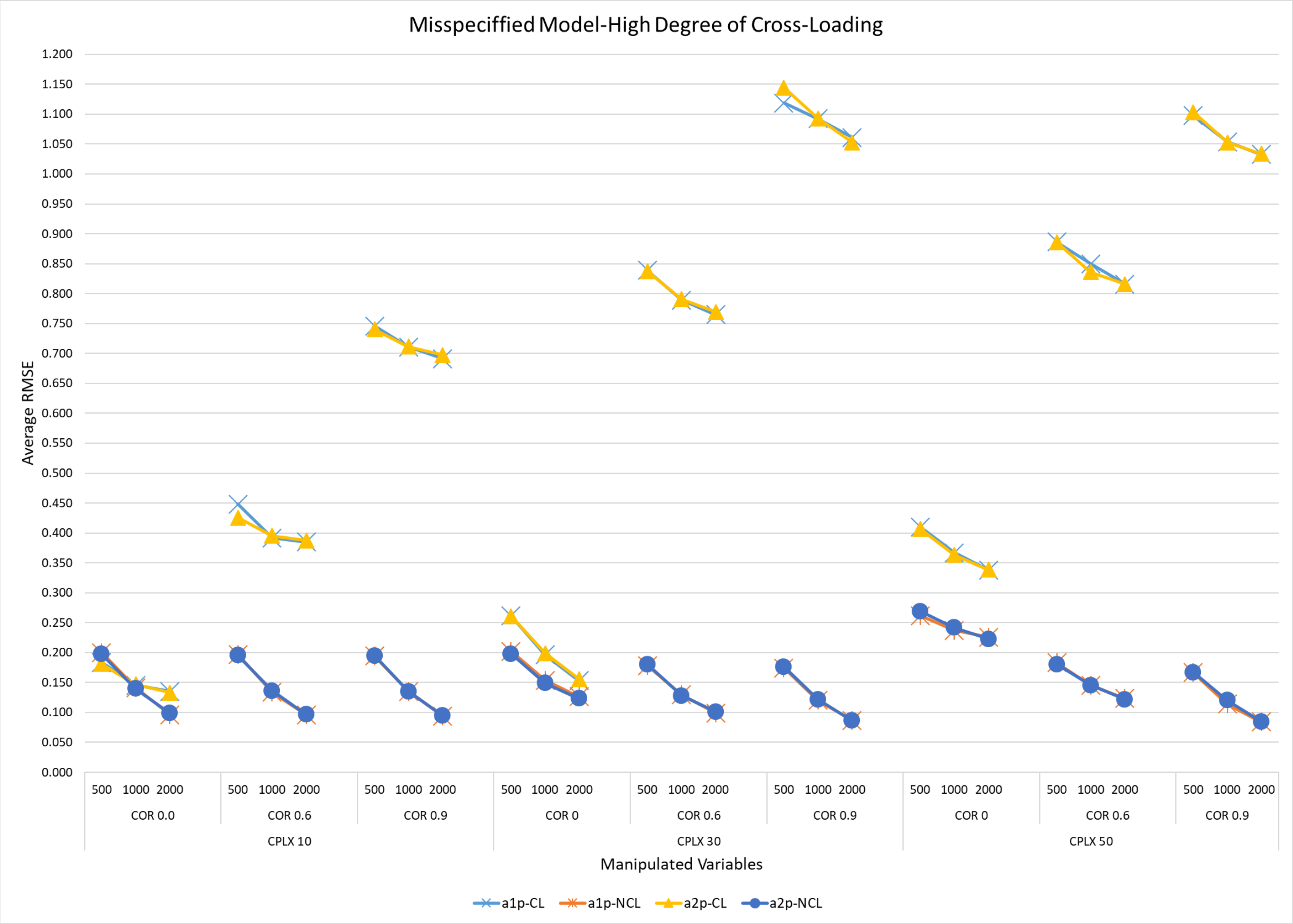

Figure 4. 6. Average RMSEs for item discrimination parameters when models were misspecified with a high degree of cross-loading

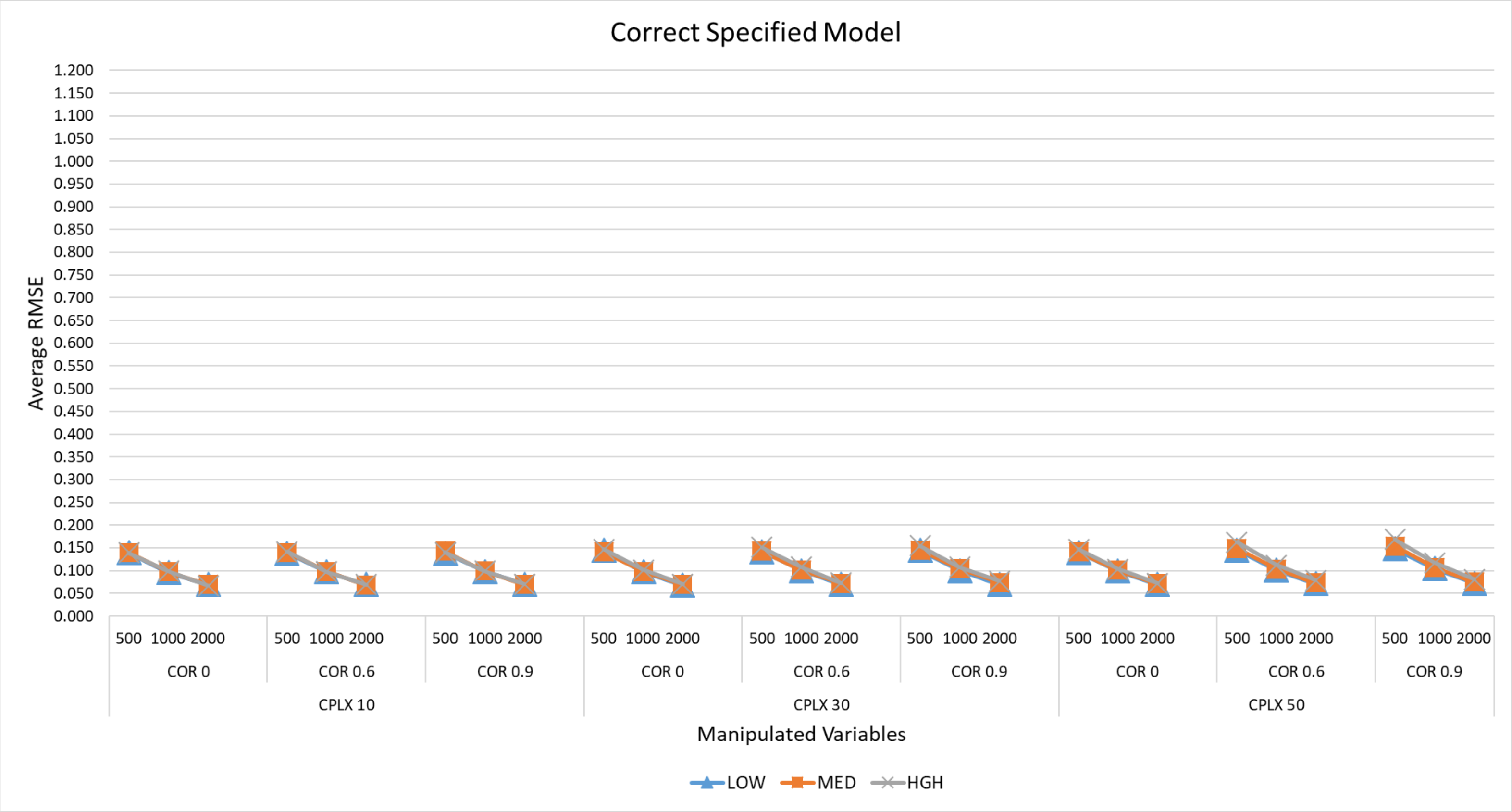

Figure 4. 7. Average RMSEs for item location parameter when models were specified correctly.

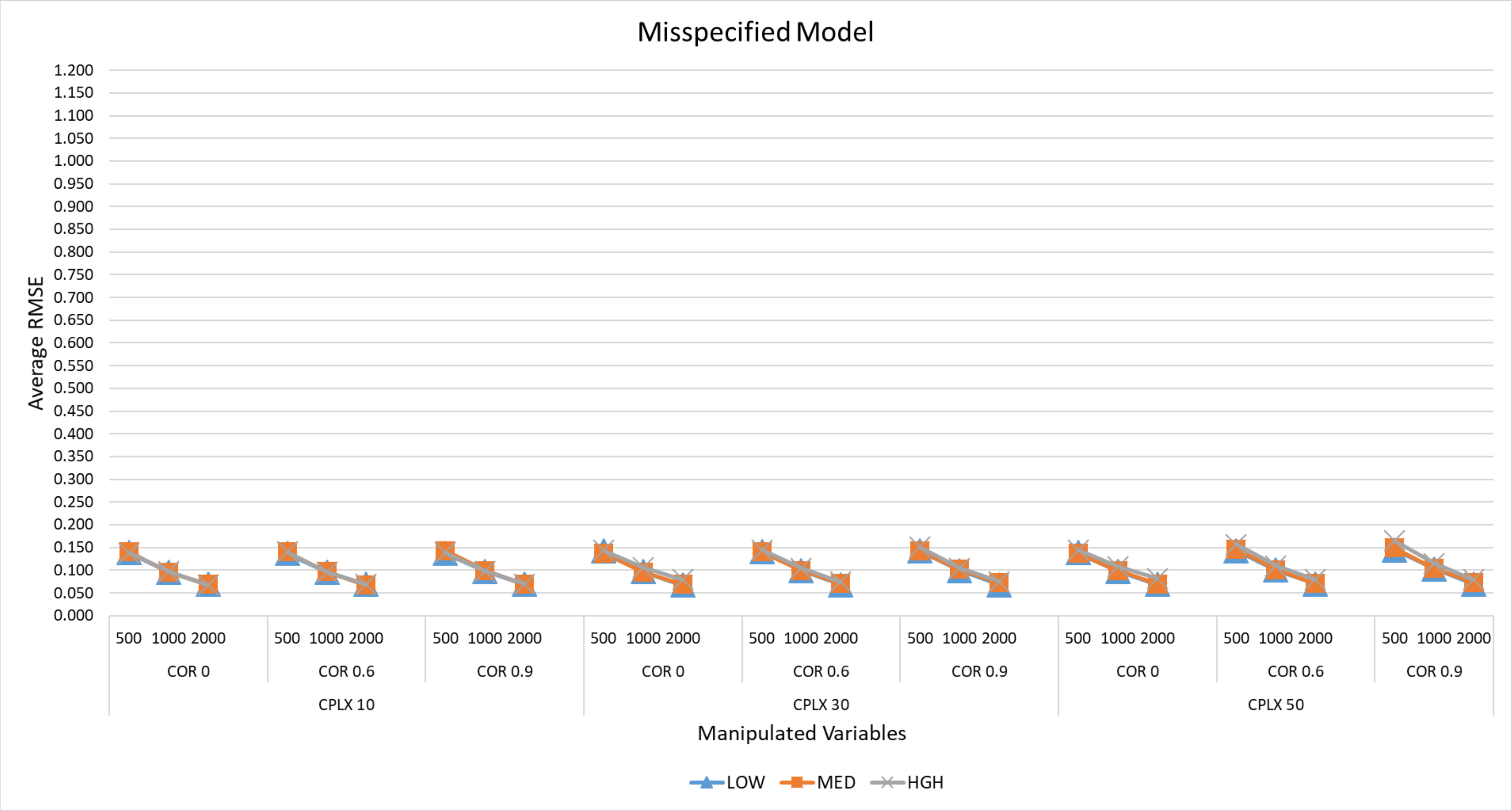

Figure 4. 8. Average RMSEs for item location parameter when models were misspecified.

Section II: Item Parameter Recovery Results in Terms of Average Bias

Correct Specified Models

Average bias in each combination of conditions was calculated comparing the true item parameter specification with the estimated item parameter. The biases were averaged for the estimated item discrimination parameters (a_1 and a_2) and the estimated item location parameter (d) within each combination of conditions. For the item discrimination parameters (a_1 and a_2) in correct specified models, the bias for the primary cross-loaded item discriminations on first and second dimensions (a_{1pcl} and a_{2pcl}) were averaged based on those items that had a primary item discrimination and were cross-loaded on both dimensions. Likewise, the biases for the primary non-cross loaded item discriminations on first and second dimensions (a_{1pncl} and a_{2pncl}) were averaged based on the items that had a primary item discrimination but were not cross-loaded on both dimensions. For the secondary item discriminations (a_{1s} and a_{2s}) the biases were averaged based on the items that had a secondary item discrimination.

Primary item discrimination parameters. Table 4.4 reports the average bias of the estimated item discrimination parameters when the models were correctly specified considering three levels of structure complexity magnitude incorporating low, medium or high degree of cross-loading. The average bias for the primary cross-loaded item discrimination parameters on the first dimension (a_{1pcl}) ranged from -0.206 to 0.242. The average bias for the primary non-cross loaded item discrimination parameters on the first dimension (a_{1pncl}) ranged from -0.071 to 0.001 across all combination of conditions. On the other hand, the average bias for the primary cross-loaded item discrimination parameters on the second dimension (a_{2pcl}) ranged from -0.212 to 0.253. The average bias for the primary non-cross loaded item discrimination parameters on the second dimension (a_{2pncl}) ranged from -0.060 to 0.003 across all combination of conditions.

Secondary item discrimination parameters. The average bias for the secondary item discrimination parameter for the first dimension (a_{1s}) ranged from -0.583 to 0.006. On the other hand, the average bias for the secondary item discrimination parameter on the second dimension (a_{2s}) ranged from -0.586 to 0.007 across all combinations of conditions. It was interesting to observe that secondary item discrimination parameters on first and second dimensions had a very similar patterns and values in terms of average bias.

Item location parameter (d). Table 4.6 reports the average bias for the item location parameter (d) for both correct specified and misspecified models across all combinations of conditions. The average bias for item location parameter when the model was correctly specified ranged from -0.157 to 0.244.

Misspecified Models

The biases were averaged for the estimated item discrimination parameters (a_1 and a_2) and the estimated item location parameter (d) within each combination of conditions for three sets of item cross-loading. For the item discrimination parameters (a_1 and a_2) in misspecified models, the biases for the *truly* primary cross-loaded item discriminations on the first and second dimensions (a_{1pcl} and a_{2pcl}) were averaged based on those items that had a primary item discrimination and were supposed to be specified as cross-loaded items on both dimensions (these are the items that were misspecified in the model). Likewise, the biases for the *truly* primary non-cross loaded item discriminations on the first and second dimensions (a_{1pncl} and a_{2pncl}) were averaged based on the items that had a primary item discrimination but were not truly cross-loaded on both dimensions. It should be noted that there was no secondary item discrimination defined for the first and second dimension in the misspecified models (a_{1s} and a_{2s}).

Primary item discrimination Parameters. Table 4.5 reports the average bias of the estimated item discrimination parameters when the models were misspecified considering three levels of structure complexity magnitude incorporating low, medium or high degree of cross-loading. As shown in table 4.5, the average bias for the truly primary cross-loaded item discrimination parameters on the first dimension (a_{1pcl}) ranged from -1.062 to 0.114. The average bias for the truly primary non-cross loaded item discrimination parameters on the first dimension (a_{1pncl}) ranged from -0.021 to 0.211 across all combinations of conditions. On the other hand, the average bias for the truly primary cross-loaded item discrimination parameters on the second dimension (a_{2pcl}) ranged from -1.080 to 0.112. The average bias for the truly primary non-cross loaded item discrimination parameters on the second dimension (a_{2pncl}) ranged from -0.018 to 0.212 across all combinations of conditions. In conclusion, It should be noted that item discrimination estimates tended to have generally negative values when the item was supposed to be specified as cross-loaded (a_{1pcl}, a_{2pcl}) suggesting that estimated primary item discrimination parameters on the *truly* cross-loaded items were somewhat smaller than their true values. On the other hand, item discrimination estimates on the primary dimension tended to have positive values generally closer to 0.000 when the item was supposed to be specified as non-cross loaded (a_{1pncl}, a_{2pncl}) suggesting that estimated primary item discrimination parameters on the *truly* non-cross loaded items were somewhat close to or slightly greater than their true values.

Item location parameter (d). Table 4.6 reports the average bias for the item location parameter (d) for both correct specified and misspecified models across all combinations of conditions. The average bias for item location parameter when the models were misspecified ranged from -0.011 to 0.010.

Table 4. 4. Average bias of the primary (cross-loaded and non-cross loaded items) and secondary discrimination on the first and second dimensions when the models were correctly specified.

			LOW						MED						HGH					
			a_1			a_2			a_1			a_2			a_1			a_2		
Complexity	Correlation	N	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}
10%	.0	500	-0.010	-0.016	-0.010	-0.013	-0.014	0.007	-0.011	-0.015	0.006	-0.011	-0.011	0.004	-0.014	-0.014	-0.003	-0.013	-0.013	-0.016
		1000	0.002	-0.009	0.002	-0.013	-0.005	-0.002	-0.014	-0.008	<0.001	-0.013	-0.007	-0.007	-0.009	-0.007	-0.003	-0.007	-0.008	-0.006
		2000	-0.004	-0.002	-0.002	0.001	-0.002	<0.001	-0.001	-0.001	0.001	-0.002	-0.001	-0.005	-0.002	0.001	-0.001	-0.004	-0.004	-0.004
	.6	500	0.038	-0.016	-0.195	0.034	-0.013	-0.198	-0.005	-0.009	-0.186	-0.019	-0.018	-0.175	-0.106	-0.022	-0.151	-0.089	-0.014	-0.163
		1000	0.046	-0.009	-0.195	0.047	-0.003	-0.199	0.004	-0.008	-0.182	0.005	-0.006	-0.184	-0.084	-0.005	-0.145	-0.084	-0.005	-0.145
		2000	0.051	-0.003	-0.194	0.045	-0.005	-0.194	<0.001	-0.005	-0.181	-0.003	-0.002	-0.174	-0.081	-0.003	-0.139	-0.083	0.000	-0.143
	.9	500	0.181	-0.014	-0.386	0.170	-0.013	-0.388	0.093	-0.014	-0.347	0.101	-0.019	-0.346	-0.087	-0.013	-0.233	-0.080	-0.019	-0.213
		1000	0.188	-0.006	-0.396	0.207	-0.006	-0.377	0.098	-0.006	-0.334	0.097	-0.005	-0.330	-0.079	-0.008	-0.224	-0.074	-0.007	-0.221
		2000	0.201	-0.003	-0.382	0.194	-0.002	-0.378	0.110	-0.001	-0.324	0.103	-0.003	-0.328	-0.067	-0.005	-0.213	-0.074	-0.004	-0.217
30%	.0	500	-0.022	-0.013	-0.001	-0.017	-0.019	-0.001	-0.014	-0.016	-0.006	-0.018	-0.012	-0.002	-0.019	-0.009	-0.009	-0.022	-0.018	-0.015
		1000	-0.006	-0.006	0.002	-0.006	-0.004	-0.007	-0.010	-0.005	<0.001	-0.006	-0.004	-0.004	-0.010	-0.009	-0.014	-0.011	-0.008	-0.006
		2000	-0.002	-0.001	-0.001	-0.004	-0.002	0.001	-0.004	-0.002	-0.001	-0.005	-0.003	-0.003	0.001	0.001	-0.005	-0.001	-0.005	-0.003
	.6	500	0.042	-0.018	-0.306	0.053	-0.021	-0.296	-0.020	-0.029	-0.282	-0.029	-0.023	-0.272	-0.169	-0.024	-0.238	-0.165	-0.030	-0.228
		1000	0.053	-0.013	-0.297	0.060	-0.010	-0.291	-0.008	-0.019	-0.277	-0.007	-0.017	-0.272	-0.147	-0.024	-0.225	-0.151	-0.023	-0.216
		2000	0.059	-0.010	-0.299	0.065	-0.011	-0.293	0.003	-0.014	-0.268	-0.004	-0.014	-0.265	-0.142	-0.017	-0.215	-0.147	-0.014	-0.218
	.9	500	0.220	-0.029	-0.564	0.234	-0.035	-0.559	0.081	-0.027	-0.495	0.083	-0.030	-0.478	-0.172	-0.033	-0.331	-0.186	-0.029	-0.333
		1000	0.242	-0.023	-0.562	0.253	-0.021	-0.564	0.110	-0.025	-0.479	0.100	-0.022	-0.482	-0.158	-0.026	-0.307	-0.184	-0.029	-0.336
		2000	0.240	-0.015	-0.558	0.246	-0.016	-0.553	0.103	-0.018	-0.477	0.107	-0.017	-0.472	-0.161	-0.021	-0.315	-0.152	-0.021	-0.310
50%	.0	500	-0.021	-0.011	-0.004	-0.021	-0.017	0.002	-0.015	-0.016	-0.006	-0.015	-0.010	-0.006	-0.020	-0.014	-0.021	-0.008	-0.013	-0.013
		1000	-0.007	-0.005	-0.001	-0.012	-0.005	-0.003	-0.011	-0.009	-0.003	-0.011	-0.008	0.001	-0.010	-0.010	-0.003	-0.007	-0.003	-0.011
		2000	-0.008	-0.001	-0.002	-0.002	0.003	0.001	-0.004	-0.005	-0.003	-0.002	<0.001	<0.001	-0.003	-0.002	-0.002	-0.003	-0.004	<0.001
	.6	500	0.041	-0.036	-0.311	0.042	-0.033	-0.316	-0.016	-0.046	-0.302	-0.022	-0.042	-0.301	-0.182	-0.060	-0.255	-0.180	-0.052	-0.255
		1000	0.054	-0.027	-0.313	0.047	-0.026	-0.305	-0.016	-0.034	-0.288	-0.012	-0.038	-0.295	-0.168	-0.044	-0.237	-0.161	-0.044	-0.246
		2000	0.057	-0.022	-0.312	0.056	-0.021	-0.312	-0.006	-0.033	-0.290	-0.007	-0.031	-0.288	-0.153	-0.037	-0.236	-0.153	-0.041	-0.235

	.9	500	0.221	-0.056	-0.583	0.217	-0.048	-0.586	0.070	-0.058	-0.515	0.075	-0.056	-0.505	-0.206	-0.071	-0.355	-0.212	-0.060	-0.359
		1000	0.223	-0.043	-0.575	0.223	-0.045	-0.576	0.086	-0.048	-0.505	0.089	-0.050	-0.496	-0.193	-0.053	-0.353	-0.181	-0.057	-0.343
		2000	0.227	-0.042	-0.570	0.225	-0.041	-0.570	0.092	-0.044	-0.496	0.093	-0.046	-0.494	-0.180	-0.049	-0.343	-0.181	-0.051	-0.342

Table 4. 5. Average bias of the truly primary (cross-loaded and non-cross loaded items) on the first and second dimensions when the models were misspecified.

			LOW						MED						HGH					
			a_1			a_2			a_1			a_2			a_1			a_2		
Complexity	Correlation	N	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}
10%	.0	500	0.001	-0.016	—	0.000	-0.014	—	0.024	-0.015	—	0.025	-0.012	—	0.108	-0.014	—	0.106	-0.014	—
		1000	0.012	-0.009	—	-0.003	-0.005	—	0.023	-0.008	—	0.022	-0.007	—	0.107	-0.008	—	0.108	-0.009	—
		2000	0.005	-0.002	—	0.010	-0.002	—	0.034	-0.001	—	0.032	-0.001	—	0.114	0.001	—	0.112	-0.005	—
	.6	500	-0.124	-0.016	—	-0.126	-0.013	—	-0.223	-0.008	—	-0.238	-0.017	—	-0.393	-0.021	—	-0.377	-0.014	—
		1000	-0.118	-0.008	—	-0.116	-0.002	—	-0.213	-0.007	—	-0.213	-0.005	—	-0.366	-0.005	—	-0.366	-0.004	—
		2000	-0.110	-0.003	—	-0.117	-0.005	—	-0.218	-0.005	—	-0.223	-0.002	—	-0.370	-0.002	—	-0.373	0.000	—
	.9	500	-0.192	-0.013	—	-0.202	-0.012	—	-0.370	-0.014	—	-0.362	-0.018	—	-0.699	-0.012	—	-0.695	-0.017	—
		1000	-0.181	-0.005	—	-0.172	-0.005	—	-0.361	-0.005	—	-0.365	-0.004	—	-0.690	-0.007	—	-0.687	-0.006	—
		2000	-0.169	-0.002	—	-0.179	-0.001	—	-0.349	-0.001	—	-0.354	-0.003	—	-0.678	-0.004	—	-0.684	-0.003	—
30%	.0	500	-0.021	-0.007	—	-0.015	-0.013	—	-0.026	0.007	—	-0.030	0.010	—	-0.115	0.083	—	-0.111	0.072	—
		1000	-0.008	0.000	—	-0.008	0.002	—	-0.025	0.018	—	-0.022	0.019	—	-0.102	0.081	—	-0.106	0.084	—
		2000	-0.005	0.005	—	-0.007	0.005	—	-0.023	0.021	—	-0.023	0.021	—	-0.095	0.090	—	-0.095	0.084	—
	.6	500	-0.195	-0.005	—	-0.186	-0.008	—	-0.368	-0.001	—	-0.384	0.006	—	-0.778	0.044	—	-0.776	0.042	—
		1000	-0.181	-0.001	—	-0.176	0.002	—	-0.356	0.008	—	-0.359	0.010	—	-0.754	0.046	—	-0.757	0.045	—
		2000	-0.179	0.003	—	-0.176	0.002	—	-0.345	0.013	—	-0.353	0.013	—	-0.746	0.052	—	-0.751	0.054	—
	.9	500	-0.288	-0.013	—	-0.277	-0.016	—	-0.546	-0.005	—	-0.551	-0.007	—	-1.062	0.004	—	-1.080	0.008	—
		1000	-0.270	-0.006	—	-0.259	-0.004	—	-0.521	-0.003	—	-0.528	-0.001	—	-1.059	0.011	—	-1.061	0.004	—

		2000	-0.264	-0.001	—	-0.262	-0.001	—	-0.521	0.003	—	-0.521	0.004	—	-1.042	0.014	—	-1.035	0.015	—
50%	.0	500	-0.034	0.004	—	-0.034	-0.001	—	-0.085	0.043	—	-0.085	0.048	—	-0.329	0.204	—	-0.326	0.212	—
		1000	-0.024	0.011	—	-0.028	0.010	—	-0.082	0.048	—	-0.082	0.050	—	-0.323	0.207	—	-0.320	0.211	—
		2000	-0.024	0.013	—	-0.020	0.018	—	-0.074	0.052	—	-0.073	0.057	—	-0.314	0.211	—	-0.315	0.209	—
	.6	500	-0.192	-0.004	—	-0.188	0.000	—	-0.377	0.021	—	-0.385	0.024	—	-0.832	0.079	—	-0.832	0.085	—
		1000	-0.175	0.004	—	-0.184	0.005	—	-0.376	0.028	—	-0.368	0.023	—	-0.820	0.090	—	-0.807	0.088	—
		2000	-0.176	0.009	—	-0.177	0.010	—	-0.364	0.028	—	-0.367	0.031	—	-0.800	0.095	—	-0.800	0.092	—
	.9	500	-0.259	-0.008	—	-0.261	-0.004	—	-0.521	-0.002	—	-0.524	0.002	—	-1.045	0.010	—	-1.047	0.015	—
		1000	-0.254	-0.001	—	-0.255	-0.001	—	-0.505	0.005	—	-0.508	0.005	—	-1.027	0.021	—	-1.024	0.021	—
		2000	-0.246	0.000	—	-0.248	0.001	—	-0.500	0.009	—	-0.501	0.008	—	-1.017	0.026	—	-1.018	0.024	—

Table 4. 6. Average bias of the item location parameter (*d*) for the correct and misspecified models.

			Correct Specified Models			Misspecified Models		
Complexity	Correlation	N	LOW	MED	HGH	LOW	MED	HGH
10%	.0	500	-0.003	-0.001	-0.005	-0.003	-0.001	-0.005
		1000	-0.002	0.003	0.005	-0.002	0.003	0.005
		2000	<0.001	0.001	-0.001	<0.001	<0.001	-0.001
	.6	500	<0.001	-0.001	-0.003	<0.001	-0.001	-0.003
		1000	0.002	0.001	0.001	0.002	0.001	0.001
		2000	0.001	0.002	<0.001	0.001	0.002	<0.001
	.9	500	-0.002	<0.001	0.002	-0.002	<0.001	0.002
		1000	0.001	<0.001	-0.005	0.001	<0.001	-0.005
		2000	-0.002	-0.003	0.005	-0.002	-0.003	0.005
30%	.0	500	0.003	-0.003	0.001	0.002	-0.003	0.001
		1000	0.003	0.004	<0.001	0.003	0.004	<0.001
		2000	-0.001	-0.001	0.003	-0.001	-0.001	0.003

	.6	500	0.002	0.001	0.010	0.002	<0.001	0.010
		1000	0.005	-0.002	0.001	0.005	-0.002	0.001
		2000	<0.001	<0.001	-0.002	0.001	0.000	-0.002
	.9	500	0.006	-0.002	0.003	0.006	-0.003	0.003
		1000	0.001	0.001	0.002	0.001	0.001	0.002
		2000	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
50%	.0	500	-0.005	0.010	0.001	-0.005	0.010	<0.001
		1000	0.003	-0.001	-0.002	0.003	-0.002	-0.002
		2000	-0.002	0.003	0.001	-0.002	0.003	<0.001
	.6	500	-0.001	0.003	0.002	-0.001	0.004	0.003
		1000	-0.004	-0.003	0.003	-0.004	-0.002	0.002
		2000	0.004	-0.001	0.002	0.004	-0.001	0.003
	.9	500	0.001	0.006	-0.012	0.001	0.006	-0.011
		1000	<0.001	-0.002	0.008	0.001	-0.001	0.008
		2000	-0.001	-0.002	<0.001	-0.001	-0.002	<0.001

Effect of Sample Size on Item Parameter Recovery in Terms of Average Bias

Correct Specified Models

Item discrimination parameters. Figures 4.9 to 4.11 show the average bias for item discrimination parameters when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figures 4.9 to 4.11, all of the item discrimination parameters including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}$ $a_{2pcl}, a_{2pncl}, a_{2s}$) had very small or no change of average bias as sample size changed. Average bias of item discrimination for correct specified models increased as the sample size increased from 500 to 2,000 across all combinations of conditions with the lowest associated with the sample size of 500 and the largest associated with sample size of 2,000.

Item location parameter (d). Figure 4.15 shows the average bias for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figure 4.15, the average bias for item location parameter was near zero, with no effect as the sample size increased from 500 to 2,000 , suggesting that the estimated item location parameter were close to the true values (ranged from -0.012 to 0.010).

Misspecified Models

Item discrimination parameters. Figures 4.12 to 4.14 show the average bias for item discrimination parameters when models were misspecified with a low, medium or high degree of cross-loading across combinations of conditions. As it was mentioned in the previous sections there was no secondary item discrimination defined in the misspecified models. As shown in figures 4.12 to 4.14, the average bias of the item discrimination parameters on the first and second dimensions ($a_{1pcl}, a_{1pncl},$ a_{2pcl}, a_{2pncl}) tended to be closer to zero as sample size increased, but differences across sample sizes were small. It should be noted that average bias of item discrimination estimates tended to be negative when the item was supposed to be specified as cross-loaded (a_{1pcl}, a_{2pcl}) suggesting that estimated primary item discrimination parameters on the *truly* cross-loaded items (misspecified in the estimations) were somewhat smaller than their true values. On the other hand, discrimination estimates of non-cross loaded items (a_{1pncl}, a_{2pncl}) tended to be near zero or have positive values suggesting that estimated primary item discrimination parameters on the *truly* non-cross loaded items (not misspecified) were somewhat close to or slightly greater than their true values.

Item location parameter (d). Figure 4.16 shows the average bias for item location parameter (d) when models were misspecified with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figure 4.16, the average bias for item location parameter was not affected by changes in sample size with average bias close to zero suggesting that the estimated item location parameter were close to the true values (ranged from -0.011 to 0.010).

Effect of Correlation between Abilities on Item Parameter Recovery in Terms of Average Bias

Correct Specified Models

Item discrimination parameters. Figures 4.9 to 4.11 show the average bias for item discrimination parameters when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. As figures 4.9 to 4.11 show, when the degree of cross-loading was low or medium, the discrimination of cross-loaded items on both dimensions ($a_{1pcl}\ and\ a_{2pcl}$) was equal to or near zero when data were truly uncorrelated; as correlation increased, the average bias of cross-loaded items departed from zero in the positive direction.. These positive values indicate that estimated cross loaded primary item discrimination parameters on both of the dimensions were somewhat greater than their true values. However, when the degree of cross-loading was high, the item discrimination of cross-loaded items on both dimensions (a_{1pcl} and a_{2pcl}) departed from zero in the negative direction as the correlation between abilities increased from .0 to .9. These negative values indicate that estimated item discrimination parameters of cross-loaded items on both of the dimensions were somewhat smaller than their true values.

For non-cross loaded items' discrimination on both dimensions (a_{1pncl} and a_{2pncl}), the average bias had a constant pattern with values close to 0.000 when data were truly uncorrelated, and approaching -0.05 as correlation increased to .9. Secondary item discrimination parameters on both dimensions (a_{1s} , a_{2s}) had a consistent changing pattern with increasing negative values as the correlation increased from .0 to .9 across combinations of conditions. These negative values indicate that estimated cross-loaded secondary item discrimination parameters on both of the dimensions were smaller than their true values.

Item location parameter (d). Figure 4.15 shows the average bias for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figure 4.15, the average bias for item location parameter had a constant trend with values close to zero as the correlation increased from .0 to .9.

Misspecified Models

Item discrimination parameters. Figures 4.12 to 4.14 show the average bias for item discrimination parameters when models were misspecified with a low, medium or high degree of cross-loading across combinations of conditions. As figures 4.12 to 4.14 show, the *truly* cross-loaded primary item discrimination on both dimensions (i.e., misspecified items a_{1pcl} and a_{2pcl}) had an increasing pattern in absolute value (with negative values) as the correlation between abilities increased from .0 to .9. These negative values indicate that estimated cross-loaded primary item discrimination parameters on both of the dimensions were somewhat smaller than their true values. When the degree of cross-loading was low or medium, non-cross loaded item discriminations on both dimensions (i.e., not misspecified, a_{1pncl} and a_{2pncl}) had a constant pattern with values close to 0.000 as the correlation between abilities increased from .0 to .9. However, when the degree of cross loading was high, non-cross loaded item discrimination on both dimensions (a_{1pncl} and a_{2pncl}) remained constant with values near zero when complexity was 10% and increased (in the positive direction) as correlation increased when complexity was 30% or 50%.

Item location parameter (d). Figure 4.16 shows the average bias for item location parameter (d) when models were misspecified with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figure 4.16, the average bias for item location parameter had a constant trend as the correlation between abilities increased from .0 to .9 with values close to 0.000 suggesting that the estimated item location parameter were close to the true values (ranged from -0.011 to 0.010).

Effect of Structure Complexity Magnitude on Item Parameter Recovery in Terms of Average Bias

Correct Specified Models

Item discrimination parameters. As shown in figures 4.9 to 4.11, by holding the correlation and sample size constant, the average bias trends considering the effect of the three levels of complexity magnitude on the precision of item discrimination parameters including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}$ $a_{2pcl}, a_{2pncl}, a_{2s}$) can be investigated. When the degree of cross-loading was low or medium, the average bias of the item discrimination parameters on cross-loaded items (a_{1pcl}, a_{2pcl}) had an increasing pattern with positive values as the structure complexity increased from 10% to 50% (holding all other variables constant). However, when the degree of cross-loading was high, as the structure complexity increased from 10% to 30% and 50%, the average bias of the item discrimination on cross-loaded item departed from zero in the negative direction, suggesting that the estimated item parameters were smaller than the true item parameters.

The average bias for the item discrimination on non-cross loaded items (a_{1pncl}, a_{2pncl}) had a constant trend with values close to 0.000 across all levels of structure complexity. The average bias for the secondary item discriminations on both dimensions (a_{1s}, a_{2s}) departed from zero in the negative direction as the structure complexity magnitude increased from 10% to 50% especially when the correlation between abilities was either .6 or .9. The average bias for secondary item discrimination parameters were generally negative suggesting that the estimated item parameters were smaller than the true item parameters. It seemed that when the correlation between the abilities was .0, as the complexity magnitude increased from 10% to 50%, the average bias of secondary item discrimination parameters had a constant trend with values closer to 0.000 in correctly specified models.

Item location parameter (d). Figure 4.15 shows the average bias for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combination of conditions. As the complexity magnitude increased from 10% to 50%, item location parameter had a constant trend in terms of average bias across all combinations of conditions with values closer to 0.000.

Misspecified Models

Item discrimination parameters. As shown in figures 4.12 to 4.14 by holding the correlation and sample size constant the average bias trends considering the effect of the three levels of complexity magnitude on the precision of item discrimination parameters including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}\ a_{2pcl}, a_{2pncl}, a_{2s}$) can be investigated. The primary item discrimination parameters on *truly* cross-loaded items (i.e., misspecified items, a_{1pcl}, a_{2pcl}) had an increasing pattern in absolute value with generally negative values as the structure complexity increased from 10% to 50%. These negative values indicate that the estimated item parameters were smaller than the true item parameters. When the degree of cross-loading was low, the average bias for the primary item discrimination on non-cross loaded items ($a_{1pncl},\ a_{2pncl}$) had a constant trend with values close to 0.000 as the structure complexity increased from 10% to 50%. However, when the degree of cross-loading was medium or high, for the primary item discrimination on non-cross loaded items ($a_{1pncl},\ a_{2pncl}$) as the structure complexity increased from 10% to 50%, the average bias increased with positive values when correlation was greater than zero, suggesting that estimated values were greater than the true values.

Item location parameter (d). Figure 4.16 shows the average bias trends for the item location parameter (d) when models were misspecified across combination of conditions. As shown in figure 4.16, as the complexity magnitude increased from 10% to 50% item location parameter had a consistent steady trend in terms of average bias with values closer to 0.000 across all combination of conditions.

Effect of Degree of Cross-Loading on Item Parameter Recovery in Terms of Average Bias

Correct Specified Models

Item discrimination parameters. Figures 4.9 to 4.11 show the average bias trends related to item discrimination parameters including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}\ a_{2pcl}, a_{2pncl}, a_{2s}$) when the models were correctly specified and the degree of cross-loading was low, medium and high, respectively. As the degree of cross-loading increased from low to medium the average bias decreased with values close to zero for the primary item discrimination parameters on cross-loaded items on both dimension (a_{1pcl}, a_{2pcl}) with generally positive values, and continued to decrease, departing farther from zero in the negative direction for high degree of cross-loading. This pattern was more obvious when the correlation between abilities was .6 or .9 compared to a correlation of .0.

The primary item discrimination parameters on non-cross loaded items on both dimension (a_{1pncl}, a_{2pncl}) had a constant trend in terms of average bias as the degree of cross-loading changed. As shown in figures 4.9 to 4.11 the average bias for the secondary item discriminations on both dimensions (a_{1s} , a_{2s}) was near zero when correlation was .0 at all levels of cross loading. When correlation was greater than zero, the average bias departed from zero in the negative direction, and to a greater degree when the degree of cross-loading was low or medium. The negative values suggested that the estimated parameter were smaller than the true item parameters.

Item location parameter (d). Figure 4.15 shows the average bias for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combination of conditions. As shown in figure 4.15, the average bias for item location parameter was near zero, with no effect as the degree of cross-loading changed , suggesting that the estimated item location parameter were close to the true values (ranged from -0.012 to 0.010).

Misspecified Models

Item discrimination parameters. Figures 4.12 to 4.14 show the average bias trends related to item discrimination parameters on the first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{2pcl}, a_{2pncl}$) when the models were misspecified and the degree of cross-loading was low, medium or high, respectively. As the degree of cross-loading increased from low to high the average bias had larger departures from zero in the negative direction for the primary item discrimination parameters on *truly* cross-loaded items on both dimension (i.e., misspecified items, a_{1pcl}, a_{2pcl}). The item discrimination parameters on non-cross loaded items on both dimension (a_{1pncl}, a_{2pncl}) remained near zero at all levels of cross-loading.

Item location parameter (d). Figure 4.16 shows the average bias trends for item location parameter (d) when models were misspecified across combination of conditions. As shown in figure 4.16, as the degree of cross-loading increased from low to high, item location parameter had a constant trend in terms of average bias with values close to zero across all combinations of conditions.

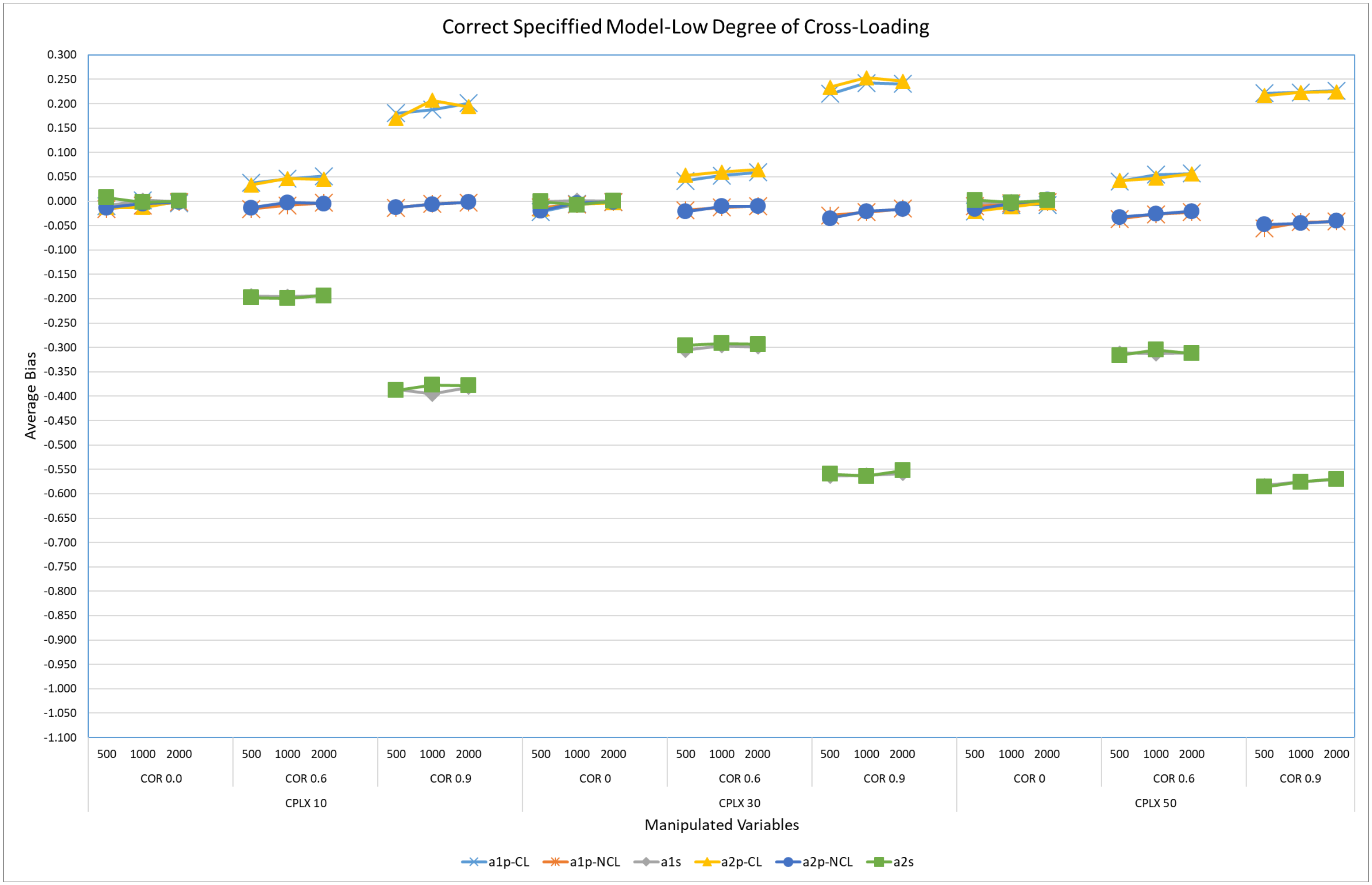

Figure 4. 9. Average bias for item discrimination parameters when models were specified correctly with a low degree of cross-loading

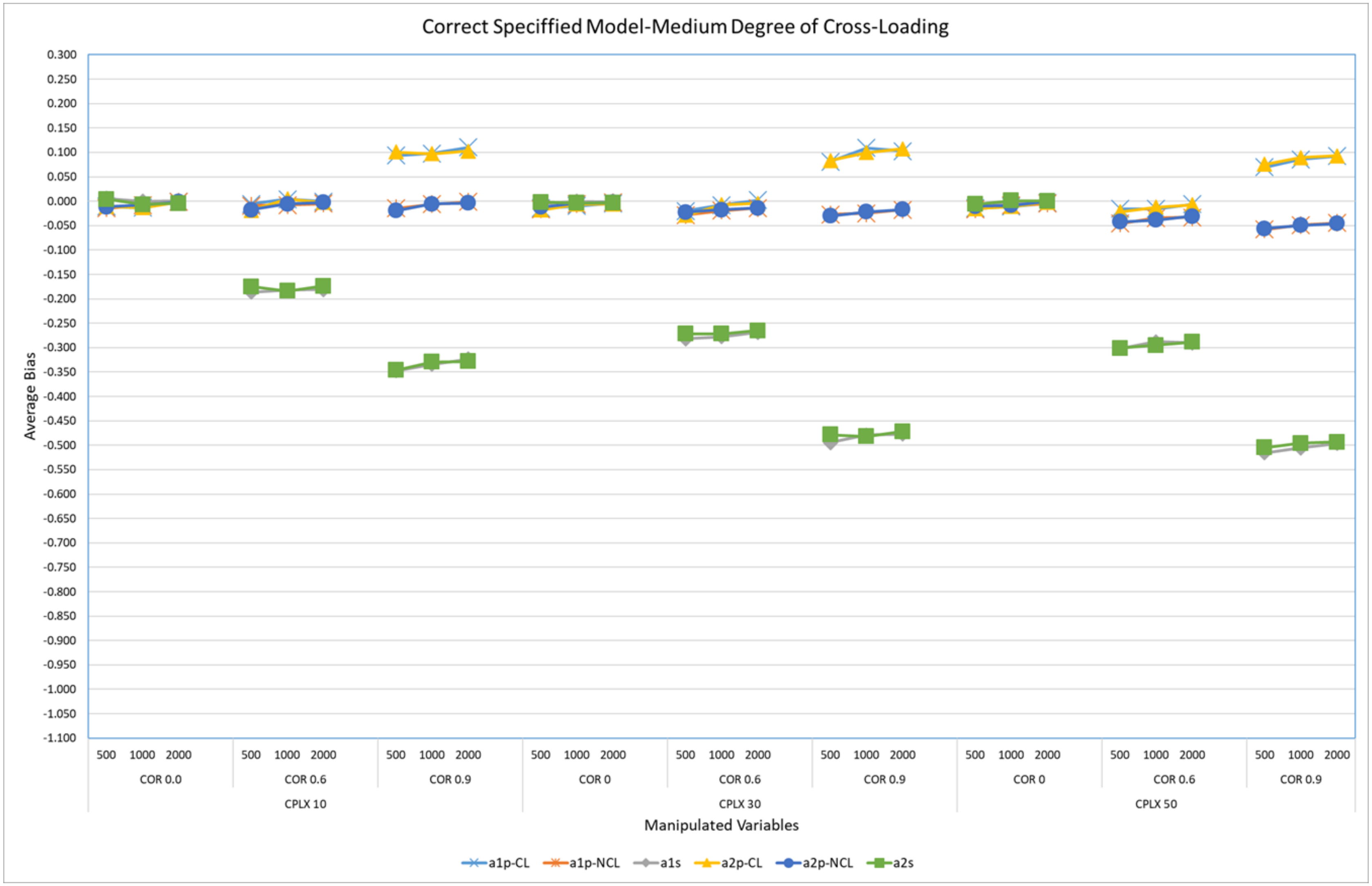

Figure 4. 10. Average bias for item discrimination parameters when models were specified correctly with a medium degree of cross-loading

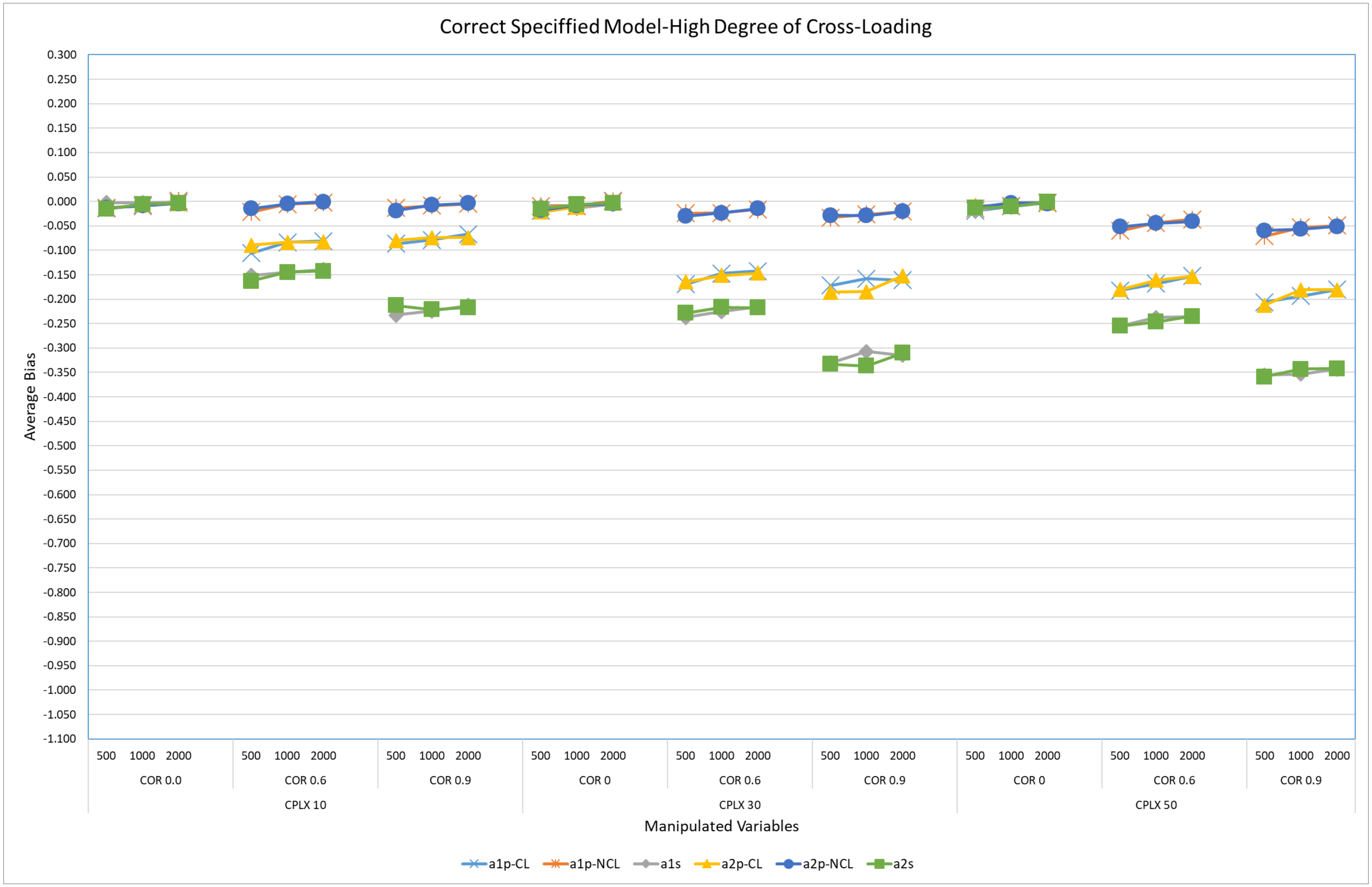

Figure 4. 11. Average bias for item discrimination parameters when models were specified correctly with a high degree of cross-loading

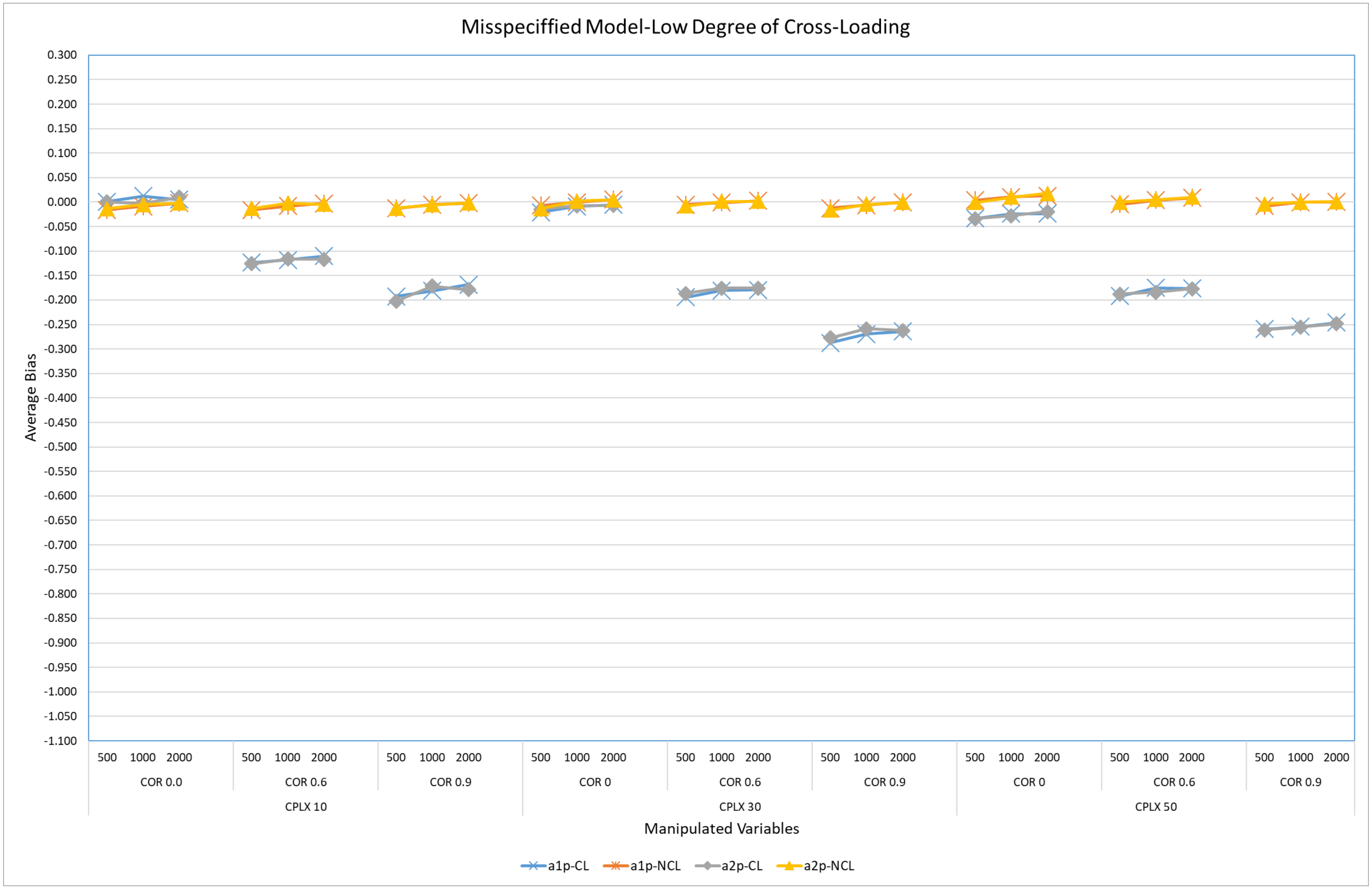

Figure 4. 12. Average bias for item discrimination parameters when models were misspecified with a low degree of cross-loading

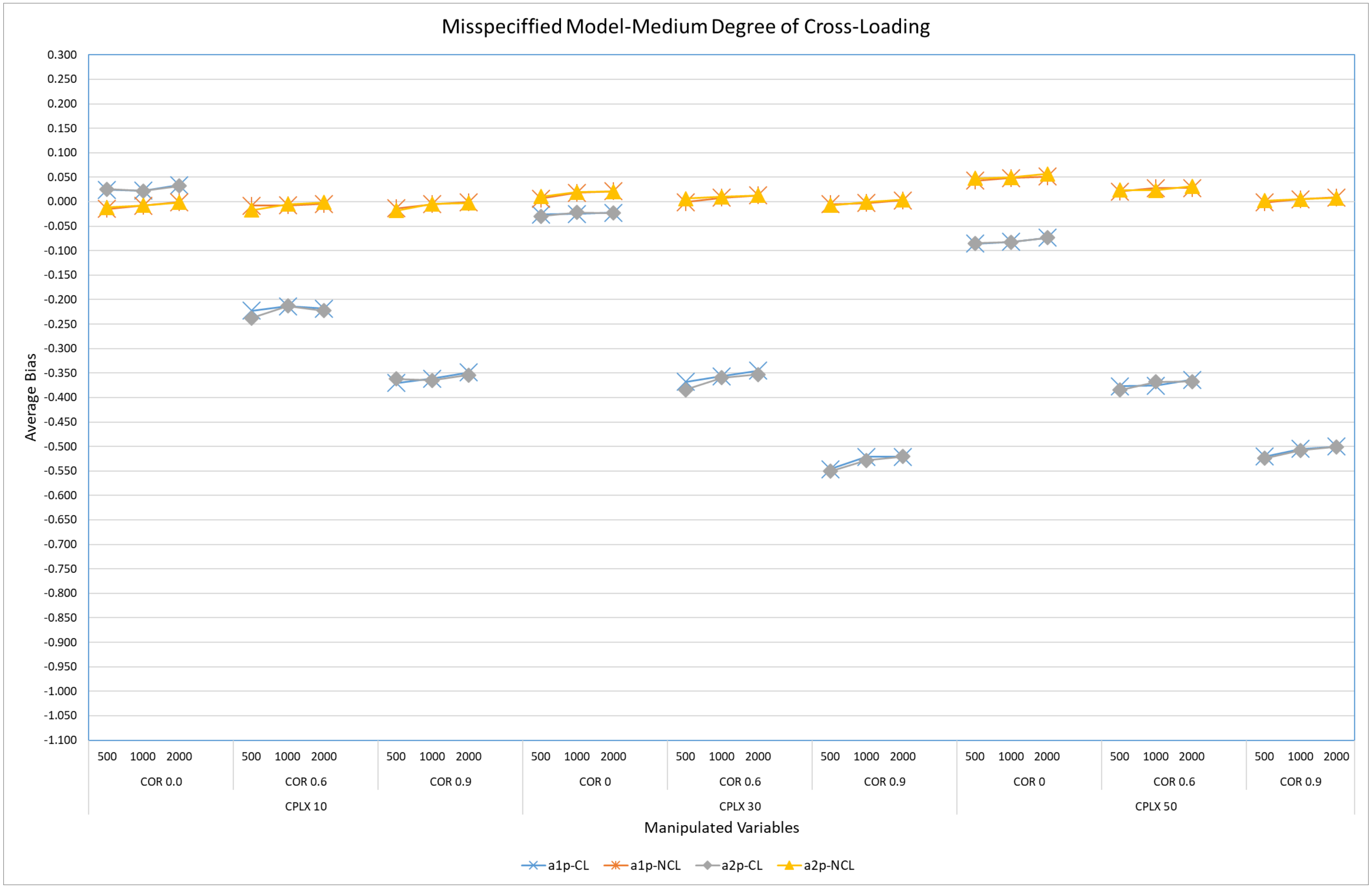

Figure 4. 13. Average bias for item discrimination parameters when models were misspecified with a medium degree of cross-loading

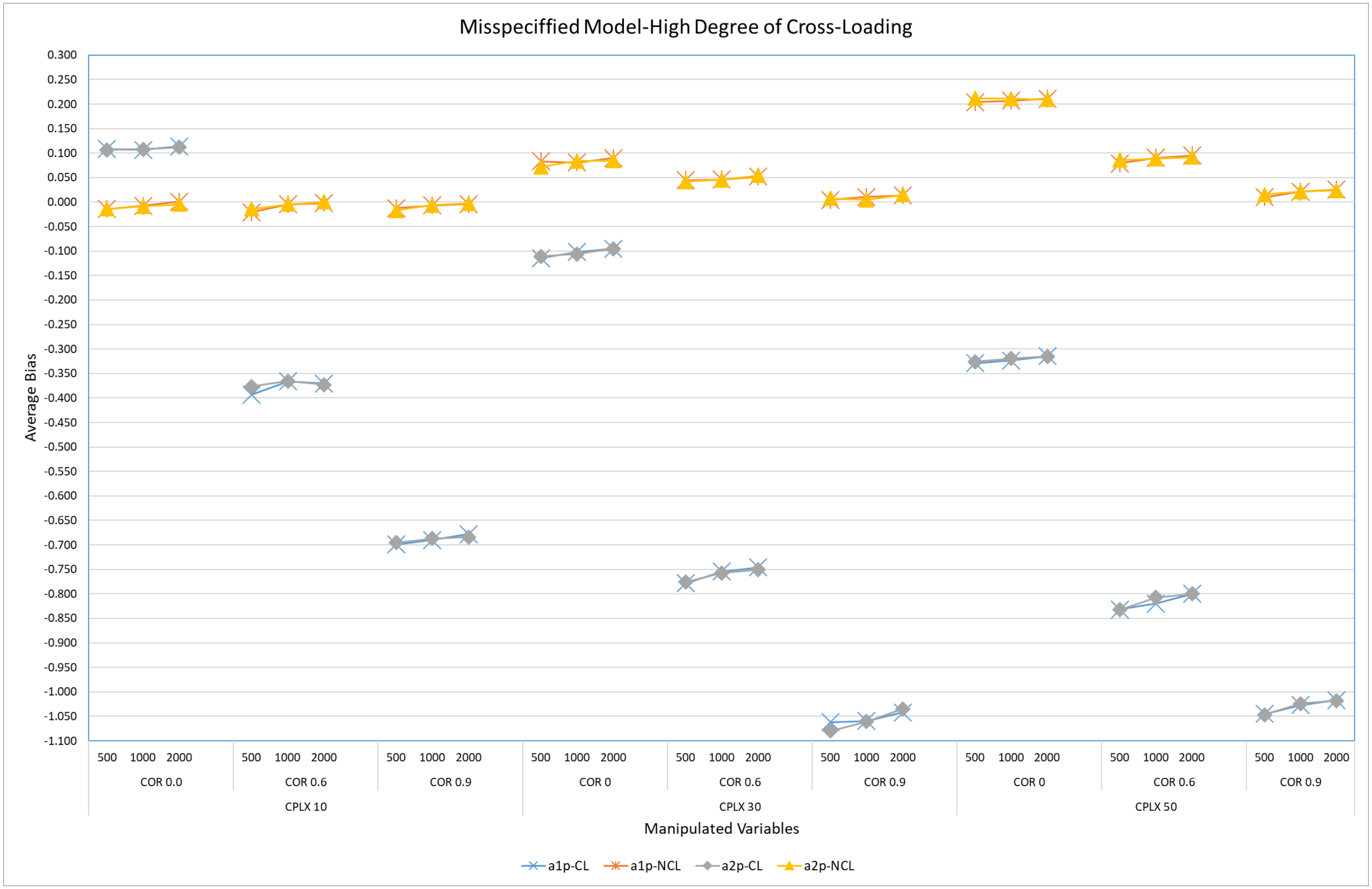

Figure 4. 14. Average bias for item discrimination parameters when models were misspecified with a high degree of cross-loading

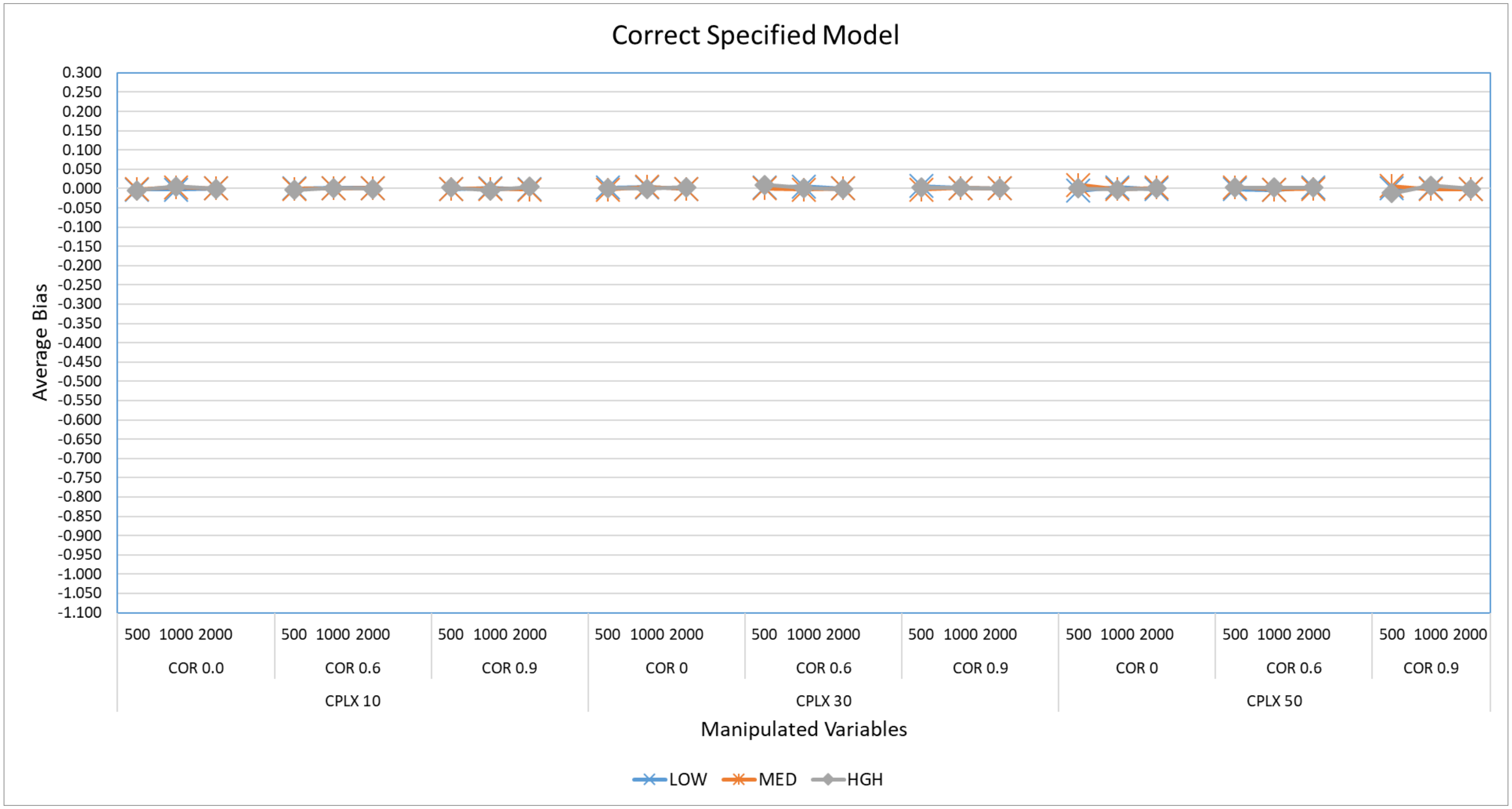

Figure 4. 15. Average bias for item location parameter when models were specified correctly.

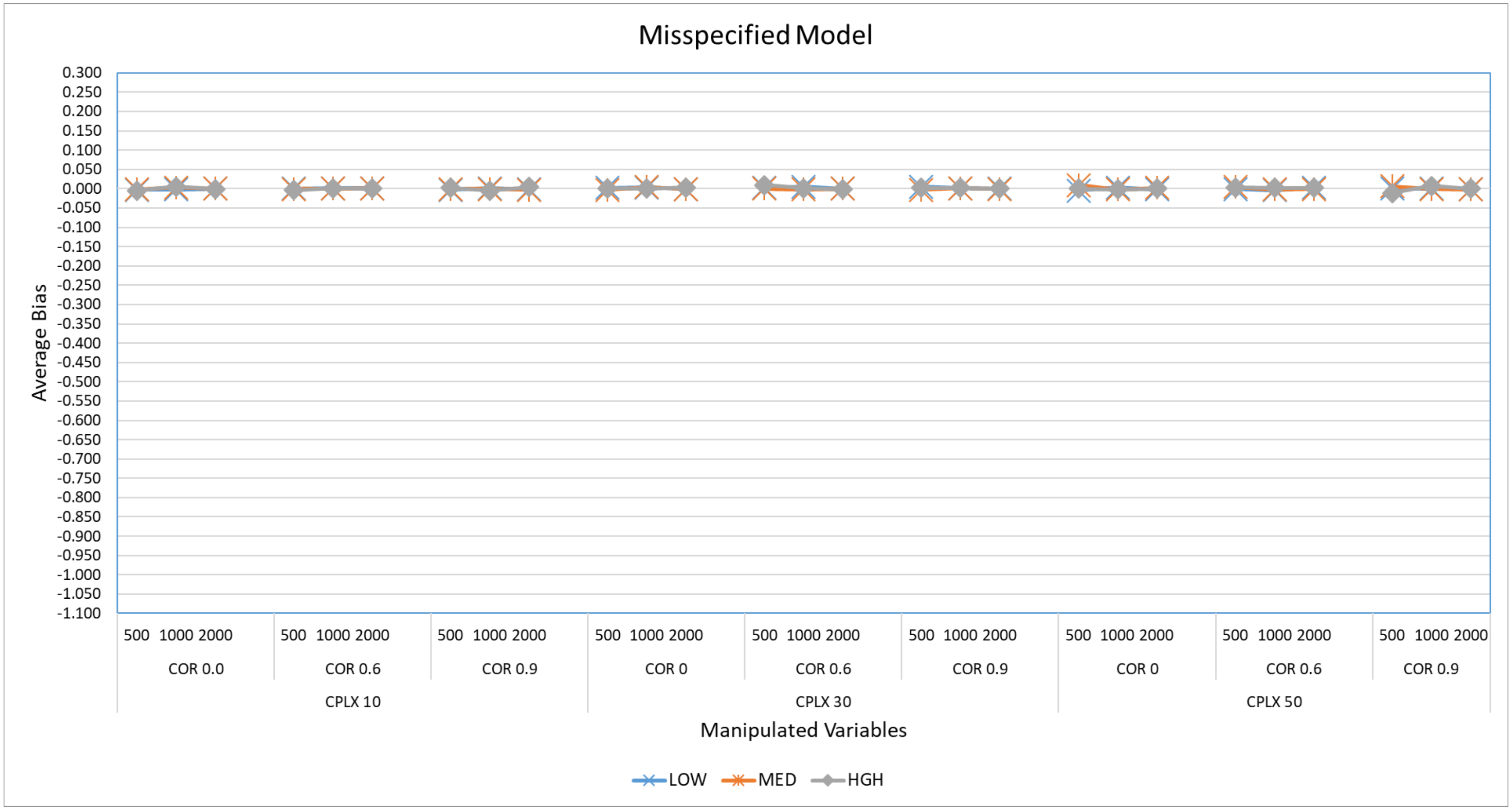

Figure 4. 16. Average bias for item location parameter when models were misspecified.

Section III: Item Parameter Recovery Results in Terms of Average Standard Error of Estimate (SEE)

Correct Specified Models

The SEEs were averaged for the estimated item discrimination parameters (a_1 and a_2) and the estimated item location parameter (d) within each combination of conditions. For the item discrimination parameters (a_1 and a_2) in correct specified models, the SEEs for the primary cross-loaded item discriminations on first and second dimensions (a_{1pcl} and a_{2pcl}) were averaged based on those items that had a primary item discrimination and were cross-loaded on both dimensions. Likewise, the SEEs for the primary non-cross loaded item discriminations on first and second dimensions (a_{1pncl} and a_{2pncl}) were averaged based on the items that had a primary item discrimination but were not cross-loaded on both dimensions. For the secondary item discriminations (a_{1s} and a_{2s}) the SEEs were averaged based on the items that had a secondary item discrimination.

Primary item discrimination parameters. Table 4.7 reports the average SEEs of the estimated item discrimination parameters when the models were correctly specified considering three levels of structure complexity magnitude incorporating low, medium or high degree of cross-loading. As reported in table 4.7, the average SEEs for the primary cross-loaded item discrimination parameters on the first dimension (a_{1pcl}) ranged from 0.078 to 0.262. The average SEEs for the primary non-cross loaded item discrimination parameters on the first dimension (a_{1pncl}) ranged from 0.091 to 0.202 across all combinations of conditions. On the other hand, the average SEEs for the primary cross-loaded item discrimination parameters on the second dimension (a_{2pcl}) ranged from 0.078 to 0.263. The average SEEs for the primary item discrimination parameters for non-cross loaded items on the second dimension (a_{2pncl}) ranged from 0.091 to 0.201 across all combinations of conditions.

Secondary item discrimination parameters. As reported in table 4.7, the average SEEs for the secondary item discrimination parameter for the first dimension (a_{1s}) ranged from 0.061 to 0.258. On the other hand, the average SEEs for the secondary item discrimination parameter on the second dimension (a_{2s}) ranged from 0.062 to 0.257 across all combinations of conditions. It should be noted that secondary item discrimination parameters on first and second dimensions had a very similar patterns and values in terms of average SEE.

Item location parameter (d). Table 4.9 reports the average SEEs for the item location parameter (d) for both correct specified and misspecified models across all combinations of conditions. The average SEE for item location parameter when the models were correctly specified ranged from 0.068 to 0.159. The lowest average SEE for item location parameter was associated with the condition when the correlation between the abilities was .0 and the sample size was 2,000. The highest average SEE for item location parameter was associated with the condition when the correlation between the abilities was .90 and the sample size was 500.

Misspecified Models

The SEEs were averaged for the estimated item discrimination parameters (a_1 and a_2) and the estimated item location parameter (d) within each combination of conditions for three sets of item cross-loading. For the item discrimination parameters (a_1 and a_2) in misspecified models, the SEEs for the *truly* primary cross-loaded item discriminations on first and second dimensions (a_{1pcl} and a_{2pcl}) were averaged based on those items that had a primary item discrimination and were supposed to be specified as cross-loaded items on both dimensions (these are the items that were misspecified in the model). Likewise, the SEEs for the *truly* non-cross loaded item discriminations on first and second dimensions (a_{1pncl} and a_{2pncl}) were averaged based on the items that had a primary item discrimination but were not truly cross-loaded on both dimensions. It should be noted that there was no secondary item discrimination defined for the first and second dimension in the misspecified models (a_{1s} and a_{2s}).

Primary item discrimination parameters. Table 4.8 reports the average SEEs of the estimated item discrimination parameters when the models were misspecified considering three levels of structure complexity magnitude incorporating low, medium or high degree of cross-loading. As shown in table 4.8, the average SEEs for the truly primary cross-loaded item discrimination parameters on the first dimension (a_{1pcl}) ranged from 0.073 to 0.344. The average SEEs for the *truly* primary non-cross loaded item discrimination parameters on the first dimension (a_{1pncl}) ranged from 0.078 to 0.197 across all combinations of conditions. On the other hand, the average SEEs for the *truly* primary cross-loaded item discrimination parameters on the second dimension (a_{2pcl}) ranged from 0.073 to 0.348. The average SEEs for the truly primary non-cross loaded item discrimination parameters on the second dimension (a_{2pncl}) ranged from 0.078 to 0.197 across all combinations of conditions. In conclusion, item discrimination estimates on the truly primary dimension tended to have larger SEEs when the item was supposed to be specified as cross-loaded than when it was non-crossloaded. In addition, it should be noted that very similar patterns and values in terms of average SEEs were observed for the truly primary item discrimination parameters on the first and second dimensions.

Item location parameter (d). Table 4.9 reports the average SEEs for the item location parameter (d) for both correct specified and misspecified models across all combinations of conditions. The average SEEs for item location parameter when the models were misspecified ranged from 0.067 to 0.160. The lowest average SEE for item location parameter was associated with the condition when the correlation between the abilities was .0 and the sample size was 2,000. The highest average SEE for item location parameter was associated with the condition when the correlation between the abilities was .9 and the sample size was 500. It is worth mentioning that the SEEs for item location parameter (d) had a very similar pattern and values comparing the misspecified and correct specified models.

Table 4. 7. Average SEE of the primary (cross-loaded and non-cross loaded items) and secondary discrimination on the first and second dimensions when the models were correctly specified.

			LOW						MED						HGH					
			a_1			a_2			a_1			a_2			a_1			a_2		
Complexity	Correlation	N	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}
10%	.0	500	0.157	0.196	0.124	0.158	0.196	0.124	0.159	0.195	0.130	0.159	0.195	0.130	0.165	0.194	0.152	0.164	0.194	0.153
		1000	0.110	0.137	0.087	0.111	0.137	0.087	0.112	0.137	0.091	0.112	0.137	0.091	0.115	0.136	0.107	0.115	0.136	0.107
		2000	0.078	0.096	0.061	0.078	0.096	0.062	0.078	0.096	0.064	0.078	0.096	0.064	0.081	0.096	0.075	0.081	0.096	0.075
	.6	500	0.167	0.196	0.140	0.167	0.196	0.140	0.173	0.195	0.150	0.174	0.196	0.148	0.187	0.194	0.175	0.186	0.194	0.177
		1000	0.117	0.137	0.098	0.117	0.137	0.098	0.121	0.137	0.104	0.121	0.137	0.105	0.130	0.135	0.123	0.130	0.135	0.123
		2000	0.082	0.096	0.069	0.082	0.096	0.069	0.085	0.096	0.074	0.085	0.096	0.074	0.092	0.095	0.087	0.092	0.095	0.087
	.9	500	0.181	0.196	0.168	0.181	0.196	0.168	0.190	0.195	0.180	0.189	0.196	0.180	0.212	0.193	0.207	0.211	0.194	0.205
		1000	0.126	0.137	0.118	0.126	0.137	0.118	0.133	0.137	0.126	0.133	0.137	0.126	0.148	0.135	0.145	0.148	0.135	0.145
		2000	0.089	0.097	0.083	0.089	0.097	0.083	0.093	0.096	0.088	0.094	0.096	0.088	0.104	0.095	0.102	0.104	0.095	0.102
30%	.0	500	0.202	0.189	0.140	0.201	0.189	0.140	0.201	0.190	0.150	0.202	0.189	0.150	0.205	0.189	0.188	0.205	0.190	0.189
		1000	0.140	0.132	0.098	0.140	0.132	0.098	0.141	0.132	0.105	0.141	0.132	0.106	0.143	0.133	0.132	0.143	0.133	0.132
		2000	0.099	0.093	0.069	0.099	0.093	0.069	0.099	0.093	0.074	0.099	0.093	0.074	0.100	0.093	0.093	0.100	0.094	0.093
	.6	500	0.209	0.191	0.159	0.208	0.192	0.159	0.217	0.192	0.177	0.219	0.191	0.176	0.239	0.189	0.226	0.239	0.191	0.225
		1000	0.146	0.134	0.112	0.145	0.134	0.112	0.151	0.134	0.124	0.151	0.134	0.123	0.165	0.133	0.157	0.166	0.133	0.157
		2000	0.102	0.094	0.079	0.102	0.094	0.079	0.106	0.094	0.087	0.107	0.094	0.087	0.116	0.093	0.110	0.117	0.093	0.110
	.9	500	0.212	0.195	0.190	0.210	0.196	0.190	0.228	0.193	0.211	0.228	0.193	0.209	0.262	0.191	0.258	0.263	0.190	0.257
		1000	0.147	0.137	0.133	0.146	0.137	0.133	0.158	0.135	0.147	0.158	0.135	0.147	0.183	0.133	0.179	0.183	0.133	0.180
		2000	0.104	0.096	0.094	0.103	0.096	0.094	0.111	0.095	0.103	0.111	0.095	0.103	0.128	0.094	0.125	0.127	0.094	0.125
50%	.0	500	0.201	0.185	0.147	0.201	0.186	0.147	0.201	0.189	0.156	0.202	0.188	0.156	0.205	0.193	0.189	0.204	0.193	0.189
		1000	0.140	0.130	0.103	0.140	0.130	0.103	0.141	0.132	0.109	0.141	0.132	0.109	0.143	0.135	0.132	0.143	0.134	0.133
		2000	0.099	0.091	0.073	0.099	0.091	0.073	0.099	0.093	0.077	0.099	0.092	0.077	0.100	0.095	0.093	0.101	0.095	0.093
	.6	500	0.209	0.193	0.165	0.209	0.193	0.166	0.216	0.195	0.182	0.218	0.194	0.181	0.236	0.195	0.224	0.237	0.194	0.225
		1000	0.145	0.135	0.116	0.146	0.135	0.116	0.152	0.135	0.127	0.151	0.136	0.127	0.165	0.136	0.156	0.164	0.136	0.157
		2000	0.102	0.095	0.082	0.102	0.094	0.082	0.106	0.095	0.089	0.106	0.095	0.089	0.115	0.095	0.110	0.115	0.095	0.110

	.9	500	0.212	0.202	0.195	0.212	0.201	0.195	0.227	0.198	0.212	0.226	0.199	0.212	0.257	0.197	0.252	0.258	0.195	0.252
		1000	0.148	0.140	0.136	0.148	0.140	0.136	0.157	0.138	0.148	0.158	0.139	0.148	0.179	0.136	0.176	0.178	0.137	0.175
		2000	0.104	0.099	0.096	0.104	0.099	0.096	0.110	0.097	0.104	0.110	0.097	0.104	0.125	0.096	0.123	0.125	0.096	0.123

Table 4. 8. Average SEE of the truly primary (cross-loaded and non-cross loaded items) on the first and second dimensions when the models were misspecified.

			LOW						MED						HGH					
			a_1			a_2			a_1			a_2			a_1			a_2		
Complexity	Correlation	N	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}	a_{1p} CL	a_{1p} NCL	a_{1s}	a_{2p} CL	a_{2p} NCL	a_{2s}
10%	.0	500	0.156	0.196	—	0.156	0.196	—	0.153	0.196	—	0.154	0.196	—	0.147	0.197	—	0.147	0.197	—
		1000	0.109	0.137	—	0.110	0.137	—	0.108	0.137	—	0.108	0.137	—	0.103	0.138	—	0.103	0.138	—
		2000	0.077	0.096	—	0.077	0.096	—	0.076	0.096	—	0.076	0.096	—	0.073	0.097	—	0.073	0.097	—
	.6	500	0.167	0.195	—	0.168	0.195	—	0.178	0.194	—	0.179	0.194	—	0.196	0.193	—	0.194	0.193	—
		1000	0.117	0.137	—	0.117	0.136	—	0.124	0.136	—	0.124	0.136	—	0.136	0.135	—	0.136	0.135	—
		2000	0.082	0.096	—	0.083	0.096	—	0.088	0.096	—	0.088	0.096	—	0.096	0.095	—	0.096	0.095	—
	.9	500	0.174	0.194	—	0.175	0.194	—	0.194	0.193	—	0.192	0.193	—	0.237	0.191	—	0.235	0.191	—
		1000	0.122	0.136	—	0.121	0.136	—	0.135	0.135	—	0.136	0.135	—	0.165	0.134	—	0.165	0.134	—
		2000	0.085	0.096	—	0.086	0.096	—	0.095	0.095	—	0.095	0.095	—	0.115	0.094	—	0.116	0.094	—
30%	.0	500	0.201	0.188	—	0.200	0.188	—	0.202	0.186	—	0.202	0.186	—	0.217	0.179	—	0.216	0.180	—
		1000	0.140	0.131	—	0.140	0.131	—	0.142	0.130	—	0.142	0.130	—	0.150	0.126	—	0.151	0.126	—
		2000	0.099	0.092	—	0.099	0.092	—	0.100	0.092	—	0.100	0.092	—	0.106	0.089	—	0.106	0.089	—
	.6	500	0.221	0.184	—	0.220	0.184	—	0.242	0.181	—	0.244	0.180	—	0.300	0.172	—	0.300	0.172	—
		1000	0.154	0.129	—	0.153	0.129	—	0.168	0.127	—	0.169	0.126	—	0.208	0.121	—	0.208	0.121	—
		2000	0.108	0.091	—	0.108	0.091	—	0.118	0.089	—	0.119	0.089	—	0.145	0.085	—	0.146	0.085	—
	.9	500	0.232	0.183	—	0.230	0.184	—	0.265	0.179	—	0.266	0.179	—	0.344	0.173	—	0.348	0.172	—
		1000	0.161	0.128	—	0.160	0.128	—	0.183	0.126	—	0.184	0.125	—	0.240	0.121	—	0.240	0.121	—
		2000	0.113	0.090	—	0.113	0.090	—	0.129	0.088	—	0.129	0.088	—	0.167	0.085	—	0.167	0.085	—

50%	.0	500	0.202	0.182	—	0.201	0.183	—	0.207	0.177	—	0.207	0.177	—	0.236	0.158	—	0.236	0.158	—
		1000	0.141	0.128	—	0.141	0.128	—	0.145	0.124	—	0.145	0.124	—	0.165	0.111	—	0.165	0.111	—
		2000	0.099	0.090	—	0.099	0.090	—	0.102	0.087	—	0.102	0.087	—	0.116	0.078	—	0.116	0.078	—
	.6	500	0.215	0.178	—	0.216	0.178	—	0.234	0.172	—	0.236	0.171	—	0.288	0.160	—	0.288	0.159	—
		1000	0.150	0.125	—	0.151	0.125	—	0.164	0.120	—	0.164	0.121	—	0.201	0.112	—	0.199	0.112	—
		2000	0.106	0.088	—	0.106	0.088	—	0.115	0.085	—	0.115	0.085	—	0.140	0.079	—	0.140	0.079	—
	.9	500	0.222	0.177	—	0.222	0.177	—	0.249	0.171	—	0.250	0.171	—	0.313	0.163	—	0.314	0.163	—
		1000	0.155	0.124	—	0.156	0.124	—	0.174	0.120	—	0.174	0.120	—	0.217	0.114	—	0.217	0.115	—
		2000	0.109	0.087	—	0.109	0.087	—	0.122	0.084	—	0.122	0.084	—	0.152	0.081	—	0.152	0.081	—

Table 4. 9. Average SEE of the item location parameter (*d*) for the correct and misspecified models.

			Correct Specified Models			Misspecified Models		
Complexity	Correlation	N	LOW	MED	HGH	LOW	MED	HGH
10%	.0	500	0.138	0.138	0.139	0.138	0.138	0.138
		1000	0.097	0.097	0.098	0.097	0.097	0.097
		2000	0.068	0.068	0.069	0.068	0.068	0.068
	.6	500	0.139	0.139	0.140	0.138	0.139	0.139
		1000	0.097	0.097	0.098	0.097	0.097	0.098
		2000	0.068	0.069	0.069	0.068	0.069	0.069
	.9	500	0.138	0.139	0.140	0.138	0.139	0.140
		1000	0.097	0.098	0.098	0.097	0.098	0.098
		2000	0.068	0.069	0.069	0.069	0.069	0.069
30%	.0	500	0.139	0.140	0.144	0.138	0.137	0.135
		1000	0.097	0.098	0.101	0.097	0.096	0.095
		2000	0.068	0.069	0.071	0.068	0.068	0.067
	.6	500	0.140	0.143	0.149	0.140	0.141	0.145
		1000	0.098	0.100	0.104	0.098	0.099	0.101

		2000	0.069	0.070	0.073	0.069	0.070	0.071
	.9	500	0.141	0.144	0.151	0.141	0.144	0.150
		1000	0.099	0.101	0.106	0.099	0.101	0.106
		2000	0.070	0.071	0.074	0.070	0.071	0.074
50%	.0	500	0.139	0.141	0.146	0.138	0.138	0.139
		1000	0.097	0.099	0.103	0.097	0.097	0.098
		2000	0.069	0.069	0.072	0.068	0.068	0.069
	.6	500	0.142	0.146	0.156	0.141	0.145	0.153
		1000	0.099	0.102	0.109	0.099	0.102	0.107
		2000	0.070	0.072	0.076	0.070	0.072	0.076
	.9	500	0.143	0.148	0.159	0.143	0.148	0.160
		1000	0.100	0.103	0.111	0.100	0.104	0.112
		2000	0.070	0.073	0.078	0.071	0.073	0.079

Effect of Sample Size on Item Parameter Recovery in Terms of SEE

Correct Specified Models

Item discrimination parameters. Figures 4.17 to 4.19 show the average SEEs for item discrimination parameters when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figures 4.17 to 4.19, all of the item discrimination parameters including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}$ $a_{2pcl}, a_{2pncl}, a_{2s}$) had a consistent decreasing trend in terms of average SEE as the sample size increased from 500 to 2,000 across all combinations of conditions. The lowest average SEE of item discrimination for correct specified models was associated with the sample size of 2,000 and the highest average SEE of item discrimination for correct specified models was associated with the sample size of 500.

Item location parameter (d). Figure 4.23 shows the average SEE trends for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figure 4.23, item location parameter had a consistent decreasing trend in terms of average SEE as the sample size increased from 500 to 2,000 across all combinations of conditions. The lowest average SEE of item location for correct specified models was associated with the sample size of 2,000 and the highest average SEE of item location for correct specified models was associated with the sample size of 500.

Misspecified Models

Item discrimination parameters. Figures 4.20 to 4.22 show the average SEEs for item discrimination parameters when models were misspecified with a low, medium or high degree of cross-loading across combination of conditions. As shown in figures 4.20 to 4.22, all of the item discrimination parameters on the first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{2pcl}, a_{2pncl}$) had a consistent decreasing trend in terms of average SEE as the sample size increased from 500 to

2,000 across all combinations of conditions. The lowest average SEE was associated with the sample size of 2,000 and the highest average SEE was associated with the sample size of 500.

Item Location parameter (d). Figure 4.24 shows the average SEEs for item location parameter (d) when models were misspecified with a low, medium or high degree of cross-loading across combination of conditions. As shown in figure 4.25, item location parameter had a consistent decreasing trend in terms of average SEE as the sample size increased from 500 to 2,000 across all combination of conditions. The lowest average SEE of item discrimination for misspecified models was associated with the sample size of 2,000 and the highest average SEE of item discrimination for misspecified models was associated with the sample size of 500.

Effect of Correlation between Abilities on Item Parameter Recovery in Terms of SEE

Correct Specified Models

Item discrimination parameters. Figures 4.17 to 4.19 show the average SEE trends considering the effect of the correlation between abilities on the precision of item discrimination parameters including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}$ $a_{2pcl}, a_{2pncl}, a_{2s}$). When the correlation between abilities varied across the conditions while the sample size and complexity magnitude were held constant the average SEE increased consistently for the primary and secondary item discriminations on both dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}$ $a_{2pcl}, a_{2pncl}, a_{2s}$) as the correlation increased from .0 to .9 across combination of all conditions. As the correlation increased from .0 to .9 the increasing pattern in terms of average SEE was greater for the primary item discrimination parameters on cross-loaded items ($a_{1pcl},\ a_{2pcl}$) and the secondary item discrimination parameters (a_{1s} , a_{2s}) compared to the item discrimination parameters on non-cross loaded items (a_{1pncl}, a_{2pncl}).

Item location parameter (d). Figure 4.23 shows the average SEE patterns for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across all combinations of conditions. As shown in figure 4.23, item location parameter had a constant trend with relatively slight increases in terms of average SEE as the correlation between abilities increased from .0 to .9 across all combinations of conditions.

Misspecified Models

Item discrimination parameters. Figures 4.20 to 4.22 show the average SEE trends considering the effect of the correlation between abilities on the precision of item discrimination parameters. When the sample size and complexity magnitude were held constant , the average SEE increased consistently for the primary item discriminations on *truly* cross-loaded items on both dimensions (i.e., misspecified items, a_{1pcl}, a_{2pcl}) as the correlation increased from .0 to .9 across combinations of conditions. For the item discrimination parameters on the truly non-cross loaded items (a_{1pncl}, a_{2pncl}) the average SEE decreased slightly as the correlation between abilities increased from .0 to .9

Item location parameter (d). Figure 4.24 shows the average SEE patterns for item location parameter (d) when models were misspecified with a low, medium or high degree of cross-loading across combination of conditions. As shown in figure 4.24, item location parameter had a constant trend in terms of average SEE as the correlation between abilities increased from .0 to .9 across all combination of conditions.

Effect of Structure Complexity Magnitude on Item Parameter Recovery in Terms of SEE

Correct Specified Models

Item discrimination parameters. By holding the correlation and sample size constant, figures 4.17 to 4.19 show the average SEE trends considering the effect of the three levels of complexity magnitude on the precision of item discrimination parameters including primary and secondary item discrimination on first and second dimensions. The average SEEs for the primary item discriminations on cross-loaded items on both dimensions (a_{1pcl}, a_{2pcl}) increased as the structure complexity magnitude increased from 10% to 30% then decreased slightly as the complexity magnitude increased from 30% to 50%. The average SEE on non-cross loaded items (a_{1pncl}, a_{2pncl}) had a constant trend with values close to zero or slight decreases as the complexity magnitude increased from 10% to 50%. The average SEE for the secondary item discrimination (a_{1s} , a_{2s}) had an increasing pattern as complexity magnitude increased from 10% to 30% and then a decreasing pattern as the complexity magnitude increased from 30% to 50%.

Item location parameter (d). Figure 4.23 shows the average SEEs for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combination of conditions. As shown in figure 4.24, item location parameter had a constant trend with slight increases in terms of average SEE as the complexity magnitude increased from 10% to 50% across all combination of conditions.

Misspecified Models

Item discrimination parameters. By holding the correlation and sample size constant, figures 4.20 to 4.22 show the average SEE trends considering the effect of the three levels of complexity magnitude on the precision of item discrimination parameters on first and second dimensions when the models were misspecified. The average SEEs for the primary item discriminations on *truly* cross-loaded items on both dimensions (a_{1pcl}, a_{2pcl}) increased as the structure complexity magnitude increased from 10% to 30% then decreased slightly as the complexity magnitude increased from 30% to 50%. The average SEE on non-cross loaded items (a_{1pncl}, a_{2pncl}) tended to slightly decrease as the complexity magnitude increased across all levels from 10% to 50%.

Item location parameter (d). Figure 4.24 shows the average SEEs for item location parameter (d) when models were misspecified with a low, medium or high degree of cross-loading across combination of conditions. As it can be seen in figure 4.25, item location parameter had a constant trend with slight increases in terms of average SEE as the complexity magnitude increased from 10% to 50% across all combination of conditions.

Effect of the Degree of Cross-Loading on Item Parameter Recovery in Terms of SEE

Correct Specified Models

Item discrimination parameters. Figures 4.17 to 4.19 show the average SEE trends related to item discrimination parameters including primary and secondary item discrimination on first and second dimensions ($a_{1pcl}, a_{1pncl}, a_{1s}\ a_{2pcl}, a_{2pncl}, a_{2s}$) when the models were correctly specified and the degree of cross-loading was low, medium and high, respectively. As the degree of cross-loading increased from low to high the average SEE increased for the item discrimination parameters on cross-loaded items on both dimension (a_{1pcl}, a_{2pcl}). This pattern was more obvious when the correlation between abilities were .6 or .9 compared to a correlation of .0. The primary item discrimination parameters on non-cross loaded items on both dimension (a_{1pncl}, a_{2pncl}) had a constant trend in terms of average SEE as the degree of cross-loading increased from low to high. The average SEE for the secondary item discriminations on both dimensions ($a_{1s}\ , a_{2s}$) increased as the degree of cross-loading increased from low to high.

Item location parameter (d). Figure 4.23 shows the average SEEs for item location parameter (d) when models were specified correctly with a low, medium or high degree of cross-loading across combinations of conditions. As shown in figure 4.23, item location parameter had a constant trend in terms of average SEE as the degree of cross-loading increased from low to high. However, when the complexity magnitude was 50% and the correlation was either .6 or .9 the average SEE had a slightly increasing pattern as the degree of cross-loading increased from low to high.

Misspecified Models

Item discrimination parameters. Figures 4.20 to 4.22 show the average SEE trends related to item discrimination parameters on the first and second dimensions when the models were misspecified and the degree of cross-loading was low, medium and high, respectively. For the primary item discrimination parameters on *truly* cross-loaded items on both dimension (a_{1pcl}, a_{2pcl}) the average SEE increased as the degree of cross-loading increased from low to high. This pattern was more obvious when the complexity magnitude was 30% or 50% compared to a complexity magnitude of 10%. On the other hand, when the complexity magnitude was 10% the primary item discrimination parameters on *truly* non-cross loaded items on both dimension (a_{1pncl}, a_{2pncl}) had a constant trend in terms of average SEE as the degree of cross-loading increased from low to high. However, when the complexity magnitude was 30% or 50% the average SEEs on *truly* non-cross loaded items had a decreasing pattern as the degree of cross-loading increased from low to high.

Item location parameter (d). Figure 4.24 shows the average SEEs for item location parameter (d) when models were misspecified with a low, medium or high degree of cross-loading across combination of conditions. As can be seen in figure 4.24, item location parameter had a constant trend in terms of average SEE as the degree of cross-loading increased from low to high. However, when the complexity magnitude was 50% and the correlation was either .6 or .9 the average SEE had slight increases as the degree of cross-loading increased from low to high.

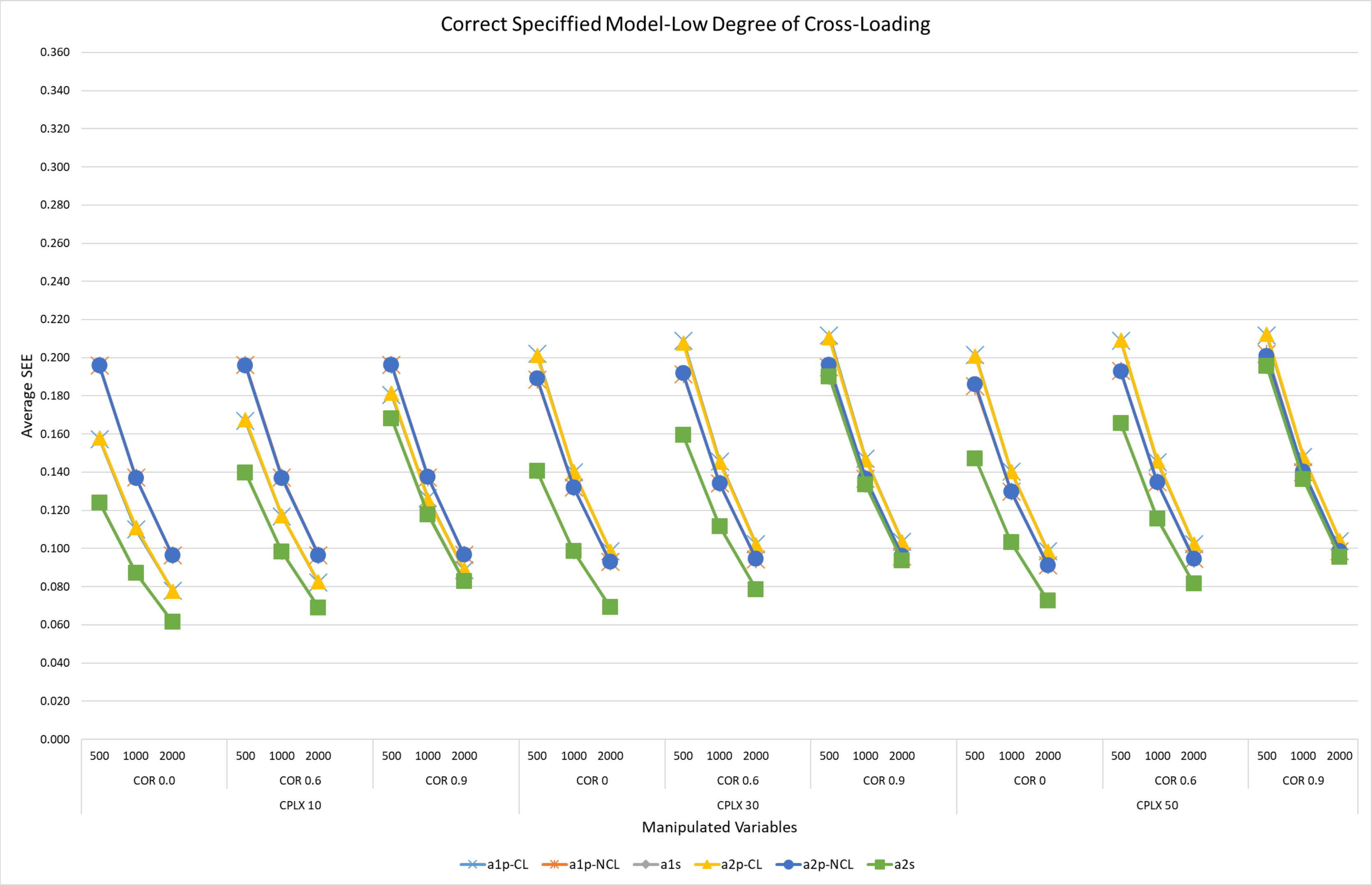

Figure 4. 17. Average SEE for item discrimination parameters when models were specified correctly with a low degree of cross-loading

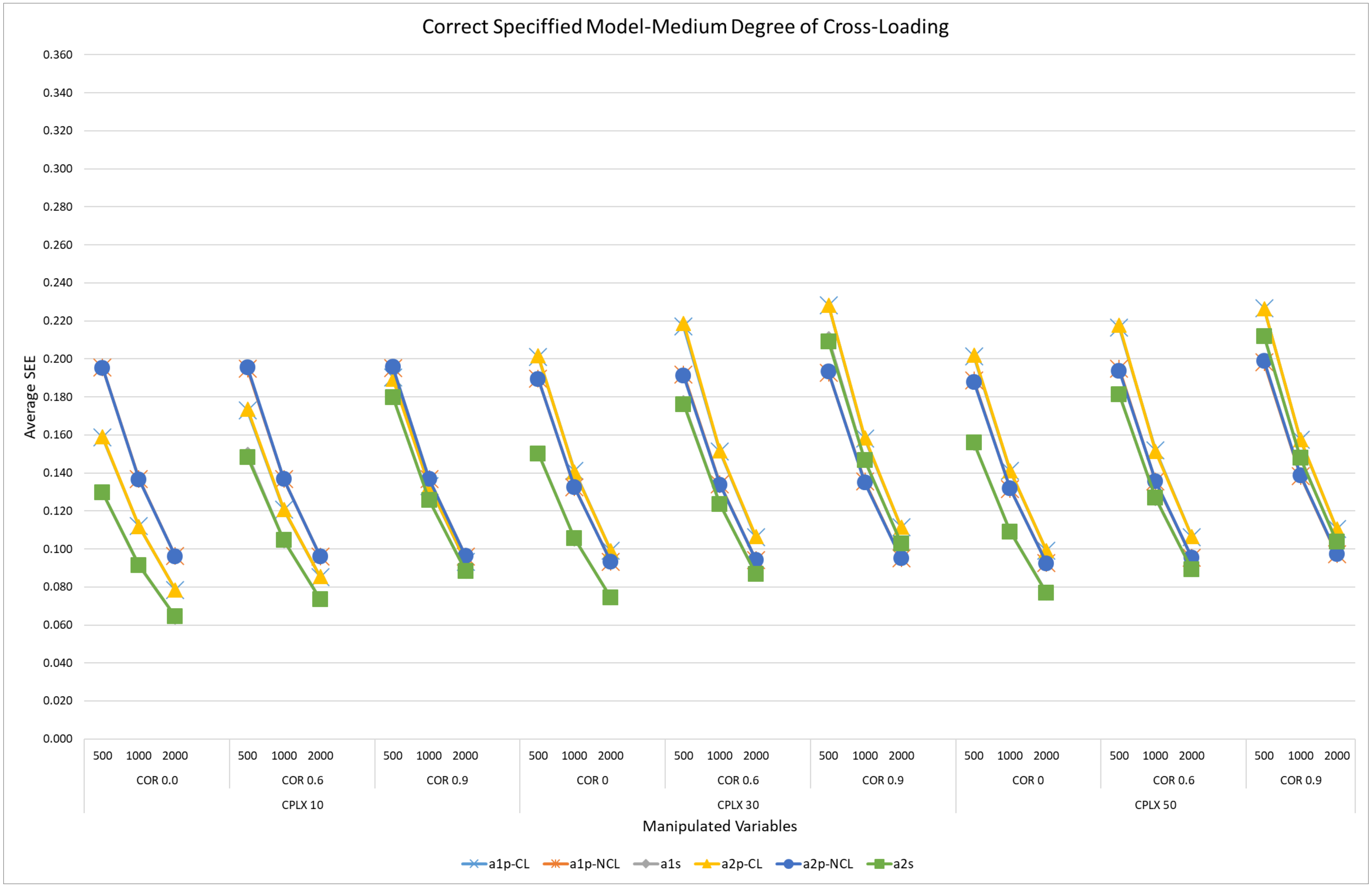

Figure 4. 18. Average SEE for item discrimination parameters when models were specified correctly with a medium degree of cross-loading

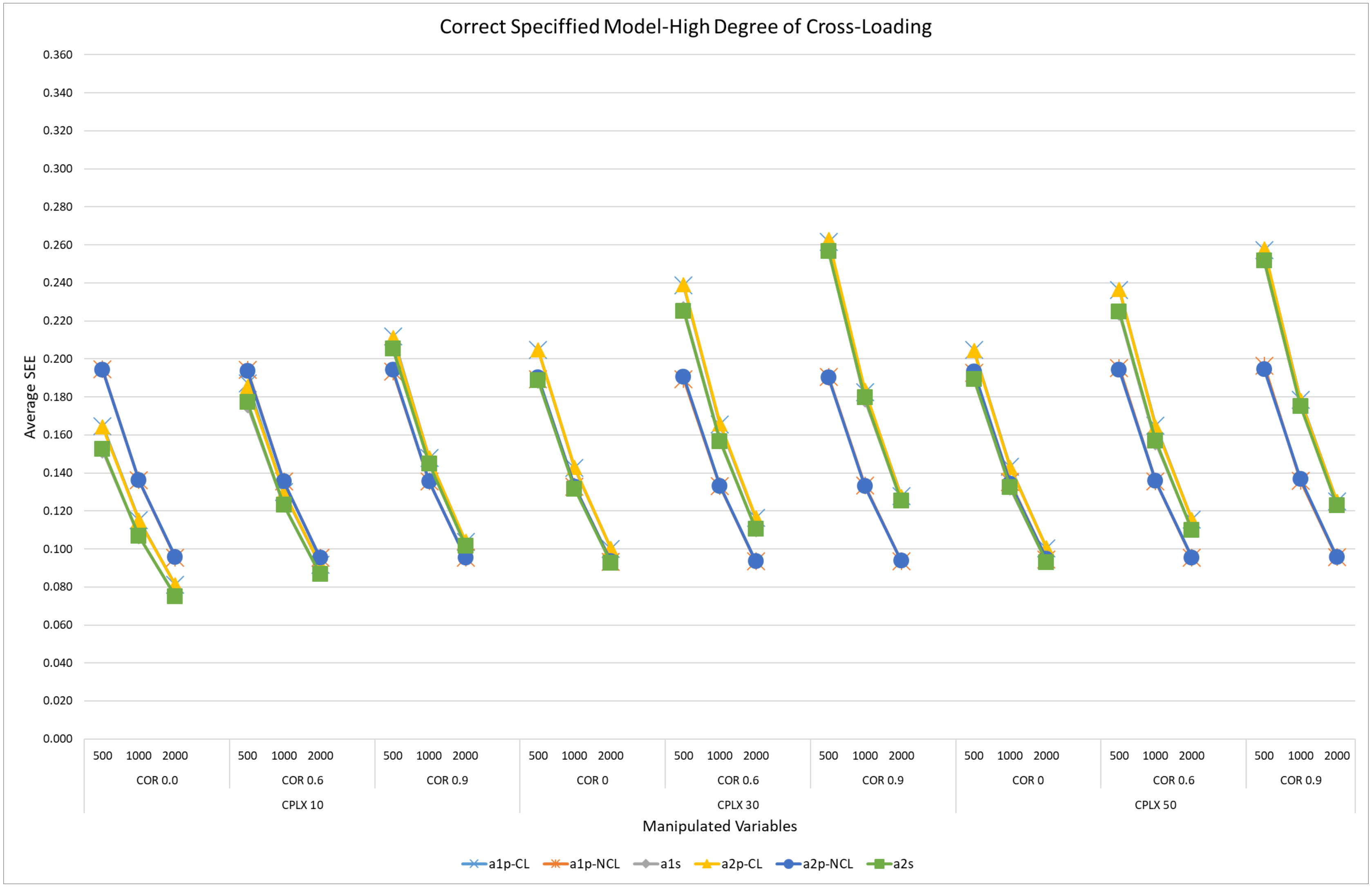

Figure 4. 19. Average SEE for item discrimination parameters when models were specified correctly with a high degree of cross-loading

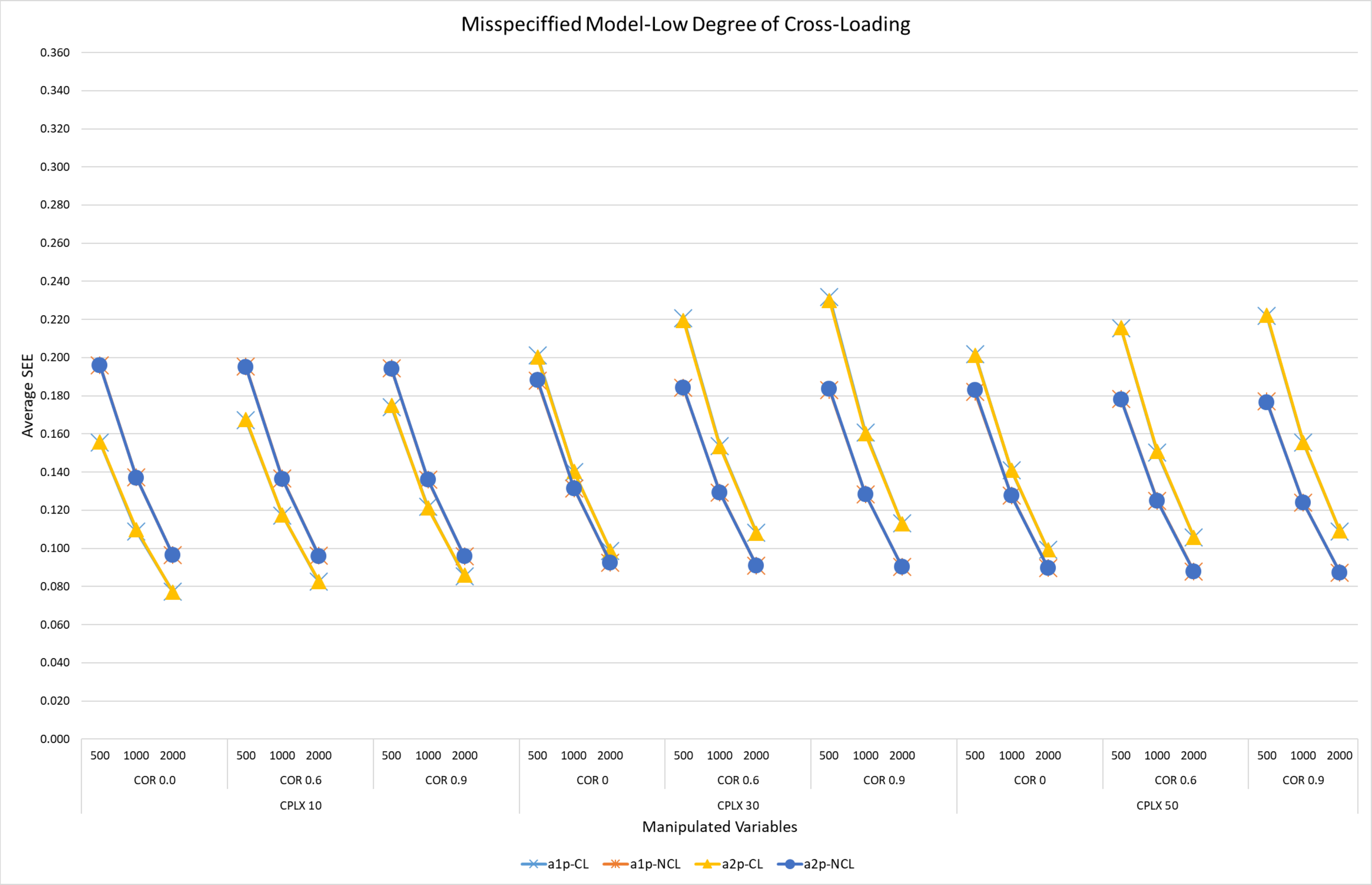

Figure 4. 20. Average SEE for item discrimination parameters when models were misspecified with a low degree of cross-loading

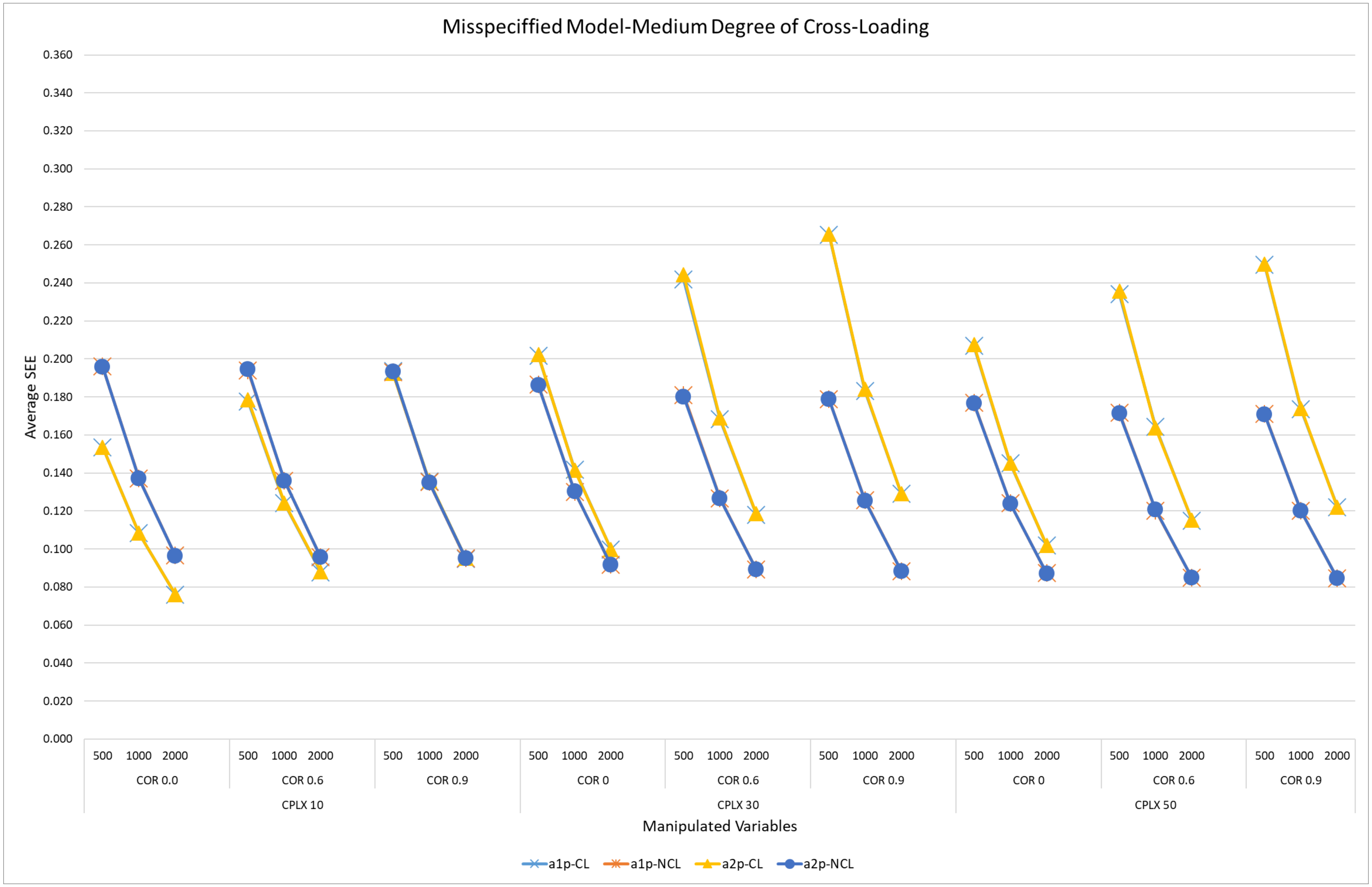

Figure 4. 21. Average SEE for item discrimination parameters when models were misspecified with a Medium degree of cross-loading

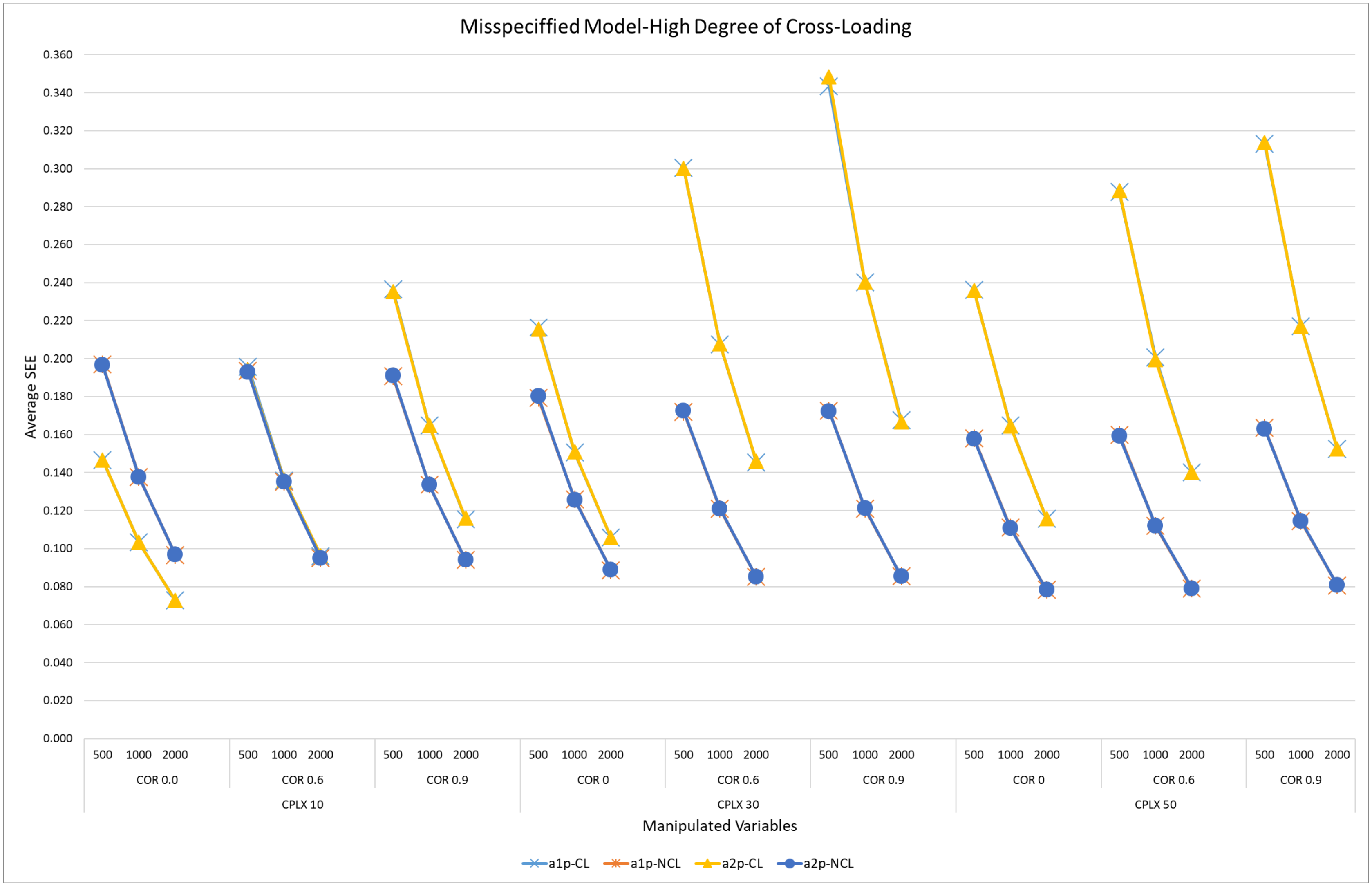

Figure 4. 22. Average SEE for item discrimination parameters when models were misspecified with a high degree of cross-loading

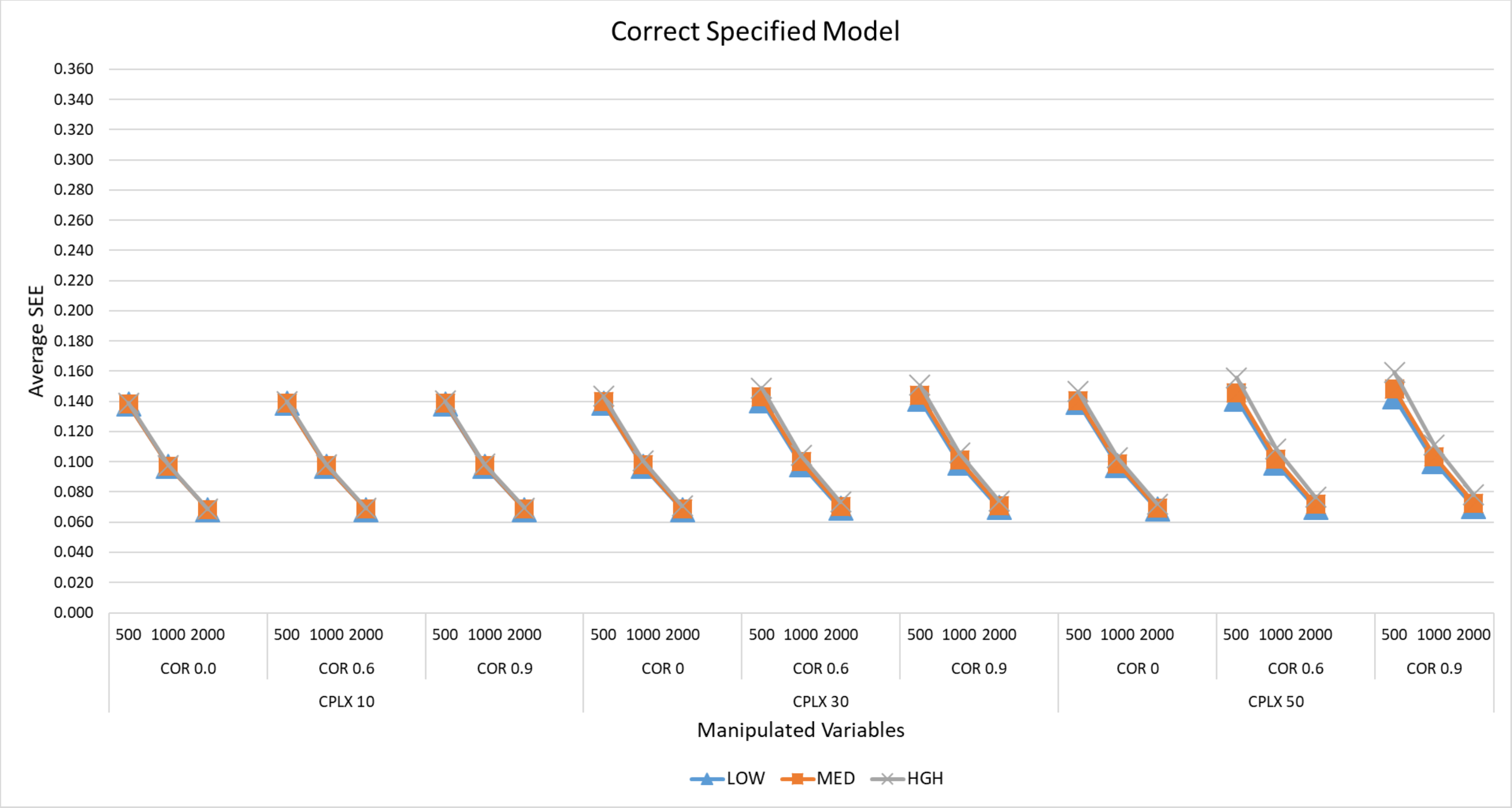

Figure 4. 23. Average SEE for item location parameter when models were specified correctly.

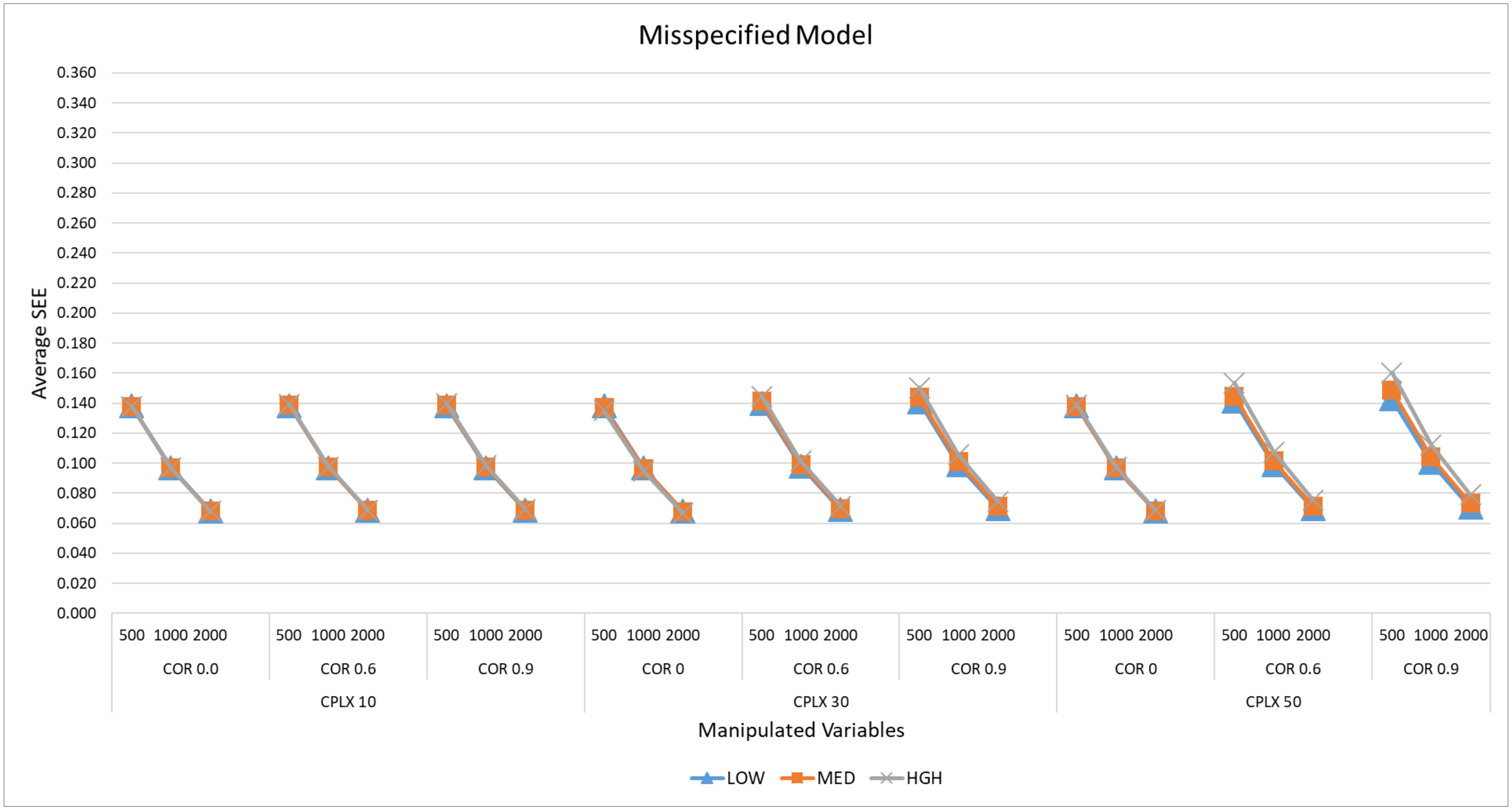

Figure 4. 24. Average SEE for item location parameter when models were misspecified.

CHAPTER V

DISCUSSION

Previous studies have investigated the effects of structure complexity (i.e., the proportion of items that are cross-loaded), the correlation between the underlying latent abilities, sample size, distribution of examinees on dimensionality assessment and item parameter recovery on complex structure MIRT models (Finch, 2011; Svetina & Levy, 2016; Svetina et al., 2017; Zhang 2012). However, these did not consider the degree of cross-loading on secondary dimension (how strongly each ability is being measured with an item, specifically the secondary dimension) and model specification (misspecified simple structure model when the data are truly complex) and their effects on item parameter estimation precision. In addition, previous studies did not collectively discuss the effects of structure complexity magnitude of the data incorporating the degree of cross-loading on secondary dimension and model specification on item parameter estimation in complex structure MIRT models.

The purpose of this study was to investigate the influence of structure complexity magnitude of the data incorporating the degree of cross-loading on secondary dimensions and model specification, especially when the model was specified as simple structure, ignoring the cross-loading, while the data are truly complex on item parameter recovery in MIRT models.

In real-world situations, multidimensional data may appear on an educational test or attitudinal survey. For example, a math test measuring algebra and geometry. Designers of the test may structure 10 items to measure algebra and 10 items to measure geometry. However, one or more of the primarily algebra items may require some secondary geometry knowledge to answer the item correctly, or some of the primarily geometry items may require some secondary algebra knowledge to answer the item correctly. In these situations, a simple structure, multidimensional model may be used to evaluate the items, ignoring the cross-loading of some items on the secondary dimension. By utilizing simulated data which replicated this scenario and then applying a simple structure model, i.e., misspecifying the model and ignoring the cross-loading, the potential consequences of the misspecification was investigated.

The primary research question of interest in this study was: what are the effects of structure complexity magnitude of the data, the degree of cross-loading on secondary dimension, and model specification on item parameter estimation in MIRT models? In order to address the research question a simulation study was designed to manipulate the variables that potentially influence the precision of item parameter estimation in the MIRT models. The manipulated variables in this study were three levels of structure complexity magnitude, three levels of the degree of cross-loading on secondary dimension, model specification, three levels of sample size and three levels of correlation between abilities. The described manipulated variables influencing the precision of item parameter estimation in this study led to 162 simulated item response data sets. Parameters were estimated using marginal maximum likelihood (MML), utilizing the expectation-maximization (EM) algorithms. A compensatory 2PL-MIRT model with two dimensions and dichotomous item response type (Reckase, 1985) was used to simulate and calibrate the data for every replication for each combination of conditions. A standard bivariate normal distribution was considered for the distribution of abilities and each combination of conditions was replicated 500 times.

Summary of Results

Effect of Structure Complexity Magnitude (RMSE, Bias and SEE)

Effects of this may be seen in application when on one form of a math test, one (of 10, 10%) of the primarily algebra items requires some secondary geometry knowledge to answer the item correctly; but, on a second form, three (of 10, 30%) of the primarily algebra items requires some secondary geometry knowledge to answer the item correctly.

Correct specified models. When the model correctly took into account the secondary loadings of cross-loaded items, the average RMSE, bias and SEE of the item discrimination on cross-loaded items on primary dimension (a_{1pcl}, a_{2pcl}) and secondary dimension (a_{1s} and a_{2s}) increased (in absolute value for bias) as structure complexity magnitude increased from 10% to 30%, but had little change when 50% of the items cross-loaded. The RMSE, bias and SEE of the non-cross loaded items on both dimensions ($a_{1pncl},\ a_{2pncl}$) had a little to no effect of the increases as the complexity magnitude increased from 10% to 50%.

Misspecified models. Item discrimination estimates when the model was misspecified and ignored the cross-loadings worsened compared to correctly specified models under most conditions of the structure complexity. The RMSE and SEE worsened as complexity increased from 10% to 30% to 50%, even when data were truly uncorrelated. Bias of the estimates of cross-loaded and non-crossloaded were adversely affected by increasing the proportion of complex items for correlated and uncorrelated data, holding all other variables constant; truly cross-loaded items had much more severe estimations.

Effect of the Degree of Cross-Loading (RMSE, Bias and SEE)

As it was mentioned before, one of the understudied variables that can affect the precision of item parameter estimation in MIRT models under complex structure of the data is the degree of cross-loading on secondary dimension. Effects of this may be seen in application when on one form of a math test, one or more of the primarily algebra items require very little secondary geometry knowledge to answer the item correctly; but, on a second form, the same cross-loaded primarily algebra items require a moderate or high secondary geometry knowledge to answer the item correctly. To our knowledge, none of the past studies regarding complex structure MIRT models addressed the possible influence of the degree of cross-loading incorporating model specification on item parameter estimation in MIRT models (e.g., Finch, 2011; Svetina et al., 2017; Zhang 2012). The results of this study indicated that item parameter recovery was influenced by the degree of cross-loading on secondary dimension.

Correct specified models. The average RMSE, bias and SEE of the item discrimination on cross-loaded items on both dimensions (a_{1pcl}, a_{2pcl}) and the secondary item discrimination parameters (a_{1s} and a_{2s}) had an increasing trend across the combinations of conditions as the degree of cross-loading increased from low to high. The average RMSEs for the primary item discrimination on non-cross loaded items on both dimensions (a_{1pncl}, a_{2pncl}) was not affected by changes in the degree of cross-loading increased from low to high.

Misspecified models. The average RMSE, bias and SEE for the item discrimination on cross-loaded items on both dimensions (a_{1pcl}, a_{2pcl}) was adversely affected when the degree of cross-loading increased from low to medium to high and the model ignored the cross-loading. The average RMSEs for the item discrimination on non-cross loaded items on both dimensions (a_{1pncl}, a_{2pncl}) was not affected by changes in the degree of cross-loading increasing from low to high. The RMSEs for the primary item discrimination parameters were much larger for the cross-loaded items compared to non-cross loaded items.

Effect of Sample Size (RMSE, Bias and SEE)

Previous studies have investigated the effect of a broad range of sample size from a small (500) to very large sample size (5,000) on the item parameter estimation in the MIRT models (e.g., Bolt & Lall, 2003; Finch, 2010, 2011; Zhang, 2012). Aligned with the previous studies considering MIRT models, the results of this study indicated that item parameter recovery performed better as the sample size increased.

Correct specified models. RMSE and SEE of the item discrimination parameters, including primary and secondary item discrimination ($a_{1pcl}, a_{1pncl}, a_{1s}\ a_{2pcl}, a_{2pncl}, a_{2s}$), had a consistent decreasing measure as the sample size increased from 500 to 1,000 to 2,000 across all combinations of conditions. Effects of changes in sample size had very little effect on average bias.

Misspecified models. Trends of RMSE, bias and SEE were similar for correct specified models and for misspecified models. RMSE and SEE of the item discrimination parameters, including primary and secondary item discrimination ($a_{1pcl}, a_{1pncl}, a_{1s}\ a_{2pcl}, a_{2pncl}, a_{2s}$), had a consistent decreasing measure as the sample size increased from 500 to 1,000 to 2,000 across all combinations of conditions. Effects of changes in sample size had very little effect on average bias.

A minimum sample size for item parameter estimation was not the primary focus of this study. However, aligned with the past studies (Bolt & Lall, 2003; Finch, 2010, 2011; Jing et al., 2016; Zhang, 2012) the results of this study indicated that as the number of examinees increased item parameter estimations were more accurate whether the model was correctly specified or misspecified. This means that in large-scale tests, test developers should consider a sufficient amount of sample size regarding item parameter estimations for designing multiple forms of a math test even if the item response data represent a complex structure (i.e., measuring algebra and geometry knowledge utilizing cross-loaded items).

Effect of Correlation (RMSE, Bias and SEE)

In practice, when a math test with algebra and geometry is administered, some may view the two subjects as related and correlated, and others may not. The data were simulated as being uncorrelated, having moderate correlation (.6) and high correlation (.9). When estimated, all conditions were assumed to be uncorrelated to replicate situations when the underlying correlation is ignored, even when dimensions may have moderate or high correlation.

Correct specified models. When other variables were held constant for each combination of conditions while the true correlation between abilities varied and the model did not account for correlation, the average RMSE, bias and SEEs increased consistently when items cross-loaded for the primary and secondary item discriminations on both dimensions $(a_{1pcl}, a_{1s}\ a_{2pcl}, a_{2s})$ as the correlation increased from .0 to .9. The non-cross loaded item discriminations $(a_{1pncl},\ a_{2pncl})$ had a consistent measure of bias across all levels of correlation.

Misspecified models. The average RMSE, bias and SEEs increased consistently when items cross-loaded for the primary and secondary item discriminations on both dimensions (a_{1pcl}, a_{2pcl}) as the correlation increased from .0 to .9. The non-cross loaded item discriminations $(a_{1pncl},\ a_{2pncl})$ had a consistent measure of bias across all levels of correlation.

Item Location

Sample size was the only variable that influenced item location estimation. As the sample size increased the RMSE and SEE of the item location decreased across all combinations of conditions. Item location parameter had a consistent measure of bias with no effect across all combinations of conditions. As the complex structure of multidimensional data in this study reflected more on item discrimination parameters (i.e., cross-loading), whether the model was correctly specified or misspecified, variables such as complexity magnitude, degree of cross-loading, and the correlation between abilities had no effect on item location parameter in terms of average RMSE, bias and SEE.

Summary of Results Compared to the Literature

It should be noted that it is difficult to compare the result of this study to those of Finch (2011), Zhang (2012) and Svetina et al. (2017) due to different structure complexity of the data in terms of item discrimination and item location specifications, structure complexity magnitude, the degree of cross-loading, model specification or the distribution of latent abilities. Comparison of the results of this study to studies of Finch (2011), Zhang (2012) and Svetina et al. (2017) should be made with caution as those studies focused on different combinations of influencing variables and conditions. For instance, Finch (2011) focused on the complex MIRT models when the distribution of latent abilities were non-normal. Zhang (2012) focused on comparing the precision of item parameter estimation in unidimensional and multidimensional estimation approaches within simple structure and mixed structure environments. This study primarily was an extension of Svetina et al. (2017) in which the authors focused on comparison of item parameter estimation under complex structure when the distribution of abilities were non-normal with balanced and imbalanced item discriminations but not incorporating the degree of cross-loading and model specification effects.

Implications and Suggestions

There are a great number of instruments, including surveys and tests that measure multiple latent abilities, which leads to a potentially multidimensional structure of item response data. In practice and in real-world situations, it is very likely that when the items in a test exhibit a complex structure with a strong loading on a primary ability and a small loading on a secondary ability (small degree of cross-loading), the data are treated as having a simple structure, ignoring the small cross-loading of some items.

The results of this study have implications for test or instrument developers and practitioners especially for those that are involved with complex structure multidimensional item response data. If the data are uncorrelated and the complex structure of the data is correctly specified, then changes in the degree of cross loading or percentage of complexity of the data has little effect on the estimations; most notably, a larger sample size produces better estimates and a smaller sample size produces worse estimates. If the data are correlated and the complex structure of the data is correctly specified but data are assumed to be uncorrelated, then changes in the degree of cross-loading has little effect on item discrimination estimations, but the percentage of complexity of the data does have an effect on the item discrimination estimations. Increases in percentage of complexity worsens as complexity increases from 10% to 30%, but having more than 30% complexity doesn't produce additional adverse effects.

It is very likely that ignoring complex structure of the multidimensional data severely impact item discrimination estimation that could result in biased item discrimination parameters. When the complex structure of the data is misspecified, whether data are correlated or uncorrelated, item discrimination parameters are adversely affected. If the data are uncorrelated and the complex structure of the data is misspecified, then a low or medium degree of cross loading and 10% or 30% of complexity of the data has little effect on the item discrimination estimations. However, as the degree of cross-loading increases or the percentage of complexity increases, the error and bias estimates of item discrimination worsen. Furthermore, if data are correlated and the correlation is not specified nor are the item cross-loadings, item discrimination estimates, specifically for the truly cross-loaded items has extremely poor error and bias estimates.

Under all circumstances, a larger sample size improves the item discrimination estimations, but even with a sample as large as 2,000, if cross-loadings and correlation are ignored and data are treated as having a simple structure, item discrimination estimates are severely adversely affected. This can ultimately result in inaccurate inferences regarding the examinees' abilities on each dimension. Therefore, it is imperative that test designers take variables such as complexity magnitude of the data incorporating the degree of cross-loading and model specification into account in order to have accurate inferences about the item parameter recovery and ultimately examinees' abilities on multiple dimensions.

First, test designers should be aware that the accuracy of item discrimination estimations could be potentially influenced by the structure complexity magnitude and correlation of the item response data. This means that as the number of cross-loaded items increases the item discrimination estimation might be less accurate. Test designers, should take into account the complexity magnitude of the data when designing multiple forms of the test. For example, in a math test one form might have three items out of the 10 (%30) that are cross-loaded (algebra and geometry) on both dimensions and the other form might have 5 items out of 10 (%50) that are cross-loaded (algebra and geometry) on both dimensions. Test designers should take into account the complexity structure of the data when designing multiple forms of a large-scale test.

Second, test designers should consider the fact that even if the model is correctly specified and the complexity magnitude is taken into account, still there are some effects of the degree of cross-loading on the secondary dimension on item discrimination estimation precision. For instance, one or some of the primarily algebra items may require some little, moderate or high secondary geometry knowledge to answer the item correctly (degree of cross-loading). Test developers should take into account the degree of cross-loading when designing multiple forms of a large-scale test.

Third, test designers should be aware of and cautious that utilizing a misspecified simple structure multidimensional model to evaluate the items, ignoring the cross-loading of some items on the secondary dimension could have serious consequences regarding item discrimination estimation accuracy. i.e., misspecifying the model and ignoring the cross-loading on the items that primarily measure algebra knowledge and require some geometry knowledge.

The results of this study support the conclusion that the complexity magnitude of the data incorporating the degree of cross-loading and model specification had an influence on the precision of item discrimination recovery that ultimately result in inappropriate inferences about the latent abilities of the examinees.

Limitations and Future Studies

In this simulation study, there exist some limitations that should be noted for future studies. First, the item response data were generated and analyzed utilizing R programing and simulation study techniques and it is possible that the results in real world situation differ when actual data from instruments such as tests and surveys are analyzed. Example of such factors could be the distribution of the examinees' abilities (i.e., depression), testing environment conditions, etc.

Second, item response data in this study were simulated considering the sample size of examinees similar to large-scale tests and surveys, and it is likely that the results differ when the number of examinees are relatively small. While many variables were manipulated within the context of this simulation study, correlation was not freely estimated; in future studies, allowing correlation to be freely estimated to (1) evaluate how well correlation is estimated and (2) understand how this may affect item parameter estimation, either better or worse.

For future studies, it should be noted that in addition to the variables manipulated in this study, there are a number of other variables that could influence the precision of item parameter recovery in MIRT models. For instance, a compensatory 2PL-MIRT model with two dimensions and dichotomous items response type was considered to simulate and calibrate the data for every replication for each condition combinations. It would be interesting to further investigate the effect of the manipulated variables in this study on other MIRT models, such as having more than two dimensions or bifactor and higher order models. In addition, it would be interesting to see how items would be recovered under a non-compensatory MIRT model considering the manipulated variables in this study.

Furthermore, in this simulation study a bivariate normal distribution was utilized for generating item response data and item recovery. However, in real world situation, the distribution of the latent ability is not always normal and it can influence item parameter recovery in MIRT models. It would be interesting to investigate violation of the normality assumption especially in the presence of skewed abilities considering structure complexity of the data incorporating the degree of cross-loading on secondary dimension and its influence on item parameter recovery.

As mentioned before, in this study, item parameters were estimated using marginal maximum likelihood (MML), utilizing the expectation-maximization (EM) algorithms. Marginal maximum likelihood (MML) method is considered generally as an effective estimation method with few dimensions (Chalmers, 2012; de Ayala, 2009; Stone, 1992). However, depending on the MIRT model specification and especially the distribution of the underlying abilities other estimation methods should also be considered for future studies.

Lastly, not only is IRT used for item evaluation, it is used for scoring respondents. Scores may be used to diagnose depression or anxiety, to classify respondents into groups, or to benchmark students in education. In any situation, it is imperative that scores provide accurate understandings of underlying abilities. Model misspecification, degree of cross loading, and structure complexity often affect item parameter estimates, which then is likely to affect estimated trait scores. Future analysis may also investigate the effects of these variables on estimated trait scores.

REFERENCES

Ackerman, P. L. (1992). Predicting individual differences in complex skill acquisition: Dynamics of ability determinants. *Journal of Applied Psychology*. 77(5), 598–614. https://doi.org/10.1037/0021-9010.77.5.598

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence*. 22(2), 227–257. https://doi.org/10.1016/S0160-2896(96)90016-1

Ackerman, T. A. (1987a). A comparison study of the unidimensional IRT estimation of compensatory and noncompensatory multidimensional item response data. *ACT Research Report Series*. (No. 87-12).

Adams, R. J., Wilson, M., & Wang, W. (1997). The Multidimensional Random Coefficients Multinomial Logit Model. *Applied Psychological Measurement*, 21(1), 1–23. https://doi.org/10.1177/0146621697211001

Ansley, T. N., & Forsyth, R. A. (1985). An Examination of the Characteristics of Unidimensional IRT Parameter Estimates Derived From Two-Dimensional Data. *Applied Psychological Measurement*, 9(1), 37–48. https://doi.org/10.1177/014662168500900104

Bolt, D. M., & Lall, V. F. (2003). Estimation of Compensatory and Noncompensatory Multidimensional Item Response Models Using Markov Chain Monte Carlo. *Applied Psychological Measurement*, 27(6), 395 414. https://doi.org/10.1177/0146621603258350

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48, 1-29.

Chen YF., Jiao H. (2013) Does Model Misspecification Lead to Spurious Latent Classes? An Evaluation of Model Comparison Indices. In: Millsap R.E., van der Ark L.A., Bolt D.M., Woods C.M. (eds) New Developments in Quantitative Psychology. Springer Proceedings in Mathematics & Statistics, vol 66. Springer, New York, NY. https://doi.org/10.1007/978-1-4614-9348-8_22

de Ayala, R. J. (2009). *The theory and practice of item response theory.* New York, NY: The Guilford Press.

Finch, H. (2006). Comparison of the Performance of Varimax and Promax Rotations: Factor Structure Recovery for Dichotomous Items. *Journal of Educational Measurement*, 43(1), 39–52. https://doi.org/10.1111/j.1745-3984.2006.00003.x

Finch, H. (2010). Item Parameter Estimation for the MIRT Model: Bias and Precision of Confirmatory Factor Analysis—Based Models. *Applied Psychological Measurement,* 34(1), 10–26. https://doi.org/10.1177/0146621609336112

Finch, H. (2011). Multidimensional Item Response Theory Parameter Estimation With Nonsimple Structure Items. *Applied Psychological Measurement*, 35(1), 67-82.

Finch, H., & Habing, B. T. (2005). Comparison of NOHARM and DETECT in item cluster recovery: Counting dimensions and allocating items. *Journal of Educational Measurement*, 42, 149-169.

Fox, G., Klein Entink, R., & Avetisyan, M. (2014). Compensatory and non-compensatory multidimensional randomized item response models. *British Journal of Mathematical & Statistical Psychology,* 67(1), 133–152. https://doi.org/10.1111/bmsp.12012

GRE (2018). Guide to the use of scores. ETS, Princeton, New Jersey.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Dow Jones-Irwin

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

Lord, F.M. (1980). Applications of Item Response Theory To Practical Testing Problems (1st ed.). Routledge. https://doi.org/10.4324/97802030566154

Matlock, K. L., & Turner, R. (2016). Unidimensional IRT Item Parameter Estimates Across Equivalent Test Forms With Confounding Specifications Within Dimensions. *Educational and Psychological Measurement*, 76(2), 258–279. https://doi.org/10.1177/0013164415589756

MCAT (2020). The MCAT Essential for Testing Year 2020. Association of American Medical Colleges

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

Reckase, M. D. (1985). The Difficulty of Test Items That Measure More Than One Ability. *Applied Psychological Measurement*, 9(4), 401 412. https://doi.org/10.1177/014662168500900409

Reckase, M. D. (1997). The Past and Future of Multidimensional Item Response Theory. *Applied Psychological Measurement*, 21(1), 25–36. https://doi.org/10.1177/0146621697211002

Reckase, M. D. (2009). Multidimensional Item Response Theory. Springer-Verlag. https://doi.org/10.1007/978-0-387-89976-3

Reckase, M. D., & McKinley, R. L. (1982). Some Latent Trait Theory in a Multidimensional Latent Space. *Item Response Theory and Computerized Adaptive Testing Conference Proceedings*, Wayzata, 27-30 July 1982.

RStudio Team (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Stone, C.A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, 16, 1-16.

Strachan, T., Ip, E., Fu, Y., Ackerman, T., Chen, S.-H., & Willse, J. (2020). Robustness of Projective IRT to Misspecification of the Underlying Multidimensional Model. *Applied Psychological Measurement*, 44(5), 362–375. https://doi.org/10.1177/0146621620909894

Svetina, D. (2013). Assessing dimensionality in noncompensatory MIRT with complex structure. *Educational and Psychological Measurement*, 73, 312–338.

Svetina, D., & Levy, R. (2016). Dimensionality in Compensatory MIRT When Complex Structure Exists: Evaluation of DETECT and NOHARM. *The Journal of Experimental Education*, 84(2), 398-420.

Svetina, D., Valdivia, A., Underhill, S., Dai, S., & Wang, X. (2017). Parameter Recovery in Multidimensional Item Response Theory Models Under Complexity and Nonnormality. *Applied Psychological Measurement*, 41(7), 530–544. https://doi.org/10.1177/0146621617707507

Sympson, J. B. (1978). A model for testing with multidimensional items. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Convergence (pp. 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Thurstone, L. L. (1947). Multiple factor analysis. Chicago, IL: University of Chicago Press.

TOEFL iBT® (2020). TOEFL iBT® Test Framework and Test Development. TOEFL® Research Insight Series, Volume 1. ETS.

Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika*, 72, 69–91.

Zhang, J. (2012). Calibration of Response Data Using MIRT Models With Simple and Mixed Structures. *Applied Psychological Measurement*, 36(5), 375–398. https://doi.org/10.1177/0146621612445904

Zhang, J., & Stout, W. (1999). The theoretical detect index of dimensionality and its application to approximate simple structure. *Psychometrika*, 64(2), 213–249. https://doi.org/10.1007/BF02294536

VITA

Mostafa Hosseinzadeh

Candidate for the Degree of

Doctor of Philosophy

Dissertation: EVALUATION OF STRUCTURE COMPLEXITY MAGNITUDE, DEGREE OF CROSS-LOADING ON SECONDARY DIMENSION AND MODEL SPECIFICATION ON MIRT PARAMETER ESTIMATION

Major Field: Educational Psychology

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Educational Psychology at Oklahoma State University, Stillwater, Oklahoma in December, 2021.

Completed the requirements for the Master of Science in Industrial and System Engineering at Islamic Azad University, Arak, Iran in 2010.

Completed the requirements for the Bachelor of Science in Biosystems Engineering at Shahid Bahonar University, Kerman, Iran in 2007.

Experience:

AUG 2019 – Present, Assessment Coordinator, University of Central Oklahoma.
AUG 2019 – Present, Graduate Teaching & Research Associate, Oklahoma State University.
AUG 2018 – AUG 2019, Statistical Analyst-GRA, Institutional Assessment, Oklahoma State University.
MAR 2018 – AUG 2018, Math Lab Associate, Oklahoma City Community College
Summer 2017, Instructor, Sooner Upward Bound, The University of Oklahoma.

Professional Memberships:

American Educational Research Association (AERA)