6-2017

Modelling Student Misconceptions Using Nested Logit Item Response Models

Mustafa Yildiz

Western Michigan University, mustafa.yildz@gmail.com

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Educational Assessment, Evaluation, and Research Commons

Recommended Citation

https://scholarworks.wmich.edu/dissertations/3123

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact wmu-scholarworks@wmich.edu.
MODELING STUDENT MISCONCEPTIONS USING
NESTED-LOGIT ITEM RESPONSE MODELS

Mustafa Yildiz, Ph.D.
Western Michigan University, 2017

Student misconceptions have been studied for decades from a curricular/instructional perspective and from the assessment/test level perspective. Numerous misconception assessment tools have been developed in order to measure students’ misconceptions relative to the correct content. Often, these tools are used to make a variety of educational decisions including students’ achievement level, instructional method effectiveness, and curriculum related achievement progress. These tools have included qualitative and quantitative assessment methods.

The quantitative analysis of misconceptions has mostly relied on classical test theory methods of test construction related to total raw score, percentage of correct responses, and/or percentage of misconception responses. More recently, researchers have begun to use modern test theory methods of test construction including item response theory and cognitive diagnostic models to assess misconceptions. However, to date, there has not been any test construction modeling that has scaled a student’s ability estimate and a student’s misconception level into a continuous metric.

The purpose of this study was to investigate if it is possible to model misconceptions, which in the latent framework have been only measured using a latent class approach, as single
or multiple factor continuous latent variables in addition to a latent variable of interest, and see if modeling misconceptions help provide additional test information. Bayesian (Markov Chain Monte Carlo, (MCMC) methods were used to estimate model parameters. This study investigated if test length, number of misconceptions, and the prior distribution specification affected model convergence, parameter estimation precision, and the value-added impact gained by the modeling of student misconceptions.

The findings indicated that overall estimation precision was satisfactory for both item and person parameters when single factor misconception was used however increasing the number of misconceptions reduces estimation precision. Increasing the number of distractors measuring misconceptions increases the test information related to the misconception.

Future research might consider test lengths other than 25 or 50 as well as different sizes of sample used in this study. The framework provided by this study could inform and guide the misconception instrument development processes.
ACKNOWLEDGEMENTS

I would like to thank to the most important two people who have had supported me during my studies, my mother and my father, Fatma and Ahmet Yildiz, for their endless encouragement and support.

Second, I would like to thank to my advisor, Dr. E. Brooks Applegate, for his guidance during my studies here at Western Michigan University. I also would like to thank to the committee members of my dissertation; Dr. Fernando H Andrade Adaniya, and Dr. Rajib Paul for their support and guidance.

Lastly, I would like to thank my friends at WMU; Jason, Ran, Diyana, Deyab, Yu, and other guys for the collaboration during our studies. I would like to send my special thanks my uncle, sister and brothers; Fazli, Emin, Elif, Saim, and Muhammed YILDIZ. I always remember the love of the four little people, Ahmet, Omer, Meleknur, and Zeynep.

Mustafa Yildiz
TABLE OF CONTENTS

ACKNOWLEDGEMENTS.................................................................................................................... ii

LIST OF TABLES.............................................................................................................................. viii

LIST OF FIGURES............................................................................................................................ ix

LIST OF EQUATIONS....................................................................................................................... xiii

CHAPTER

I. INTRODUCTION .......................................................................................................................... 1

   Statement of the Problem....................................................................................................... 1

   Background.................................................................................................................................. 7

       Item response theory......................................................................................................... 7

       Nested logit item response models ................................................................................... 9

   Test information functions ................................................................................................. 10

       Bayesian estimation approach......................................................................................... 11

   Proposed Study .................................................................................................................... 12

   Significance of the study ..................................................................................................... 16

   Research Questions ............................................................................................................ 16
**Table of Contents – Continued**

**CHAPTER**

II. REVIEW OF LITERATURE .................................................................................................................. 19
   A general overview of the epistemology of misconception research ............................... 19
   Piagetian epistemology ................................................................................................................. 20
   The philosophy of science ............................................................................................................. 21
   Systematic errors ............................................................................................................................ 21
   A chronological synopsis of the past of quantitative misconception studies .......... 26
   Misconception Modeling ............................................................................................................... 31
   Nested logit item response models ............................................................................................. 34
   Bayesian and ML item response models ....................................................................................... 35

III. METHODOLOGY ............................................................................................................................ 37
   The design of data simulation ....................................................................................................... 37
   Number of examinees ..................................................................................................................... 39
   Number of items ............................................................................................................................. 39
   IRT parameter values ................................................................................................................... 41
   Misconception level ....................................................................................................................... 41
   Use of non-informative priors ....................................................................................................... 42
   Simulation evaluation .................................................................................................................... 43
   Convergence assessment ............................................................................................................. 43
   Item parameter bias and RMSE .................................................................................................... 44
Table of Contents – Continued

CHAPTER

Person parameter bias and RMSE................................................................. 45
Average absolute error .............................................................................. 47
Additional test information from modeling misconceptions .................. 48

IV. RESULTS ................................................................................................. 50

RQ 1: What is the degree of estimation precision in a 2PL-2PL-NLMM when a single LVI and MLV are modeled?.......................................................... 50

RQ 1.1: What are the convergence rates when 2PL-2PL-NLMM was used with a single MLV? ................................................................. 51

RQ 1.2: What is the degree of item parameter estimation bias when 2PL-2PL-NLMM was used with a single MLV? ................................................. 54

RQ 1.3: What is the degree of average absolute errors when 2PL-2PL-NLMM was used with a single MLV? .......................................................... 61

RQ 1.4: What is the degree of person parameter bias when 2PL-2PL-NLMM was used with a single MLV? .......................................................... 66

RQ 1.5: What is the test information contribution of modeling a single MLV? ...... 71

RQ 2: What is the degree of estimation precision in a 2PL-2PL-NLMM when a single LVI and two MLVs are modeled? ............................................. 75

RQ 2.1: What are the convergence rates when 2PL-2PL-NLMM was used with two MLVs? ................................................................. 77
# Table of Contents – Continued

**CHAPTER**

RQ 2.2: What is the degree of item parameter estimation bias when 2PL-2PL-NLMM was used with two $MLVs$? ................................................................. 80

RQ 2.3: What is the degree of average absolute errors when 2PL-2PL-NLMM was used with two $MLVs$? ................................................................. 92

RQ 2.4: What is the degree of person parameter bias when 2PL-2PL-NLMM was used with two $MLVs$? ................................................................. 98

RQ 2.5: What is the test information contribution of modeling two $MLVs$? ...... 107

RQ 3: What is the degree of estimation precision in a 2PL-2PL-NLMM when a single LVI and three MLVs are modeled? ......................................................... 114

RQ 3.1: What are the convergence rates when 2PL-2PL-NLMM was used with two $MLVs$? ................................................................. 115

RQ 3.2: What is the degree of item parameter estimation bias when 2PL-2PL-NLMM was used with three $MLVs$? ................................................................. 118

RQ 3.3: What is the degree of average absolute errors when 2PL-2PL-NLMM was used with three $MLVs$? ................................................................. 131

RQ 3.4: What is the degree of person parameter bias when 2PL-2PL-NLMM was used with three $MLVs$? ................................................................. 136

RQ 3.5: What is the test information contribution of modeling three $MLVs$? ...... 149

V. CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS ................................................. 156
Table of Contents – Continued

CHAPTER

Discussion.................................................................................................................................. 161
Limitations.................................................................................................................................. 165
Recommendations for Future research ...................................................................................... 166

REFERENCES.................................................................................................................................. 168

APPENDICES.................................................................................................................................. 178

A.............................................................................................................................................. 178
B.............................................................................................................................................. 188
C.............................................................................................................................................. 192
# LIST OF TABLES

1. Notation Table .................................................................................................................................................. 37

2. The decomposition of distractors for the test length of 25 by the dimensionality of $MLV_s$ .................................................................................................................. 40

3. The decomposition of distractors for the test length of 50 by the dimensionality of $MLV_s$ .................................................................................................................. 40

4. Distributional characteristics of parameters for data generation, and prior specification ........................................................................................................................................... 42

5. Observed responses, and corresponding ability estimates for $LVI$ and $MLV$ ........................................................................................................................................ 52

6. Percent non-Convergence Assessment * ........................................................................................................ 53

7. RMSE for item parameter estimates ................................................................................................................ 55

8. Average absolute error for person parameter estimates ................................................................................. 62

9. Person parameter bias and RMSE* .................................................................................................................. 67

10. Observed responses, and corresponding ability estimates for $LVI$, $MLV_1$, and $MLV_2$ .......................................................................................................................................... 77

11. Percent non-Convergence Assessment* ....................................................................................................... 78

12. RMSE for item parameter estimates ................................................................................................................ 81

13. Average absolute error for person parameter estimates ................................................................................. 93

14. Person parameter bias and RMSE* .................................................................................................................. 99

15. Observed responses, and corresponding ability estimates for $LVI$, $MLV_1$, $MLV_2$, and $MLV_3$ .......................................................................................................................................... 115

16. Percent non-Convergence Assessment* ....................................................................................................... 116

17. RMSEs for the item parameter estimates ....................................................................................................... 119

18. Average absolute error for person parameter estimates ................................................................................. 132

19. Person parameter bias and RMSE* .................................................................................................................. 137
### LIST OF FIGURES

1. An illustration of the hierarchical structure of the model in terms of how the probability of a misconception response is estimated .......................................................... 15
2. An abstraction for measurement model of a NLMM with 3 MLVs ................................. 38
3. Study design (Design cube) .......................................................................................... 43
4. Example MCMC plots for five distinct parameters ....................................................... 54
5. Estimation Bias for Discrimination Parameter of LVI (a1) ........................................... 56
6. Estimation Bias for Difficulty Parameter of LVI (b1) .................................................. 57
7. Estimation Bias for Difficulty Parameter of MLV (b2) ................................................ 59
8. Estimation Bias for Discrimination Parameter of MLV (a2) ........................................ 61
9. Average Absolute Error for the Shorter Test Length (25 items) .................................. 63
10. Average Absolute Error for the Longer Test Length (50 items) .................................. 65
11. Bias for the estimates of LVI and MLV when a short test was used ............................. 69
12. RMSE for the estimates of LVI and MLV .................................................................... 71
13. Illustrative Test Information Functions for Normal Distribution Conditions ............... 72
14. Illustrative Test Information Functions for Uniform Distribution Conditions ............ 73
15. The peak of the test information functions by replications ......................................... 75
16. Example MCMC history plots for seven distinct parameters ....................................... 80
17. Bias values for the location parameter of the LVI ....................................................... 83
18. Bias values for the discrimination parameter of LVI .................................................. 85
19. Bias values for the location parameters of MLV_1 .................................................... 86
20. Bias values for the location parameters of MLV_2 ..................................................... 87
21. Bias values for the discrimination parameters of MLV_1 ........................................... 89
22. Bias values for the discrimination parameters of MLV_2 .......................................... 91
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>Average absolute error for short test conditions when MLV is 2-dimensional</td>
<td>95</td>
</tr>
<tr>
<td>24</td>
<td>Average absolute error for long test conditions when MLV is 2-dimensional</td>
<td>97</td>
</tr>
<tr>
<td>25</td>
<td>Bias for the estimates of LVI and MLV when short tests used</td>
<td>101</td>
</tr>
<tr>
<td>26</td>
<td>Bias for the estimates of LVI and MLV when long tests used</td>
<td>103</td>
</tr>
<tr>
<td>27</td>
<td>RMSE for the estimates of LVI and MLV when short tests used</td>
<td>105</td>
</tr>
<tr>
<td>28</td>
<td>RMSE for the estimates of LVI and MLV when long tests used</td>
<td>107</td>
</tr>
<tr>
<td>29</td>
<td>Illustrative Test Information Functions for Normal Distribution Conditions (25 items)</td>
<td>108</td>
</tr>
<tr>
<td>30</td>
<td>Illustrative Test Information Functions for Normal Distribution Conditions (50 items)</td>
<td>109</td>
</tr>
<tr>
<td>31</td>
<td>Illustrative Test Information Functions for Uniform Distribution Conditions (25 items)</td>
<td>110</td>
</tr>
<tr>
<td>32</td>
<td>Illustrative Test Information Functions for uniform Distribution Conditions (50 items)</td>
<td>111</td>
</tr>
<tr>
<td>33</td>
<td>Maximum amount of information provided by replications</td>
<td>113</td>
</tr>
<tr>
<td>34</td>
<td>Example MCMC history plots for eight distinct parameters</td>
<td>118</td>
</tr>
<tr>
<td>35</td>
<td>Estimation bias of the difficulty of LVI when there were 3 MLVs</td>
<td>120</td>
</tr>
<tr>
<td>36</td>
<td>Estimation bias of the discrimination parameter of the LVI</td>
<td>121</td>
</tr>
<tr>
<td>37</td>
<td>Estimation bias for the discrimination parameter of MLV₁ (a2)</td>
<td>123</td>
</tr>
<tr>
<td>38</td>
<td>Estimation bias for the discrimination parameter of MLV₂ (a3)</td>
<td>124</td>
</tr>
<tr>
<td>39</td>
<td>Estimation bias for the discrimination parameter of MLV₃ (a4)</td>
<td>126</td>
</tr>
<tr>
<td>40</td>
<td>Estimation bias for the difficulty parameter of MLV₁ (b2)</td>
<td>127</td>
</tr>
<tr>
<td>41</td>
<td>Estimation bias for the difficulty parameter of MLV₂ (b3)</td>
<td>128</td>
</tr>
<tr>
<td>42</td>
<td>Estimation bias for the difficulty parameter of MLV₃ (b4)</td>
<td>130</td>
</tr>
<tr>
<td>43</td>
<td>Average absolute error for short test conditions when MLV is 3-dimensional</td>
<td>133</td>
</tr>
<tr>
<td>44</td>
<td>Average absolute error for long test conditions when MLV is 3-dimensional</td>
<td>135</td>
</tr>
</tbody>
</table>
List of Figures – Continued

45. Bias for the estimates of LVI and MLV when short tests were used (normal prior specification) ................................................................................................................... 139

46. Bias for the estimates of LVI and MLV when short tests were used (uniform prior specification) ................................................................................................................... 141

47. Bias for the estimates of LVI and MLV when long tests were used (normal prior) ...... 142

48. Bias for the estimates of LVI and MLV when long tests were used (uniform prior) ..... 144

49. RMSE for the estimates of LVI and MLV when short tests used ................................. 147

50. RMSE for the estimates of LVI and MLV when long tests used ............................... 149

51. Illustrative Test Information Functions for Normal Distribution Conditions (25 items) 150

52. Illustrative Test Information Functions for Normal Distribution Conditions (50 items) 151

53. Illustrative Test Information Functions for Uniform Distribution Conditions (25 items) 152

54. Illustrative Test Information Functions for Uniform Distribution Conditions (50 items) 153

55. Maximum amount of information by replications ...................................................... 155

56. Item/distractor characteristics curves as functions of latent variables when there was an LVI and an MLV ...................................................................................................................... 188

57. Item/distractor characteristics curves as functions of latent variables when there was an LVI and two MLVs ...................................................................................................................... 189

58. Item/distractor characteristics curves as functions of latent variables when there was an LVI and three MLVs ...................................................................................................................... 191

59. Trait correlations when there was a single MLV ....................................................... 193

60. Trait correlations for short test forms when there were two MLVs (Normal prior) ...... 194

61. Trait correlations for long test forms when there were two MLVs (Normal prior) ...... 194

62. Trait correlations for short test forms when there were two MLVs (Uniform prior)..... 195

63. Trait correlations for long test forms when there were two MLVs (Uniform prior)..... 196

64. Trait correlations for short test forms when there were three MLVs (Normal prior) ... 197
List of Figures – Continued

65. Trait correlations for long test forms when there were three MLVs (Normal prior) ..... 198
66. Trait correlations for short test forms when there were three MLVs (Uniform prior) .. 199
67. Trait correlations for long test forms when there were three MLVs (Uniform prior).... 200
## LIST OF EQUATIONS

1. 1PL IRT MODEL ............................................................................................................................ 8
2. 2PL IRT MODEL ........................................................................................................................... 8
3. NESTED LOGIT ITEM RESPONSE MODEL .................................................................................. 10
4. ITEM INFORMATION FUNCTION .............................................................................................. 10
5. TEST INFORMATION FUNCTION ............................................................................................. 11
6. 2PL-2PL-NLMM ......................................................................................................................... 14
7. LIKELIHOOD FUNCTION FOR 2PL-2PL-NLMM ........................................................................... 14
8. LOG-LIKELIHOOD FUNCTION FOR 2PL-2PL-NLMM ................................................................. 14
9. BOCK’S (1972) NOMINAL RESPONSE MODEL ........................................................................... 32
10. SICM MODEL ........................................................................................................................... 33
11. COVARIANCE MATRIX FOR LATENT VARIABLES ..................................................................... 41
12. RMSE FOR ITEM PARAMETER ESTIMATES.............................................................................. 45
13. BIAS FOR PERSON PARAMETER ESTIMATES .......................................................................... 46
14. RMSE FOR PERSON PARAMETER ESTIMATES .......................................................................... 46
15. AVERAGE ABSOLUTE ERROR .................................................................................................. 48
16. TEST INFORMATION FUNCTION ............................................................................................. 48
17. COLLATERAL INFORMATION DUE TO MODELING MISCONCEPTIONS ................................. 49
CHAPTER I

INTRODUCTION

Statement of the Problem

Learning *misconceptions*\(^1\) are a concern for teachers and have been discussed in the testing literature for over four decades (Smith, diSessa, & Roschssle, 1993\(^2\); Confery, 1990; Hestenes, Wells & Swackhamer, 1992) and have been described many different ways in the literature (Bell, Swan, & Taylor, 1981; Clement, 1981; Halloun & Hestenes, 1985). Attempts to define misconceptions usually start with the idea that students come into a classroom with some pre-experience or knowledge on the topics taught (Bell, Swan, & Taylor, 1981). Caramazza, McCloskey, and Green (1981) illustrated the assumption of (pre)existing student knowledge in a study related to motion in which real world experiences played a crucial role for the existence of a misconception. A large number of studies has shown that children develop their own understanding of nature, such as how the physical world works which gets revised and reinterpreted with the new information that is experienced or delivered via a classroom instruction. At any time, a student’s understanding of their world may not always be compatible with scientific consensus (Mulford & Robinson, 2002). The resulting misunderstanding of the phenomena can be expressed as *misconception*.

---

\(^1\) Key words in this dissertation are shown in italic font and are defined in the Definition section.
\(^2\) All citations in this dissertation follow the *APA Publications Manual, 6*\(^{th}\) ed.
For example, the algebraic expression \((x+7)/(x+9)\) asking student to find a simpler form sometimes gets answered as \(7/9\), which is not just incorrect but also is likely to happen again when a student faces a similar equation problem. Another example from physics is that some students believe that an object would be moving at a constant speed if it were under a stable force. This is also an incorrect response while being a misconception.

Previous research on misconceptions has crossed a variety of disciplines including: Physics (Clement, 1981; Caramazza, McCloskey, & Green, 1981; Halloun & Hestenes, 1985; Kolcak, Mogol, & Unsal, 2014); mathematics (Halloun & Hestenes, 1985; Bell, Swan, & Taylor, 1981), electricity (Pesman & Eryilmaz, 2010; Turgut, Gurbuz, & Turgut, 2011); astronomy (Favia, Comins, & Thorpe, 2013; Sadler, Coyle, Miller, Cook-Smith, Dussault, & Gould, 2010); chemistry (Mulford & Robinson, 2002); and statistics (Jendraszek, 2008; Khazanov, 2008). It has been named in different ways as its conceptualization developed over time: For example, preconceptions (Clement, 19822; Glaser, Bassok, 1989; Wiser, 1989); alternative conceptions (Hewson & Hewson, 1984); naïve beliefs (McCloskey, Caramazza, & Green, 1981); alternative beliefs (Wiser, 1989); alternative frameworks (Driver, 1983; Driver, & Easley, 1978); naïve theories (McCloskey, 1983; Resnick, 1983); and misconceptions (Smith, diSessa, & Roschlsle, 1993). The term *misconception* will be used in this dissertation as a concept that encompasses all the above different labels.

Bell, Swan, and Taylor (1981) studied a variety of misconceptions common in less-able 14-year-olds. For instance, misconceptions included interpretation of decimal places, result of a multiplication and/or division, units associated with numbers, and familiarity with operator names. A very common example of a misconception for this group was “multiplication makes
bigger and division smaller” (p. 405). A test question asking how much it would cost for 8.6 gallons of gas when the gas price was 1.17 versus when the gas price was 0.22 made a difference on student responses because of the misconception that “multiplication always make bigger” (p. 405). When a student has the misconception “multiplication always make bigger”, her/his total price expectation for cost of the 8.6 gallon of gas is a number that is greater than 8.6 since she/he has to multiply 8.6 with the cost of 1 gallon of gas. When the cost per gallon is a value which is less than 1$, students are more likely to have difficulty understanding why the resulting value is less than 8.6 because of the belief that makes them to think that “multiplication always make bigger”.

Similarly, in physics, Hestenes, Wells, and Swackhamer (1992) developed a well-known inventory, Force Concept Inventory (FCI) to assess misconceptions related to Newtonian concepts in physics. The inventory investigates the misconceptions in the following content domains: kinematics, impetus, active force, action/reaction pairs, concatenation of influences, and motion. For example, students sometimes do not have a well-developed understanding of kinematics in which they cannot discriminate velocity from acceleration. The researchers stated that students often get confused among the concepts such as “force”, “energy”, “power”, and often they use them interchangeably. Another example of misconception found in FCI is “heavier objects fall faster” (p. 154).

Khazanov (2008) studied misconceptions related to learning probability concepts and reported that there were seven very common misconceptions. One of them was representativeness bias that is when estimating the likelihood of a dichotomous outcome with two events as in the case of tossing a coin, students often get confused with the
representativeness of the outcome in the population. For example, the likelihood that six times tossing a coin would result in HHHTTT vs TTTTTT is assumed in the favor of HHHTTT just because 50-50 outcome is a better representation of the population (head or tail). In reality both, HHHTTT and TTTTTT have the same likelihood.

An important characteristic of misconceptions distinguishes them from simple errors (Khazanov, 2008). Simple errors would be simply due to any of carelessness, guessing, answer copying, fatigued responding, creative responding, random responding, speeded responding, anxiety, high or low motivation (Rupp, 2013). Misconceptions, on the other hand, produce errors systematically due to its existence in the respondent’s misunderstanding of the nature (Smith, diSessa, & Roschslle, 1993). From a student perspective, in the case of chemistry education, if a student has some degree of misconception on a topic, the new information being presented by the instruction may be “ignored, rejected, disbelieved, deemed irrelevant, or held for consideration for a later time” (Mulford & Robinson, 2002). Caramazza, McCloskey, and Green (1981) reported that typical classroom instruction does not always yield a true understanding of basic principles of motion in physics, and the instruction that does not take the idea of misconception into account could be responsible for the failure of student learning problems. Clement (1981) pointed out that students start with real world experiences to develop their conceptual understanding on physics and that psychological state that is very resistant to change. However, Smith, diSessa, & Roschslle, (1993) disagreed with Clement (1981) because that view assumes a discontinuity between the student beliefs and the expert scientist consensus. According to Smith, diSessa, & Roschslle, (1993), misconceptions are not
the flawed things to be replaced, instead, these are knowledge elements to be refined, and reorganized.

Previous research on misconceptions has treated the misconception as something that is held by a student; or that is developed/thought by the student; or somehow believed by the student that does not align with scientific consensus. The way that student believes these ideas is coherent on its own and make sense to the student, but unfortunately, lead to incorrect outcomes. The inferences made from simple misconceptions to more complex ideas are artifacts of the misconceptions. A decision made by student in choosing an incorrect response during a test can happen in two ways. One is that some external unsystematic thing happens to make the student to choose that incorrect option (carelessness, anxiety, fatigue, motivation); or there is some belief, thought, or experience that makes student to believe in a certain way was which is likely to remain unchanged.

While researchers have tried to define and to understand misconceptions (Confrey, 1990; Smith, diSessa, & Roschelle, 1993) they have also tried to numerically express or quantify misconception (Bradshaw & Templin, 2014; Sadler, 1998; Khazanov, 2008; Hestenes, Wells, & Swackhamer, 1992). For instance, some efforts were made to quantify misconceptions that include classical test theory (Crocker and Algina, 1986) related approaches. Mostly, these efforts included either the percentage of correct answers versus percentage of answers that reflected misconception, or the change of misconception selection percentages from a pre-test to a post-test (Mulford & Robinson, 2002; Hestenes, Wells & Swackhamer, 1992; Khazanov, 2008). Some other studies employed item response theory (IRT) models. For instance, Sadler (1998) attempted use Bock’s (1972) nominal response model with option characteristic curve
(OCC) to understand how each item choice that measured a misconception looked like with respect to the true latent trait continuum. This model offers visual displays of misconception categories with regard to the correct response category that measures a latent trait of interest. It does not offer a framework in which the misconceptions could be scored. Bradshaw and Templin (2014) modeled misconceptions as categorical latent variables within the framework of cognitive diagnostic models (CDM, Rupp, Templin, & Henson, 2010). These researchers assumed that misconceptions could be modeled as categorical latent variables, and that theoretical evidence for this can be represented in the Q-matrix of the CDM model (Tatsuoka, 1990). The model proposed in Bradshaw and Templin (2014) was a bi-factor model that had two components: while one component used a regular IRT (2PL) model to scale student ability into a continuous metric, the other component used a CDM model to classify student misconceptions into dichotomous latent class categories. It assumed that the relationship between a misconception and the ability distribution was zero.

I propose a different approach for measuring misconceptions that builds up on the past literature and expands the way misconceptions are modeled. This different approach measures misconceptions as a continuous latent variable using a nested logit 2PL Item Response Theory (IRT; de Ayala, 2009) model (McFadden, 1981; Suh and Bolt, 2010). Single and multiple misconception models will be developed, estimated, and simulated examinee test data fitted. The nested logit IRT models will be estimated using a Bayesian Markov Chain Monte Carlo (MCMC) algorithm (Albert, 1992; Patz & Junker, 1999a, 199b; Kim & Bolt 2007). The models described in the following sections are designed to model both the misconception as well as a primary latent ability simultaneously, and can be used to locate on the latent continuum the
following parameters: item parameters for latent variable of interest ($LVI$), person parameters for $LVI$, distractor parameters for misconception latent variable ($MLV$), and person parameters for $MLV$.

Background

Item response theory

Item response theory (IRT) is a large family of models for understanding the relationship between a test score, its item elements, and the inferred psychological construct underlying test performance differences among examinees. The most common form of test data modeled within an IRT framework is item data from dichotomously scored multiple choice items. Dichotomous or binary scored test item data are often modeled, depending on the test developer needs with the Rasch (Rasch, 1960) or one-parameter logistic (1PL) model, two-parameter logistic (2PL) model, or the three-parameter logistic model (3PL; Birnbaum, 1968; Lord, 1980). For example, the 1PL model postulates that the probability that an examinee $j$ will respond correctly to item $i$, conditional on their ability estimate ($\theta_j$) is strictly a function of item $i$’s difficulty parameter ($\beta_i$). In the 2PL model this probability is a function of two item parameters: item difficulty ($\beta_i$) and item discrimination ($\alpha_i$). The difficulty parameter of an item represents its location on the latent variable continuum. The discrimination parameter represents how well an item differentiates high performing examinees from low performing examinees. The 3PL model adds a third item parameter known as the lower asymptote to the 2PL model. This IRT model postulates that examinee $j$ will respond correctly to item $i$, conditional on their ability estimate ($\theta_j$) is a function of item difficulty ($\beta_i$), item discrimination
(α), and a lower asymptote, also referred to as a pseudo-guessing, parameter (c). Equation 1 illustrates a 1PL IRT model.

\[
P(X_{ij} = 1|\theta_j, \beta_i) = \frac{e^{\theta_j - \beta_i}}{1 + e^{\theta_j - \beta_i}} \quad (1)
\]

Where \(X_{ij}\) represents the observed response, (0 or 1) for binary scored item data for the \(i^{th}\) item and \(j^{th}\) individual. The ability level of the \(j^{th}\) individual is \(\theta_j\), and \(\beta_i\) is the difficulty level of the \(i^{th}\) item, respectively. Equation 2 illustrates the 2PL IRT model with an item discrimination parameter (\(a_i\)). Equation 1 and Equation 2 are two examples of item response models that could be used for binary response (correct or incorrect) item response data.

\[
P(X_{ij} = 1|\theta_j, \alpha_i, \beta_i) = \frac{e^{\alpha_i \theta_j + \beta_i}}{1 + e^{\alpha_i \theta_j + \beta_i}} \quad (2)
\]

One of the assumptions of the IRT models illustrated above is that there is only one latent trait causing the observed variation among individual’s responses, known as unidimensionality. Although unidimensional IRT models are the most common, there are multidimensional IRT models appropriate there is more than one latent trait measured by a test (Reckase 1985, 2009). Multidimensional IRT (MIRT) models estimate probability of a correct response based on how the dimensionality of the test/instrument is specified. MIRT models could be used for many different applications such as test-let IRT models (Wainer, Bradlow, Wang, 2007), bi-factor IRT models (Gibbons & Hedeker, 1992), cognitive diagnostic models (Rupp, Templin, & Henson, 2010) among many others.
Nested logit item response models

Nested-logit models were first introduced by McFadden (1981, 1982) as a framework for modeling choice behaviors when choice decisions are assumed to happen in consecutively. The use of nested-logit models in the latent trait framework was first used by Suh and Bolt (2010). They investigated how the hierarchical relationship of probabilities could be specified based on an a-priori pattern from which the responses are assumed to happen. For example, the probabilities were estimated in two levels by Suh and Bolt (2010): higher level and lower level. The higher level introduces branches that disentangle incorrect options from the correct option whereas the lower level introduces levels that disentangle among the incorrect options. The lower level model is used to make distinctions among incorrect responses and depends on specifics of the model selected, the nature of the test, and/or the assumptions regarding to the underlying latent trait. For example, if there is a consecutive order to the occurrences of A, B, and C, assuming that either A or B happens first. Then C can only happen if B occurred. Then it is plausible to use nested logit item response models. Bolt, Suh and Wollack (2012) employed Bock’s (1972) nominal response model (NRM) at the lower level in order to estimate a second latent trait in addition to the latent trait of interest that was modeled at the first level. Equation 3 illustrates Bolt, Suh, and Wollack’s (2012) two-dimensional 2-parameter-logistic nested-logit model (2D-2PL-NLM),

\[ P(u_{ij} = 0, d_{ijv} = 1 | \theta_j) = P(u_{ij} = 0 | \theta_j) P(d_{ijv} = 1 | u_{ij} = 0, \theta_j) = \]
where, 

\[ Z_{iv}(\eta_j) = \zeta_{iv} + \lambda_{iv}\eta_j \]

where \(i\), \(j\), \(v\) represent items, individuals, and response categories, respectively. The higher level (the left-side bracketed term) is a 2PL model (Equation 2) that reflects the probability of an incorrect response whereas the lower level (the right side bracketed term) reflects the category selection probabilities given a response is incorrect. Their definition for the \(\theta\) was some ability (e.g. math or science ability), and for the second latent variable modeled at the lower hierarchy, \(\eta\), was “a latent construct that influenced category selection” (p. 342). So, they estimated a slope and an intercept parameter for each of the distractors in the test.

Test information functions

A useful characteristic of item response theory is that one can evaluate the usefulness of an item with respect to the latent continuum (Baker, 2001). The amount of information modeling by a single item is a function of correct response probability, and the discrimination power of an item as expressed in Equation 4.

\[ I_i(\theta_j) = a_i^2 P_i(\theta_j) Q_i(\theta_j) \]  

(4)

The test information function is another unique characteristic feature that is often used in IRT applications to quantify the usefulness of test over the range of the ability continuum. The test information function is the sum of the individual item information values. It is especially useful
when developing instruments that measure psychological constructs. One can see the
information against the amount of error being delivered using a test, and make arrangements,
if necessary, such as modifying item, using easier or more difficult item, deleting items,
changing foil structure etc. Test information is quantified as

\[ I(\theta_j) = \sum_{i=1}^{N} I_i(\theta_j) \]  

(5)

Bayesian estimation approach

There are several different ways of estimating IRT model parameters. Each comes with
advantages and assumptions (pros and cons), and often with computational demands. For
typical IRT models, the most common estimation methods employed are (a) joint maximum
likelihood (JML; Birnbaum, 1968; Lord, 1980, 1986; Wright & Stone, 1979), (b) marginal
maximum likelihood (MML; Bock & Aitkin, 1981; Thissen, 1982), (c) conditional maximum
likelihood (Andersen, 1970, 1973), and (d) Bayesian Markov Chain Monte Carlo (MCMC)
estimations (Patz & Junker, 1999). Although, Cohen, and Lee (2002) reported that there were
no considerable differences between MML and MCMC when a complex model, such as Bock’s
(1972) nominal response mode (NRM), were estimated in terms of differences between
population parameters and their estimates Bayesian/MCMC estimation tends to be preferred
when an IRT model gets more complex such (Wollack, Bolt, Cohen, & Lee, 2002). An advantage
of MCMC was stated in Patz and Junker (p. 147, 1999) as “it is difficult to incorporate
uncertainty (standard errors) into the item parameter estimates in calculations of uncertainty
(standard errors) about inferences for examinees, and there is no way to assess the extent to
which standard errors for examinee inferences are overly optimistic because of this...” and
MCMC is “a method of building IRT models that allows for more complete uncertainty
calculations”. This incorporation is not possible in MML estimation because of the sequential nature of the estimation: first estimate items parameters then estimate person parameters. Bayesian models also have advantages over MML approach in accommodating small sample and/or non-random samples. Another advantage of Bayesian estimation is the ease of implementation and availability of free software such as WINBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), OPENBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2007), JAGS (Plummer, 2015), as well as several applications in other well-known commercial such as SAS (SAS Institute, 1985), MPLUS (Muthén & Muthén, 2007), SPSS (Pallant & Manual, 2001), or noncommercial R (Core Team, 2014; Kim & Bolt, 2007). In fact, there are a number of R packages that could be used as an interface to conduct simulation studies such as R2WINBUGS (Sturtsz, Ligges & Gelman, 2005), R2Openbugs (Sturtz, Ligges, & Gelman, 2005, 2010), R2JAGS (Su, & Yajima, 2012), and Stan (McElreath, 2016).

This dissertation will use IRT (de Ayala, 2009) to model misconceptions (Confrey, 1990) by employing a Bayesian IRT estimation algorithm (Patz & Junker, 1999a, 1999b) within the framework of nested logit models (McFadden, 1981; Suh and Bolt, 2010).

Proposed Study

The purpose of this study is three-fold, first to model misconceptions as a unidimensional latent construct. Previous research has established that misconceptions can be modeled as a latent class, but has not considered modeling a misconception as a continuous latent variable. This is important because past misconception assessment practices included some form of assessment that required pretest versus posttest comparisons or comparison among groups (e.g. grade, gender). Continuous scores provide more opportunity for detecting
change than the categorical scores from a pretest to a posttest after intervention (e.g. teaching, curriculum). Continuous IRT scaled scores might also be useful when developing misconception instruments. Second, this study seeks to explore the sensibility to model multiple misconceptions within one test. This is important because some misconception assessment applications included multiple misconceptions measured within a single test administration. Lastly, this dissertation seeks to explore how test information can be estimated increases when the distractors are designed such that they measure misconceptions as continuous latent variables. This is important because when developing instruments that measure psychological constructs, it is often relevant to know how much information about the latent trait is being produced by a test as well as which part of the latent continuum is being measured.

The model under consideration builds upon the traditional unidimensional (Hambleton & Swaminathan, 1992; de Ayala, 2009) IRT models. More specifically, it uses a unidimensional 2PL IRT for modeling the latent variable of interest \( LV_1 (\theta) \) which estimates item response probabilities based on the correct responses versus incorrect responses from a multiple-choice test. An example would be general math, physics or science ability. Likewise, the misconception latent variables (MLVs) were modeled based on the probability that a response is a misconception given a response is incorrect. The model is labeled as a 2PL-2PL nested-logit misconception model (2PL-2PL-NLMM). Equation 6 illustrates how the 2PL-2PL-NLMM estimates the probabilities of selecting correct options. The left side bracketed term represents the overall probability of an incorrect response (higher level). In other words, when the overall probability of a correct response was subtracted from 1, it gives overall the probability of an incorrect response. That probability is then multiplied by the right side bracketed term (lower
level) that the probability of selecting a response category that measures a misconception given
the response is an incorrect one. In the Equation 6, \( \theta \) represents the \( LVI \), for instance, some
sort of ability or achievement. The \( \eta \) represents the \( MLV \) that will be modeled by some specific
incorrect options among a set of multiple-choice items. Subscripts \( i, j \) and \( v \) represent items,
individuals, and misconception categories, respectively. Item parameters are \( a \) & \( \beta \) represent
item discrimination parameter, item difficulty for the 2PL model. The parameter \( \zeta \) is the
location parameter of a distractor measuring a specific misconception on the latent continuum
of the \( MLV \) (\( \eta \)) with a discrimination denoted as \( \lambda \). Lastly, \( k \) represents the number of
misconception latent variables.

\[
P(u_{ij} = 0, d_{ijv} = 1|\theta_j, \eta_{jk}) = P(u_{ij} = 0|\theta_j)P(d_{ijv} = 1|u_{ij} = 0, \eta_{jk}) =
\]

\[
\left[ 1 - \frac{e^{(\beta_i + \alpha_i \theta_j)}}{1 + e^{(\beta_i + \alpha_i \theta_j)}} \right] \left[ \frac{e^{(\zeta_{ik} + \lambda_{ik} \eta_{jk})}}{1 + e^{(\zeta_{ik} + \lambda_{ik} \eta_{jk})}} \right]
\]

(6)

The Likelihood Function is

\[
L_j = P([U_j, D_j]|\theta_j, \eta_j, \omega) = \prod_{i=1}^{n} [P_i(\theta_j)^{u_{ij}} \prod Q_i(\theta_j)^{d_{ij}} P_{i|u=0}(\eta_j)^{d_{ij}}]
\]

(7)

where \( Q_i = 1 - P_i \), \( U_j \) is the response vector for the correct responses, and \( D_j \) is the incorrect
response vector for the examinee \( j \). The log-likelihood function is

\[
log L_i = \sum_{i=1}^{n} \{u_{ij} \log P_{ij} + \sum d_{ij} [\log Q_{ij} + \log P_{ij|u=0}]\}
\]

(8)

This is the log-likelihood function of the 2PL-NLMM. The independence of the observations was
assumed.
Figure 1 illustrates on how the 2PL-2PL-NLMM works. Suh and Bolt (2010) expressed that if the decision on a choice includes a sequential/hierarchical order of several choices, the probabilities in each hierarchy is multiplied by the other levels of the hierarchy in order to find the final probability of a choice.

The 2PL-2PL-NLMM estimates parameters only for the distractors that measure a misconception; there are no parameters for the other distractors that do not measure a misconception. The probability of a correct response is attained from a traditional 2PL IRT model conditional on examinee ability, item discrimination, and item difficulty. Then a
probability of a misconception response in the lower level of the hierarchy is estimated given the distractor discrimination ($\lambda$), distractor difficulty ($\zeta$), and the $MLV$ ($\eta$).

Significance of the study

To the best of my knowledge, this study is the first to model misconception as a continuous latent variable where misconceptions are scaled onto an IRT metric. This study allows one to estimate both the ability and misconception latent variables simultaneously. A discrimination and a difficulty parameter were estimated for both the correct answer, and a distractor measuring a misconception. The characteristics of this study were informed by both IRT applications (Suh and Bolt, 2010; Bolt, Wollack & Suh, 2012) as well as quantitative misconception assessment (Halloun & Hestenes, 1985; Sadler, 1998; Mulford & Robinson, 2002; Wang & Bao, 2010; Favia, Comins, & Thorpe, 2012; Bradshaw & Templin, 2014), and reflect the common characteristics of quantitative misconception assessment while extending this theory to model misconceptions as continuous latent variables.

Research Questions

This dissertation has three major research questions each one related to each of the three main objectives of this study. Research question 1 (RQ 1) was designed to address the first and the third purposes of this dissertation. Specifically, it considers the case when there were a single $LVI$ and a single $MLV$. RQ 1.1 to RQ 1.4 addresses the estimation precision of the parameters (item, person, probability), while RQ 1.5 addresses the additional test information gained by modelling a single $LVI$ and a single $MLV$. RQ 2 was designed to address the second and the third purposes of this dissertation. Specifically, it considers case when there is a single $LVI$ and two $MLV$. RQ 2.1 to RQ 2.4 addresses the estimation precision of the parameters
(item, person, probability), while RQ 2.5 addresses the additional test information gained by modelling a single \textit{LVI} and two \textit{MLVs}. RQ 3 was also designed to address the second and the third purposes of this dissertation. Specifically, it considers the case when there is a single \textit{LVI} and three \textit{MLVs}. RQ 3.1 to RQ 3.4 addresses the estimation precision of the parameters (item, person, probability), while RQ 3.5 addresses the additional test information gained by modelling a single \textit{LVI} and three \textit{MLVs}.

**RQ 1:** What is the degree of estimation precision in a 2PL-2PL-NLMM when a single LVI and MLV are modeled?

RQ 1.1: What are the convergence rates when 2PL-2PL-NLMM was used with a single \textit{MLV}?

RQ 1.2: What is the degree of item parameter estimation bias when 2PL-2PL-NLMM was used with a single \textit{MLV}?

RQ 1.3: What is the degree of person parameter bias when 2PL-2PL-NLMM was used with a single \textit{MLV}?

RQ 1.4: What is the degree of average absolute errors when 2PL-2PL-NLMM was used with a single \textit{MLV}?

RQ 1.5: What is the test information contribution of modeling a single \textit{MLV}?

**RQ 2:** What is the degree of estimation precision in a 2PL-2PL-NLMM when a single LVI and two MLVs are modeled?

RQ 2.1: What are the convergence rates when 2PL-2PL-NLMM was used with two \textit{MLVs}?
RQ 2.2: What is the degree of item parameter estimation bias when 2PL-2PL-NLMM was used with two $MLVs$?

RQ 2.3: What is the degree of person parameter bias when 2PL-2PL-NLMM was used with two $MLVs$?

RQ 2.4: What is the degree of average absolute errors when 2PL-2PL-NLMM was used with two $MLVs$?

RQ 2.5: What is the test information contribution of modeling two $MLVs$?

RQ 3: What is the degree of estimation precision in a 2PL-2PL-NLMM when a single LVI and three MLVs are modeled?

RQ 3.1: What are the convergence rates when 2PL-2PL-NLMM was used with two $MLVs$?

RQ 3.2: What is the degree of item parameter estimation bias when 2PL-2PL-NLMM was used with three $MLVs$?

RQ 3.3: What is the degree of person parameter bias when 2PL-2PL-NLMM was used with three $MLVs$?

RQ 3.4: What is the degree of average absolute errors when 2PL-2PL-NLMM was used with three $MLVs$?

RQ 3.5: What is the test information contribution of modeling three $MLVs$?
CHAPTER II

REVIEW OF LITERATURE

This chapter reviews the studies that examined misconceptions from an epistemological point of view. The most important knowledge elements revised which are the definition of misconceptions, why they occur, and how it has been studied in previous research applications. Next, the studies that included some form of quantitative (purely quantitative or qualitative-quantitative mixed) misconception research will be examined in greater detail. The review of these manuscripts is chronologically ordered starting from early 80s up to the present. The design of each study, the use of instruments and items, the description and the purpose of the experiment, and the measurement processes. Additionally, especial attention was placed on the scoring of misconception although some studies did not provide detailed information related to misconception scoring, so then only what is available in the study will be reviewed for the literature review of this study.

A general overview of the epistemology of misconception research

Confery (1990) conducted a very detailed review of the literature investigating student conceptions, and misconceptions. Confery (1990)’s study was a succinct, 55-page literature review that attempted to reorganize the research prior to 1990 on both student conceptions, and student misconceptions. The study covered misconception stuties from math (including probability and statistics), science, and programming domains. The idea of misconception was developed from the idea of conception. A conception was defined as “children develop ideas
about their world, develop meanings for words used in science, and develop strategies to obtain explanations for how and why things behave as they do (p. 491 Osborne, and Wittrock, 1983)” (p. 16). Throughout the period of years covered in Confery’s review, a variety of labels were used to describe conceptions which were children held about science (Osborne & Freyberg, 1985), children’s arithmetic (Ginsburg, 1977), mathematics of a tribe (Steffe, 1988), preconceptions (Ausubel, novak, & Hanesian, 1978), conceptual primitives (Resnick, 1983), private concepts (Clement, 1982), and alternative frameworks (Driver, 1981). However, when a student’s conceptions contradict with the expert knowledge, they were called misconceptions. A unique part of Confrey’s study was that it categorized the previous research in terms of epistemology. Three major epistemologies were present in the literature base as described by Confrey: Piagetian genetic epistemology, philosophy of science, and systematic errors.

Piagetian epistemology

Piagetian epistemology, with its roots in genetic epistemology, is “to study the development of particular concepts over time in children” (p. 15, Confery, 1990). Piaget argued that “knowledge is a process, not a state” (p. 15, as cited in Confrey, 1990), so the focus of research was on the development of student conceptions. Thus, the way that the world is seen by the learner was assumed to be the starting point for any educational reform such as curriculum design or teaching/instruction strategies. From this tradition, “researchers focus on the development of microstructures, ... researchers seek to examine thorough tasks how a child acts, perceives, and operates, ... mental operations form the basic roots of conceptual development, ... mental operations are embedded in schemes, ... the construction, refinement, and internalization of these schemes occur within a theory-building approach complete with
experimentation” (p. 16). Therefore, instead of working with misconceptions, the purpose, in this tradition, was to address ideas such as conceptions, conceptual difficulties, or preconceptions.

The philosophy of science

The philosophy of science tradition has its roots in the work of Khun (1970) who rejected the idea that science grows by “simple progressive accretion of scientific fact” (as cited in Confery, p. 16). The ideas such as paradigm, scientific revolution, normal science, and anomaly changed the way that science education researchers’, and some of the math education researchers’ approached research practices related to misconceptions (Clement, 1981; Ausubel, Novak, & Hanesian, 1968; Novak, 1985). Based on Khun’s view of science, the factors that were attractive to these researchers were: a) “it allowed these researchers to critique the underlying inductive conception of science, which permeated the textbooks in the form of “the scientific method” (p.18); b) “it rejected theoretically neutral observations, and hence, could support the position that students enter instruction with firmly held beliefs or preconceptions” (p. 18); and c) “it strongly supported the claim that student conceptions relied on a configuration of beliefs, commitments, and expectations and thus to alter these preconceptions and misconceptions would require intellectual transformations akin to those that accompanied transitions in paradigms, a weak view of “ontogeny recapitulates phylogeny” (p. 18).

Systematic errors

The main purpose of the research in this tradition focused on the study of errors and focused on procedural knowledge. This tradition found the majority of studies focusing on mathematics and computer programming. Confery (1990)’s interpretation on this line of
research goes as: “Research in this area is largely concerned with diagnosing errors and remedying them through exposure and rejection. No learning theory is seen as necessary to account for these errors, beyond the recognition that the errors represent overgeneralizations on the part of the students” (p. 42). For instance, Confery (1990) reported how Radatz (1979) categorized errors, and “related them to an information-processing approach” (p. 32). For example, a brief summary of Radatz’s (1979) description of “various causes of errors that cut across mathematical content topics can be identified by examining the mechanisms used in obtaining, processing, retaining, and reproducing the information in mathematical tasks”: a) errors due to processing iconic representations, b) errors due to deficient mastery prerequisite skills, facts, and concepts, c) errors due to incorrect associations or rigidity of thinking leading to inadequate flexibility in decoding and encoding new information and the inhibition of processing new information, and d) errors due to the application of irrelevant rules or strategies.

Newman (1977) classified these errors into: reading, comprehension, transformation, process, and encoding in addition to carelessness and motivation. Some other researchers made a distinction between error and a misconception such that misconceptions were defined as systematic errors. Researchers in this tradition believe that “the errors are unavoidable and necessary in the development of knowledge” (as cited in Confery, p.42).

As a result, Confery’s (1990) review provided a summary of the epistemological foundations of the literature of misconception research. The three major traditions (Piagetian epistemology, philosophy of science, and systematic errors) were identified with their relationship to misconceptions. Researchers who follow Piagetian genetic epistemology
tradition preferred to study conceptions instead of misconceptions. Researchers who followed
the philosophy of science tradition or the systematic errors tradition, however, have a different
position on how they understand misconceptions, and therefore, they preferred to study
misconceptions. Although the general framework described in the previous paragraph seems to
be comprehensive of the underlying epistemological foundations; there were ideas in the
literature starting from 1970s to date (1990) that evolved as researchers’ theoretical
understanding matured. Preconceptions were one of the ideas that were used by a large
number of researchers because it was assumed that what student knows would determine how
instruction should begin. Another concept, conceptual structure was used to understand the
relationship among complex ideas. Researchers, for example, used conceptual maps, semantic
networks in order to understand how what is known by the student is organized. In addition,
conceptual change was another concept that referred to the conditions “under which students
will choose to modify, reject, or extend their conceptions” (p. 22, Confery, 1990). The
distinction between formal and informal knowledge was another important idea that came out
of this framework. Formal knowledge refers to the knowledge gained in a setting where some
form of education was delivered via a structured and organized institution. Informal education
is the opposite, referring to the knowledge gained by unstructured ways (e.g. every day
experience with physical world). The attractiveness of the distinction between formal and
informal education was that if the contrast between formal and informal education was large,
then students would be more likely to have their own ideas and they that would be firmly held.
Biology and statistics are the examples where the contrast between formal and informal
education exists. For example, students are likely to gain knowledge or develop some form of
understanding about the motion topic physics in their day to day experiences with the move of the objects around themselves by simple observations. When it comes to medicine or chemistry, it is not the case since these two areas of expertise are not practically involved in human day to day activities. Lastly, sense data versus theory was another construct used by some researchers, linked to Khun, to understand “how students relate their sensorial experiences to their formal knowledge” (p. 24, Confery, 1990). There were some other less important ideas developed, and used by researchers to understand conceptions, preconceptions and misconceptions (Confery, 1990).

Following Confery's review, Smith, diSessa, and Roschelle (1993) attempted to summarize the general misconception applications and their conclusions in order to better make sense of how the phenomena was understood in the literature and how it can be seen from a constructivist point of view, student misconceptions came are essential to learning, and that learning is a process not a state. Smith et al. (1993) argued that misconception research focused too much on the description of student ideas such as how student misconceptions changed as an intervention was applied. Therefore, their criticism for this line of research was about the weakness of the interpretations of how misconceptions relate to learning theories. Smith et al. identified seven general assertions in the misconception literature:

a) Students have misconceptions

b) Misconceptions originate from prior learning

c) Misconceptions can be stable and widespread among students. Misconceptions can be strongly held and resistant to change

d) Misconceptions interfere with learning
e) Misconceptions must be replaced
f) Instruction should confront misconceptions
g) Research should identify misconceptions

Smith et al. (1993) also pointed out that the typical general assertions found in misconception research were compatible with constructivist theory. However, some of the assertions did not seem to fit constructivist theory. For instance, constructivist approach assumes that learning is a process not a state and the ultimate purpose of misconception research should be to change (replace) the naïve student ideas (misconceptions) with the expert knowledge (scientific consensus). On the other side, the constructivist approach requires adaptation of prior knowledge in order to construct new knowledge. Therefore, using faulty knowledge (misconceptions) as prior knowledge is not compatible with constructivist theory. Another assertion that was criticized by Smith et al, was the idea that instruction should confront misconceptions. Smith et al. (1993) argued that there are strengths and weaknesses to confronting students’ faulty knowledge with expert knowledge. The strength was that presenting opposite or conflicting ideas may create a productive/dynamic learning environment. A possible weakness was that it would be difficult to argue why expert ideas would win against student misconceptions.

The review of the different epistemological approaches to conceptions/misconceptions revealed that researchers who follow constructivist theory tend not to study student misconceptions even though Smith et al. (1993) tried to show that only some of the assertions of misconception research practices were not compatible with constructivist theory. From an instruction perspective, the systematic error tradition focuses on student’s cognitive errors;
therefore, this tradition is consistent the how misconceptions are conceptualization in this study. The philosophy of science perspective is also a tradition that conceptualizes student misconceptions in a manner compatible with this study because in this tradition instruction starts from what a student knows. In this study, it was assumed that misconceptions are systematic sources of errors on the tests purporting to measure a LVI. From a measurement perspective, the responses to incorrect choices are due to existence or absence of a latent variable called misconception. The systematic tendency of a examinee to mark an incorrect response is hypothesized to be due to the existence of a misconception latent variable. Thus, this study is most closely aligned with systematic errors tradition.

A chronological synopsis of the past of quantitative misconception studies

Total test score approaches.

Caramazza, McCloskey and Green (1981) conducted a study investigating the kinds of concepts that university students develop from their personal life experiences. The researchers recruited 50 undergraduate students for their study and each student was asked questions on the topic of projectile motion. Methodologically, this study used a qualitative approach to evaluate the responses of the students. The assessment included visually-based questions asking students how a pendulum would move if the string was released/cut at a given-certain location. Responses were then classified into six general basic response types. The percentages of the students falling into each response type was reported accompanied with students’ background in physics. The study authors noted considerable variability among the student responses and concluded that humans’ physical experiences with the real world does not always lead to an inference that is compatible with concurrent physical laws. Although there
was uncertainty as to whether the student’s inferences were due to inductive or deductive reasoning, the study authors concluded student responses were random. The systematic sophistication that was found in the incorrect responses were defined as misconception (Caramazza et al.).

Bell, Swan, and Taylor (1981) investigated a problem associated with solving verbal problems that contained mathematical decimal places. Their sample consisted of 20 14-year-olds who were less able in the topic. Their investigation included interviews with students, accompanied with evidence gathered from previous research applications regarding misconceptions that students have about the following operations: multiplication, division, summation and subtraction. They prepared and applied a teaching intervention that was designed to remedy the previously identified misconceptions. A pre-test and a post-test single group design with a diagnostic assessment instrument was administered to assess if the teaching method was successful in improving students’ understanding of decimal places. Although the focus of this research was on teaching method that would eventually be used to remedy the misconceptions, they acknowledged that misconceptions such as multiplication always make it bigger or division always make it smaller still existed in the at the time of the post-test.

In an effort to study misconceptions in the context of physics, Clement (1982) explored how Newtonian principles and models and key concepts such as mass, acceleration, momentum, charge, energy, potential difference, and torque were perceived by college students. The author noted that many students struggle understanding these concepts at conceptual level which makes it more difficult for them to understand more complex
mathematical formulations that are based on these concepts. Therefore, it would be more difficult for students to learn any other higher order, or more complex concepts, when they have difficulty in establishing a solid understanding of these more fundamental concepts. In Clement’s study, the statistical method used to assess the misconceptions was based on percentage scores, a classical test theory approach. According to the author, the results showed that misconceptions were very resistant to change, and that there needs to be alternative instructional strategies and better approaches to correct misconceptions.

Halloun and Hestenes (1985) attempted to design and validate an instrument that measures the knowledge state of beginning physics students’ math and physics beliefs. The instrument was designed to be used as a placement exam, a diagnostic test, and as an evaluation tool for classroom instruction. Using the test at pre-test and mechanics ability post-test, Halloun and Hestenes interpreted misconceptions from the view of common sense theory that suggests students, through their experiences make their own interpretation of how the physical world around them works. In other words, each student starts a course with two things: 1) a system of beliefs, and 2) intuitions. Halloun and Hestenes (1985) suggested that instruction that does not take into account the existence of common sense beliefs is largely responsible for incomprehensibility of introductory physics. To measure their instructional outcomes, Halloun and Hestenes developed a set of three assessments which were implemented in three stages. First a math assessment was used to determine the level of mathematics achievement of their students. Second, a qualitative physics assessment tool, which included a scoring rubric for identification of common misconceptions was administered. Both the math and physics assessments were administered as a pre-test. Third, following
instruction, a multiple choice mechanics post-test was administered to assess the effectiveness of the instruction. Intervention evaluation to determine whether there was an improvement in student ability and a reduction in student misconception was based solely on descriptive statistics. The change in the mean scores were used as an evidence that the instrument detected some change, as well as the intervention helped to reduce the misconception levels of students.

A well-known instrument developed by Hestenes, Wells, and Swackhamer (1992) to help physics teachers assess and probe commonsense beliefs (misconceptions) is the Force Concept Inventory (FCI). The inventory has 24 multiple choice questions with a single correct response. The rest of the distractors are designed to measure misconceptions and the majority of the item distractors measure a single but separate misconception. The FCI yields counts of corrects responses as well as the distractors counts which measure misconceptions. The FCI is often administered in a pre- post-test setting and administered for multiple purposes such as a diagnostic tool to classify and identify the misconceptions that students have, for evaluating instruction, and as a placement exam.

Mulford and Robinson (2002) attempted to develop an inventory to measure alternate conceptions (misconceptions) in chemistry education called the Chemistry Concepts Inventory (CCI). This is a 22-item inventory comprised of multiple choice items designed to measure both the ability level, and the misconception level of students. Scoring takes place at the item level and compares the percentage of examinees correctly answering the item to the percentage of examinees incorrectly responding – or responding to a misconception. The purpose of the study was to show that the inventory was capable of detecting a change in student’s ability and the
level of misconception at the end of a semester course. While the CCI did detect a change over time, this observation is no sufficient validation that the CCI is actually measuring a change from misconception thinking to correct conception thinking in the student.

The measurement of misconceptions above all asks students to find the correct answer and as part of the item, foils include one or more misconceptions. Thus the instrument’s focus was on the LVI, not the MLV. Generally, a students’ total raw score was interpreted as their LVI trait amount and the fil percentage as their MVL trait amount consistent with CTT. Thus these authors are implicitly treating misconceptions as a continuous variable, but unfortunately in a rather unsophisticated manner. Percentage correct/incorrect or total score interpretations lack the item to test linkage afforded by IRT.

IRT based approaches.

Favia, Comins, and Thorpe (2012) analyzed data from an instrument called the Astronomy Misconception Inventory (AMI) where each item is a statement about a misconception in astronomy. The instrument was administered to a group of college students at the end of a semester of instruction in astronomy. The participants who took the exam did not know that each of the items in the instrument was a misconception. The options of the test were fixed across the questions and included the following six categories:

a) if you believed it only as a child
b) if you believed it through high school
c) if you believe it now
d) if you believed it but learned otherwise in AST 109 (an astronomy class)
e) if you never thought about it before, but it sounds plausible or correct to you
f) if you never thought about it before, think it is wrong now

After dichotomizing student responses as to whether a misconception was endorsed or not, Favia et al. (2012) employed a variety of test analysis techniques including 1PL IRT, 2PL IRT, nominal response IRT, and principle component analysis to examinee misconceptions. One of the unique characteristic of this study was that the authors called the latent misconception variable irrationality, a term was borrowed from Thorpe, McMillan, Sigmon, Owings, Dawson, and Bouman (2007). The AMI’s purpose was different from the previously mentioned instruments because it measured only misconceptions. In fact, students who took the instrument did not know that all of the statements were actually misconceptions and the irrationality latent variable modeled was whether student believed in a specific misconception. The purpose of the AMI was to determine the location of the misconceptions on the continuum of irrationality so misconception test items could be ordered from the easiest to the hardest.

Misconception Modeling

One of the first advancements in modeling misconceptions was by Saddler (1998) when he used Bock’s (1972) NRM to model misconceptions as nominal latent variable through visual analysis of option characteristic curves. Options characteristic curves depict the probability that an incorrect option would be selected as a function of a latent variable. Used in this manner, Bock’s NRM is an exploratory model, in the sense that it was used to produce option characteristic curves to see how a distractor performed with respect to the underlying latent trait. Saddler’s (1998) investigation of option characteristic curves included only a visual exploration of the test item choices that measured misconceptions and did not have anything to say about an individual’s cognitive state regarding the misconception. Equation 9 illustrates
NRM where $m_i$ represents the $m^{th}$ category of the $i^{th}$ item, and $a_{ik}$ and $c_{ik}$ represent the discrimination and difficulty parameters, respectively. This model divides a logit by the sum of the all of the logits in order to estimate the probability that a foil choice would get selected. Hence, there would be two parameters to be estimated for each distractor of a test item. A limiting feature of Saddle’s implementation of the NRM was that a number of restrictions needed to be imposed on the model in order identify the model parameters. Without lacing restrictions on the model, parameters would be unidentified and the scale of the foils may not be interpretable.

$$P(X_{ij} = k | \theta_i) = \frac{e^{(c_{ik} + a_{ik}\theta_i)}}{\sum_{k=1}^{m_i} e^{(c_{ik} + a_{ik}\theta_i)}}$$  \hspace{1cm} (9)

Model restrictions were that the sum of $c_{ik} + a_{ik}\theta_i = 0$, the sum of $a_{ik} = 0$, and the sum of $c_{ik} = 0$. These restrictions result in one of the $a_{ik}$ parameters will get the largest positive value, whereas another gets the largest negative value. While the former monotonically increases, the latter monotonically decreases and the $a_{ik}$ that monotonically increases is hopefully the correct response.

Bock’s (1972) NRM serves a crucial role in the misconception literature and has appeared twice. The first time it was used was by Saddler (1998) who incorporated the option characteristics curves in order to understand misconception distractors from an exploratory perspective. Second, more recently Bradshaw and Templin (2014) used Bock’s (1972) NRM in a model where the NRM was incorporated for modeling misconceptions as categorical latent variables.
One of the most remarkable, complex, and interpretable attempts to model misconceptions within the general framework of IRT was by Bradshaw and Templin (2014). Their purpose was to develop a model that could provide feedback to students on a given topic including possible misconceptions that students might have. The model was called ‘scaling individuals, and classifying misconceptions (SICM)’ model. SICM model used combined IRT and cognitive diagnostic models within a bi-factor framework (Gibbons & Hedeker, 1992) where ability was modeled as a continuous latent general factor and misconceptions were modeled as categorical latent variables as the sub-factors. Their model was specified so there was no correlation between the general factor and the sub-factors which indicated that students with the same misconception response pattern may have differing ability estimates. Equation 10 illustrates Bradshaw and Templin’s SICM model such that i and j subscripts represent items and examinees, respectively. SICM model produces two components as $a$ which represents the diagnostic pattern of the misconception latent variables and student ability, $\theta$, which represents a unidimensional continuous latent trait of interest. Equation 10 provides only a general representation of the model, details of the model can be found in Bradshaw and Templin (2014).

$$P(X_e = x_e) = \int_{-\infty}^{\infty} \sum_{e=1}^A \sum_{i=1}^{I_i} \sum_{j=1}^{J_j} \prod_{i=1}^{I_i} \prod_{j=1}^{J_j} \pi_{n_{ij}|a_e, \theta} P(\theta) d(\theta)$$

(10)

Although there is a value in modeling misconceptions as categorical latent variables this approach may not fully meet the needs of the researcher. For instance, the majority of quantitative research investigating misconceptions has actually focused on changing student levels of the misconception via instruction as measured by a pre-test to a post-test design (Mulford & Robinson, 2002; Hestenes, Wells & Swackhamer, 1992; Khazanov, 2008).
categorical latent variables of the SICM model would fit well when there is a need to deliver the feedback to students about their state of a certain content domain. However, when the purpose is to reduce or eliminate the misconceptions via an intervention, e.g., instruction, and/or to see the amount of reduction over time, then the categorical representation of the misconception provided by Bradshaw and Templin’s SICM model may not be sufficient.

Nested logit item response models

As previously noted, nested logit item response models were first introduced by Suh and Bolt (2010) to estimate a slope and a location parameter for each distractor. The latent trait for the distractors were called a trait that influenced category selection. Bolt, Wollack and Suh (2012) extended the use of nested logit item response models to a multidimensional structure where the lower level of the model could contain two different latent traits as $\theta_{j1}$ and $\theta_{j2}$ where each latent trait was based on a number of test items with individual slope and location parameters for each distractor that was assumed to measure each latent variable. The nested logit item response models were, for example, used for modelling a latent variable (mathematics placement exam) in addition to two latent variables modeled using distractors. The two latent variables were English usage ($\theta_{j1}$) and sentence correction($\theta_{j2}$) which were the characteristics that the distractors of the test assumed to measure (Bolt, Suh, Wollack, 2012). In misconception instruments, the distractors were designed in a way that they can measure specific misconceptions. Misconceptions could be assumed to be the latent variables that influenced the selection of the distractors that measured specific misconception latent variables.
Bayesian and ML item response models

Bayesian methods offer an estimation method for item response models (Patz & Junker 1999b; Béguin & Glas, 2001; Fox & Glas, 2001; Bolt & Lal, 2003) and Albert (1992) was the first to study the Bayesian method for item response models, specifically, for a 2PL normal ogive model. Lord (1986) pointed out that Bayesian method will return smaller mean squared error for theta estimates than the maximum likelihood methods. The reason was explained as “when item parameters are known, the MLE of ability assigned to a given response pattern must always be the same. In Bayesian methods, however, the ability estimate assigned to a given response pattern depends on the characteristics of the entire group analyzed. It is this additional flexibility that allows Bayesian methods to obtain a smaller MSE” (p. 159, Lord, 1986). As a result, several advantages were reported by the use of Bayesian estimation for item response models. One of the most important feature is that item or person parameters would be in a reasonable range, especially item discrimination and person ability parameters (Lord, 1986).

Patz and Junker (1999a, 1999b) also used Bayesian estimation for item response models citing two major reasons for their choice. First was about the uncertainty (standard errors) associated with parameter estimates. One has to estimate item parameters first, then the person parameters have to be estimated marginal maximum likelihood (MML). However, item and person parameters were simultaneously estimated by the Bayesian or JML estimation. In other words, the standard errors associated with the item parameter estimates would not be incorporated in to the standard errors associated with person parameter estimates in MML estimation. Bayesian estimation can do both at the same time. Although the JML estimation
method also would estimate item and person parameters simultaneously, the Bayesian method does not suffer from unreasonable person or item estimates as the JML does. JML is known to be unstable in terms of estimating item/person parameters that are at the far ends (too easy/difficult item or too high/low achieving test takers) of a latent continuum.

There were many applications of item response models found in literature that used Bayesian approach for parameter estimation (Kim, & Bolt 2007). The Bayesian estimation requires the specification of prior distribution in order to estimate a posterior distribution. Prior distribution specifications for model parameters usually involves standard normal distributions for location parameters (person or item location) with mean zero. The variances of the location parameters (item or person location) were for instance, either 1.0 for person location or 2.0 for item locations in BILOG (Mislevy, & Bock, 1990). If an item has a discrimination value zero or negative, it means that the question is not able to well discriminate the high achieving students from the low achieving students. Therefore, theoretically, a discrimination parameter should be bigger than zero. The specification of a discrimination parameter distribution varied from a study to another in the Bayesian applications of item response models (Beguin, & Glass, 2001; Bolt, & Lall, 2003; Kim, & Bolt 2007).
CHAPTER III

METHODOLOGY

To better understand the methods, results and discussion sections, I present the nomenclature/notation (see Table 1) used to describe the statistical parameters in Chapters III and IV.

Table 1. Notation Table

<table>
<thead>
<tr>
<th>Formulas*</th>
<th>Text**</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta )</td>
<td>( LVI )</td>
<td>Latent variable of interest</td>
</tr>
<tr>
<td>( \eta )</td>
<td>( MLV )</td>
<td>Misconception latent variable</td>
</tr>
<tr>
<td>( a )</td>
<td>( a_1 )</td>
<td>Item discrimination</td>
</tr>
<tr>
<td>( \beta )</td>
<td>( b_1 )</td>
<td>Item location</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>( a_2, a_3, a_4 )</td>
<td>Distractor discrimination</td>
</tr>
<tr>
<td>( \zeta )</td>
<td>( b_2, b_3, b_4 )</td>
<td>Distractor location</td>
</tr>
</tbody>
</table>

*This notation will be used for the formulas and equations. **This notation will be used inside the text.

The design of data simulation

Fifty data replications were simulated for the certain conditions explained below. Test length, number of misconception latent variables, and the prior distribution specifications were manipulated in order to assess their impact on the estimation of the parameters of the model proposed. To address the research questions of this dissertation, misconception dimensionality was manipulated from one to three. Although, this study simulated a maximum of three dimensions, it could be extended to a larger dimensionality. The total number of test items
measuring the $LVI$ was set to 25 for the short test length and 50 for the long test length. In each instance, either all 25, or all 50 items measured a single latent variable of interest ($LVI$). In other words, there was only one underlying latent trait (e.g. ability or achievement) that distinguished correct from incorrect responses. The number of distractors measuring a misconception latent variable ($MLV$) was also manipulated as a function of the test length. Figure 2 depicts the structural representation of a test with 25 items measuring the $LVI$, and distractors measured three distinct $MLVs$. Lastly, MCMC simulations are partially governed by a-priori distribution characteristics, thus this feature’s influence on the study outcomes was examined by investigating two different a-priori distributions.

![Diagram](image)

**Figure 2. An abstraction for measurement model of a NLMM with 3 MLVs**
Number of examinees

In order to determine the number of examinees to be used in this dissertation, similar studies were investigated. For example, Bolt, Wollack and Suh (2012) used 3000 examinees in their simulation study in which they employed Bock’s (1972) nominal response model in the lower level of the nested logit item response model. In their model, two parameters (discrimination and difficulty) for each correct response, and two parameters (discrimination and difficulty) for each category were estimated. In other words, they estimated a total of 10 parameters for each question. Their conclusion was that a sample size of 3000 was adequate for the recovery of items. DeMars (2003) studied the sample size and parameter recovery of nominal response IRT models. Their study pointed out that any sample size between 600 and 2400 was large enough to get accurate (root mean square error less than 0.10) estimates of discrimination and difficulty parameters. DeMars (2003) estimated either 6 or 12 parameters per item. In their study, two parameters per correct response, and two parameters per distractor measuring a misconception were estimated. Thus, four parameters per item were estimated. Based on Bolt, Wollack and Suh (2012) and DeMars (2003), a sample size of 2000 was set.

Number of items

The total number of items was set to either 25 or 50. Each item had only one correct response, and four incorrect responses of which one represented a misconception. When the total number of items were 25, there were 25 correct responses that measured an $LV_1$ and 25 distractors measuring the $MLV$. For the case where there were two $MLV$ dimensions, 25 items measured $LV_1$ and 12 distractors measured $MLV_1$ and 13 items measured $MLV_2$. Lastly, if there
were \(MLV\)s, then \(MLV_1\), \(MLV_2\), and \(MLV_3\) were measured by 8, 8, and 9 items, respectively.

The decomposition of the distractors for the short test length is described in Table 2.

Table 2. The decomposition of distractors for the test length of 25 by the dimensionality of \(MLV\)s

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>25 item-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(LVI) - 1(MLV)</td>
<td>(LVI = 25, MLV =25)</td>
</tr>
<tr>
<td>1(LVI) - 2(MLV)s</td>
<td>(LVI = 25, MLV_1 =12, MLV_2 =13)</td>
</tr>
<tr>
<td>1(LVI) - 3(MLV)s</td>
<td>(LVI = 25, MLV_1 =8, MLV_2 =8, MLV_3 =9)</td>
</tr>
</tbody>
</table>

When the total number of items were 50, there were 50 correct responses that measured the \(LVI\). The \(MLV\) was measured by 50 distractors when there was one \(MLV\). For two \(MLV\)s, 50 items measured \(LVI\), and each \(MLV\) was measured by 25 items. Lastly, if there were three \(MLV\)s, then \(MLV_1\), \(MLV_2\), and \(MLV_3\) were measured by 17, 17, and 16 items, respectively. The decomposition is described in Table 3.

Table 3. The decomposition of distractors for the test length of 50 by the dimensionality of \(MLV\)s

<table>
<thead>
<tr>
<th>Dimensionality</th>
<th>50 item-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(LVI) - 1(MLV)</td>
<td>(LVI = 50, MLV =50)</td>
</tr>
<tr>
<td>1(LVI) - 2(MLV)s</td>
<td>(LVI = 50, MLV_1 =25, MLV_2 =25)</td>
</tr>
<tr>
<td>1(LVI) - 3(MLV)s</td>
<td>(LVI = 50, MLV_1 =17, MLV_2 =17, MLV_3 =16)</td>
</tr>
</tbody>
</table>
IRT parameter values

In order to generate the true response data, two things had to be specified. First, the parameter specification of the \( LVI \) and second the parameter specification of \( MLV \). The values for \( LVI \) were informed by previous IRT literature (Harwel, Stone, & Kirisci, 1996); however, specification for \( MLV \) values was not available in the extent literature to date. Thus, for this study, \( MLV \) specifications was distributed as typical of IRT model parameter specification.

Previous IRT simulation studies served to set the parameter ranges for a 2PL. Person ability (\( LVI \)) was generated from a random normal distribution, as \( N(0,1) \) (Harwel, Stone, & Kirisci, 1996). Item discrimination parameter (\( \alpha \)) was generated from a lognormal distribution with minimum = 0.0, maximum = 4.50, \( M = 0.01 \), and \( SD = 0.5 \) (Bolt, Wollack, & Suh, 2012). Item difficulty parameters (\( b \)) were generated from a random normal distribution with \( M = 0 \) and \( SD = 1.0, N(0,1) \).

Misconception level

The population parameters for \( MLV \), were generated from a normal distribution, \( N(0,1) \). When multiple misconceptions were estimated. Simulated misconceptions were generated from an independent multivariate normal distribution. An independent multivariate normal distribution sets all off diagonal elements to 0.0, and all diagonal elements to 1.0 in their correlation matrix. This modeling specification presupposes independent misconceptions in the vector \( \{ MLV_1, MLV_2, MLV_3 \} \) as illustrated in Equation 11.

\[
\Sigma_{\text{misconception}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}
\]

(11)

The distractor location parameter of the \( MLV \) (\( b_2 \)) was generated from a normal distribution with \( M=0.50 \), and \( SD=1 \). The reason for choosing a mean of 0.5 instead of 0 was to
make sure that those distractors are a bit more attractive so that they would be selected more often than the other categories in the simulation. Lastly, the discrimination parameters for $MLV (a_2)$ were assumed to be from a lognormal distribution with minimum = 0.0, maximum = 4.50, $M=0.01$, and $SD=0.5$.

Use of non-informative priors

Bayesian estimation requires a distributional prior specification for the parameters being estimated. Two conditions were added for evaluating the prior distribution specifications: normal and uniform. Table 4 below is a summary of the parameters that were used to generate the data, and the prior distributions for the Bayesian estimation algorithms.

Table 4. Distributional characteristics of parameters for data generation, and prior specification

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True values</th>
<th>Prior 1*</th>
<th>Prior 2*</th>
</tr>
</thead>
<tbody>
<tr>
<td>$LVI$</td>
<td>Normal (0,1)</td>
<td>Normal (0,1)</td>
<td>Normal (0,1)</td>
</tr>
<tr>
<td>$MLV$</td>
<td>Normal (0,1)</td>
<td>Normal (0,1)</td>
<td>Normal (0,1)</td>
</tr>
<tr>
<td>Item Discrimination ($a_1$)</td>
<td>Lognormal (0, 0.5)</td>
<td>Uniform (0.5, 2.5)</td>
<td>Normal (1.5, 0.3)</td>
</tr>
<tr>
<td>Distractor Discrimination ($a_2$)</td>
<td>Lognormal (0, 0.5)</td>
<td>Uniform (0.5, 2.5)</td>
<td>Normal (1.5, 0.3)</td>
</tr>
<tr>
<td>Item Location ($b_1$)</td>
<td>Normal (0,1)</td>
<td>Normal (0,1)</td>
<td>Normal (0,1)</td>
</tr>
<tr>
<td>Distractor Location ($b_2$)</td>
<td>Normal (0.5, 1)</td>
<td>Normal (0.5 ,1)</td>
<td>Normal (0.5 ,1)</td>
</tr>
</tbody>
</table>

*Prior distribution specifications.

Overall, this study followed a 2X2X3 design factorial. Figure 3 is a visual representation of this study. Fifty replications for each of the 12 design conditions were simulated. Parameter
simulation commenced after a period of 500 burn-in iterations with a maximum number of iterations set to 5500.

Prior1 and Prior2 were described in the previous table.

Figure 3. Study design (Design cube)

Simulation evaluation

Convergence assessment
RQ 1.1, RQ 2.1, and RQ 3.1 refer to the model convergence. One of the critical factors in reporting Bayesian estimations is to make sure that the sequence of the states in a Markov
chain has arrived to a stationary distribution (Kim & Bolt, 2007). In order to make sure that a chain arrived to a stationary distribution, the diagnostics regarding to convergence should be computed and reported. There are many ways of monitoring convergence. This study considered the statistics that were frequently used in psychometric literature, namely, Geweke’s (1992) $Z$, and Raftery and Lewis’s (1992) $I$ were considered here to monitor convergence.

Geweke’s (1992) $Z$ criterion computes a $z$-score that compares the mean of the first 10% of the states to the mean of the last 50% of the states (Kim & Bolt, 2007). A non-significant difference (-1.96 ≤ $Z$ ≤ 1.96) is evidence of convergence. Percentage of the non-converged parameters will be reported for every simulated dataset, averaged across replications. The output statistic will be reported for the average of an entire simulation condition.

The convergence-monitoring statistic developed by Raftery and Lewis (1992) will also be used as an index of model convergence. This index is usually refereed as $I$, (Kim & Bolt, 2007). If the statistic value is less than 5, it will be used as an indication of convergence. The percentage of parameters that did not converge will be reported.

Item parameter bias and RMSE

RQ 1.2, RQ 2.2, and RQ 3.2 refer to item parameter estimation bias. Root mean square errors (RMSE) was computed for item parameters across replications in order to assess how much true and estimated item parameters differ. RMSE can be understood as the relative distance between the generated and the estimated parameter given a model.

$$RMSE = \left[ \frac{\sum_{r=1}^{R} \left( \sum_{i=1}^{I} (y_{ir} - \hat{y}_{ir})^2 / R \right)^{1/2}}{R} \right]$$

(12)
The parameters denoted by $y_{ir}$ and $\hat{y}_i$ represent true and estimated parameters, respectively. R is the total number of replications for a condition, and $I$ is the total number of any type of item/distractor parameters. For example, if the parameter under consideration is a difficulty parameter ($\beta$) for $LVI$ or the location parameter ($\zeta$) for an $MLV$ ($\eta$), the RMSE could be interpreted in terms of the units of a standard normal distribution. Most of these values for the difficulty/location parameters will range from -4 to +4. So, the best value of RMSE is obviously zero. The worst value, however, could be 8 assuming that the true location parameters was 4, and estimated as -4. As a result, expectation of RMSE for difficulty/location parameters is in the range of [0,8]. For instance, an RMSE value of 1 represents a discrepancy as large as a one standard deviation unit between a true and an estimated location parameter. RMSE for each item type will be averaged within a replication.

If the parameter under consideration is a discrimination parameter ($\alpha$ or $\lambda$), then the units range from 0 to 4.5. Therefore, the values of RMSEs for discrimination parameter are expected to range from 0 to 4.5. Smaller values of RMSEs are preferred. An RMSE value of 0.5 in terms of item/distractor discrimination represents a difference of a one standard deviation because when the data was generated, 0.5 was the standard deviation.

Person parameter bias and RMSE

RQ 1.3, RQ 2.3, and RQ 3.3 refer to person parameter bias, which is the difference between true and estimated person abilities. Since one of the most important aspect of this study was to locate individuals on the $MLV$, it is essential to evaluate the extent to which the person ability estimates have estimation bias. In order to assess person ability estimation bias, person ability residuals were computed, then averaged within each replication. Aggregating
over replications, the average of the entire condition for person ability estimates were reported. Equation 13 illustrates the calculation of person parameter bias. It is expected that the average of the residuals will equal zero. Considerable deviations above or below zero are indication of person parameter estimation bias.

$$Bias = \left[ \frac{\sum_{r=1}^{R}\sum_{j=1}^{J}(\theta_{jr} - \hat{\theta}_{jr})/N}{R} \right]$$  

(13)

Person parameter RMSE is similar to the person parameter bias. The only difference is that the residuals were squared, which is the raw difference between the true and the estimated person parameter. Then, the square root of the average of the residuals was computed. Then, the average RMSE across 50 replications on a given condition was reported. Person parameters values were generated from a standard normal distribution.

$$RMSE = \left[ \frac{\sum_{r=1}^{R}\sum_{j=1}^{J}((\theta_{jr} - \hat{\theta}_{jr})^2/N)^{1/2}}{R} \right]$$  

(14)

The $\theta$ and $\hat{\theta}$ are true and estimated person ability parameters, respectively. $R$ represents replications, and is set to 50. The $j$ is the index for person parameters, and $N$ is the sample size. Both the $LVI$ and $MLV$ were simulated from a standard normal distribution, $N(0,1)$. Most of these values for the ability parameters will range from -4 to +4. So, the best value of RMSE is obviously zero; the worst value could be 8 assuming that a true location as 4, would be estimated as -4. As a result, expectation of RMSE for difficulty/location parameters is in the range of [0,8]. An RMSE of 1, would indicate a one standard deviation difference on average between the true and the estimated person ability parameter.
Average absolute error

RQ 1.4, RQ 2.4, and RQ 3.4 refer to average absolute error (AAE) of response probabilities. Average absolute error is a person response probability measure that describes the discrepancy between the true and the estimated person response probabilities. It is essential to monitor the response probabilities because in this modelling framework, each probability is a function of a set of parameters. Each parameter is estimated with some amount of error. Therefore, monitoring the impact of estimation on a probability value informs us on the precision of estimation.

In this study, there are two distinct probabilities being computed, one is the probability that can be seen on the left side of the Equation 6, which is the probability of a correct response given the examinee ability, item difficulty, and item discrimination. The second probability being computed is the one that distinguishes the probability of selecting a misconception distractor given the examinee’s level of misconception, the discrimination power of that distractor, and the location of that distractor in the misconception latent variable continuum, which is the right side of the Equation 6. Wollack, Bolt, Cohen & Lee (2002) defined AAE as “the total unsigned difference between the estimated and true probabilities of selecting an alternative at a particular theta value”.

\[
AAE = \frac{\sum_{r=1}^{R} | \hat{P}_{ir}(\theta_j) - P_{ir}(\theta_j) | / N}{R}
\]

(15)

AAE is the absolute value of the difference between the true and the estimated probabilities. In Equation 15, \( P(\theta) \) represents the true probability that a response is a correct response or a misconception response given the parameters that were associated with them. Likewise, the estimated probability that a response is a correct one or a misconception response was
denoted by $\hat{P}(\theta)$. A probability estimate may take values ranging from 0 to 1. Therefore, the minimum value for AAE is zero since AAE is an absolute value. The maximum, however, could be 1, assuming, for instance, that a true response probability that was as 0 was estimated as 1. Smaller values of AAE are preferred because a small value represent a better estimation precision.

Additional test information from modeling misconceptions

RQ 1.5, RQ 2.5, and RQ 3.5 refer to visualization of additional test information. The 2PL-2PL-NLMM model produce two different information functions. One is the information gained from modeling $LVI(\theta)$ by the use of the probability that a response is correct given item and person parameters. The other one is from modeling $MLV(\eta)$ by using the distractors that measured an $MLV$. Information function for the $LVI$ will have the form which could be seen in the standard IRT literature (Embertson & Reise, 2000). Equation below illustrates the information function for the probability that a response is correct:

$$I_i(\theta) = a^2 P_i(\theta)(1 - P_i(\theta))$$  

(16)

The other source of information was gained from the use of distractor to model $MLV$. Quantification of the distractors would produce the information function as studied by Bolt, Suh, & Wollack (2012). Following the notion introduced in Table 1, $\eta$ is the $MLV$, $\lambda$ is discrimination parameter for a distractor measuring a misconception, $\zeta$ is a location parameter for a distractor measuring a $MLV$. $P'(\eta)$ and $P''(\eta)$ are first and second derivatives of $P(\eta)$ with respect to $\eta$, respectively. The following equations show how the information function was derived using $P'(\eta)$ and $P''(\eta)$. 

48
\[ I_i(\eta) = \frac{[P'(\eta)]^2}{P(\eta)} - P''(\eta), \quad \text{where,} \]

\[ P(\eta) = \frac{\exp(\zeta + \lambda \eta)}{1 + \exp(\zeta + \lambda \eta)}, \quad P'(\eta) = \frac{\lambda \exp(\zeta + \lambda \eta)}{(1 + \exp(\zeta + \lambda \eta))^2}, \]

and \[ P''(\eta) = \frac{-\lambda^2(\exp(\zeta + \lambda \eta) - 1) \exp(\zeta + \lambda \eta)}{(1 + \exp(\zeta + \lambda \eta))^3}; \text{ therefore,} \]

\[ I_i(\eta) = \frac{\lambda \exp(\zeta + \lambda \eta)}{(1 + \exp(\zeta + \lambda \eta))^2} - \frac{-\lambda^2(\exp(\zeta + \lambda \eta) - 1) \exp(\zeta + \lambda \eta)}{(1 + \exp(\zeta + \lambda \eta))^3} \]

\[ = \frac{\lambda^2 (\exp(\zeta + \lambda \eta))^2}{(1 + \exp(\zeta + \lambda \eta))^3} \]

\[ = \lambda^2 [P(\eta)]^2 [1 - P(\eta)] \quad (17) \]

As a result, \( I_i(\eta) = \lambda^2 [P(\eta)]^2 [1 - P(\eta)] \) will be used to calculate additive information gained by the use of distractors in order to model \( MLVs \). Test information plot examples for the \( MLV \), and \( LVI \) will be provided. Equation 17 is the information function that was developed from Bolt, Suh, and Wollack (2012), where they modeled a response probability by estimating two parameters for each of the distractors on a test item. That probability was actually based on Bock’s (1972) nominal response model. The only difference between Bolt et al. (2012) and this study is that this function is for binary outcome (misconception versus other distractors). Hence, this is just a simpler form of what was modeled in Bolt et al. (2012).
CHAPTER IV

RESULTS

There are three major sections in this chapter, each representing a specific dimensionality of \( MLV \). The results are displayed in the order of the research questions. Each section will start with a description of the structure of the model (e.g. number of items, number of distractors, sample sizes etc.). Convergence diagnostics, RQ n.1, and example MCMC history plots are presented. RMSE estimates of item and distractor parameters within each simulation condition are tabulates along with accompanying residual plots, RQ n.2. Person parameter residual estimates both for \( LVI \)s and \( MLV \)s are reported in the form of bias and RMSE (RQ n.3). AAE is a person response probability residual that describes the discrepancy between the true (population) and the estimated person response probabilities, RQ n.4. The average AAE for the 50 data replications will be reported in a table as summary statistics. In addition, a plot that shows the average AAE for each data replication within a simulation conditions is displayed. Lastly, test information functions for \( LVI \) and \( MLV \) are presented, RQ n.5.

RQ 1: What is the degree of estimation precision in a 2PL-2PL-NLMM when a single LVI and MLV are modeled?

Four study conditions were manipulated: Two conditions for varying prior distribution specifications, and two dimensions of test length (25 and 50 items) as shown in Figure 3. In both short and long length test, both the \( LVI \) and the \( MLV \) were measured by all the items. Table 5 illustrates part of the observed response pattern for the first six simulated individuals: their item responses and \( LVI \) and \( MLV \) estimates. A plus (+) sign stands for a correct response, a minus (-) sign stands for a distractor measuring the \( MLV \), and the rest are the other
distractors that did not measure anything. It can be seen that each individual has two ability estimates: $LVI$ and $MLV$.

Table 5. Observed responses, and corresponding ability estimates for $LVI$ and $MLV$

<table>
<thead>
<tr>
<th>ID</th>
<th>$LVI$</th>
<th>$MLV$</th>
<th>Item1</th>
<th>Item2</th>
<th>Item3</th>
<th>Item4</th>
<th>Item5</th>
<th>Item6</th>
<th>Item7</th>
<th>Item8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.03</td>
<td>1.60b</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>+</td>
<td>-</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>1.43</td>
<td>-0.72</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>B</td>
<td>+</td>
<td>+</td>
<td>B</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>-0.75</td>
<td>0.84</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>C</td>
<td>+</td>
<td>+</td>
<td>C</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>-0.38</td>
<td>1.04</td>
<td>A</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>B</td>
</tr>
<tr>
<td>5</td>
<td>-2.16</td>
<td>0.17</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-0.97</td>
<td>-1.67</td>
<td>B</td>
<td>C</td>
<td>A</td>
<td>B</td>
<td>+</td>
<td>B</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

RQ 1.1: What are the convergence rates when 2PL-2PL-NLMM was used with a single $MLV$? Model convergence was assessed based on the number of parameters that did not converge divided by the total number of parameters. Table 6 displays the non-converged parameter percentages for the 2PL-2PL-NLMM with one $MLV$. Convergence results were summarized by Geweke’s (1992) $Z$, and Raftery and Lewis’s (1992) $I$. Findings indicate that in all four conditions, convergence was very high based on the Raftery and Lewis’s (1992) $I$, never below 5 and $Z < |1.96|$. However, it was noticed that the shorter tests evidenced just a little bit less non-converging parameters than the longer test, although the differences (0.057 and 0.073) are ignorable.
Table 6. Percent non-Convergence Assessment*

<table>
<thead>
<tr>
<th>Conditions</th>
<th>I</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>25XNormal</td>
<td>&lt;0.001</td>
<td>0.057</td>
</tr>
<tr>
<td>25XUniform</td>
<td>&lt;0.001</td>
<td>0.057</td>
</tr>
<tr>
<td>50XNormal</td>
<td>&lt;0.001</td>
<td>0.073</td>
</tr>
<tr>
<td>50XUniform</td>
<td>&lt;0.001</td>
<td>0.062</td>
</tr>
</tbody>
</table>

*Values represent the percentage of the non-converged parameters. Number of items X simulated sample size. I: Raftery and Lewis (1992), Z: Geweke (1992)

Figure 4 illustrates prototypical MCMC history plots for five different parameters. There are two discrimination parameters ($a_1, a_2$), two difficulty parameters ($b_1, b_2$), and two latent variables ($t_1$ and $t_2$). Subscripts 1 represents a parameter that belongs to the $LVI$ while subscript 2 represents a parameter that belongs to $MLV$, respectively. The plots illustrate that the Markov chains arrived to a stationary distribution after 5000 iterations. Sinharay (2003) states that “if different segments of a time series plot for a parameter seems to have traversed different parts of the sample space or if there is a clear pattern in such a plot, the MCMC algorithm may not have converged” (p. 9). Sinharay, further suggests that if the density plot is constantly increasing or decreasing it may indicate convergence problems (2003). The trace plots in Figure 4 indicate a stationary distribution was achieved.
Figure 4. Example MCMC plots for five distinct parameters

RQ 1.2: What is the degree of item parameter estimation bias when 2PL-2PL-NLMM was used with a single MLV?

RMSE is the square root of the average of squared differences between the estimated and the true parameters for a given replication. Table 7 shows the RMSE for the item and distractor parameters \( a_1, b_1, a_2, b_2 \) when modeling an LVI an MLV.
Table 7. RMSE for item parameter estimates

<table>
<thead>
<tr>
<th>Conditions*</th>
<th>( a_1 )</th>
<th>( b_1 )</th>
<th>( a_2 )</th>
<th>( b_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>25XNormal</td>
<td>0.125</td>
<td>0.064</td>
<td>0.198</td>
<td>0.105</td>
</tr>
<tr>
<td>25XUniform</td>
<td>0.135</td>
<td>0.070</td>
<td>0.185</td>
<td>0.106</td>
</tr>
<tr>
<td>50XNormal</td>
<td>0.084</td>
<td>0.065</td>
<td>0.144</td>
<td>0.104</td>
</tr>
<tr>
<td>50XUniform</td>
<td>0.165</td>
<td>0.070</td>
<td>0.194</td>
<td>0.107</td>
</tr>
</tbody>
</table>

*Number of items X simulated sample size.

RMSEs for the discrimination parameter of \( LVI \) \( (a_1) \) ranged from 0.084 to 0.165. The largest RMSE occurred with a long test and a uniform prior distribution specification was used (0.165). The smallest RMSE occurred when a long test with a normal prior distribution specification was used (0.084). The shorter test with a uniform prior distribution specification showed slightly higher RMSE (0.135) than the test with a normal prior distribution specification (0.125). Similarly, the longer test with a uniform prior distribution specification had higher RMSE than the longer test with normal prior distribution specification (0.084 vs 0.165). Figure 5 illustrates the estimation bias for the discrimination parameter of \( LVI \). The first row of Figure 5 represents the short test length, and the second row represents the long test length. It is clear that there was more residual variation in the uniform distribution condition than in the normal. Furthermore, deviant residuals in the uniform condition were all negative; an indication of underestimation bias of this item parameter under the uniform distribution prior distribution condition.
The red line represents the mean of a particular condition.

Figure 5. Estimation Bias for Discrimination Parameter of LVI ($a_1$)

Examination of the $b_1$ parameter in Table 7 revealed that this had the lowest RMSE relative to the other three parameter types within all of the four conditions, ranging from 0.064 to 0.070. Moreover, the RMSE for $b_1$ was smallest when the prior distribution specification was normal for the short length test. The uniform distribution specification evidenced higher RMSE values than the normal distribution specification; however, the differences occurred in the second decimal places (0.064 vs 0.070 or 0.065 vs 0.070).
Figure 6 plots the estimation bias for the difficulty parameters of $LVI (b_1)$.

The vertical axis in each plot of the figure represents the replication datasets while the horizontal axis represents the residual values associated with the difficulty parameter of $LVI$. The normal prior distribution specification did not seem to be associated with estimation bias; however, the uniform distribution conditions revealed some parameters with larger residuals. Other than a few parameters being deviant in the uniform distribution specification conditions, the majority of the difficulty parameters of $LVI$ were estimated without bias.

The red line represents the mean of the estimate in a particular condition.

Figure 6: Estimation Bias for Difficulty Parameter of $LVI (b_1)$
Table 7 also listed the RMSE of the location parameter for the $MLV (b_2)$. The RMSE of these parameters ranged from 0.104 to 0.107. Although the normal prior distribution specifications had smaller RMSE values than the uniform distribution specifications (0.105 vs 0.106 or 0.104 vs 0.107), the differences are at the third decimal places. The same sort of comparison applies when comparing the short test length to the long test length. The length of the test does not seem to be source of variation in terms of RMSE. Figure 7 shows the bias for the difficulty parameter ($b$) of $MLV$. The first row of the figure depicts shorter test length and the second row of the figure longer test length. For the uniform distribution specifications, Figure 7 plots show that few replications were associated with negatively estimated parameters compared to the normal prior distribution conditions. In addition, the normal prior distribution specification conditions seem to be symmetrical in terms of bias. As a result, it can be concluded that in all of the four conditions estimation bias is minimal and a minimal source of concern.
The red line represents the mean of a particular condition.

Figure 7. Estimation Bias for Difficulty Parameter of $MLV (b_2)$

Finally, Table 7 presented that the discrimination parameter of the distractors ($a_2$). For the shorter test length, normal distribution specifications had (0.198) greater RMSE than the uniform distribution condition (0.185). However, for the longer test length, uniform distribution specification had (0.194) greater RMSE than normal distribution specification (0.144). For the normal distribution specifications, the short test length produced higher RMSE (0.198 vs 0.144). In contrast, for the uniform distribution conditions, the longer test length produced higher RMSE (0.185 vs 0.194). Figure 8 illustrates the estimation bias for the discrimination parameter.
of $MLV$ ($a_2$). As can be seen in this figure, residuals from the uniform distribution conditions. All plots are similar to each other within the condition and the residuals from the normal distribution conditions are similar to each other within that condition terms of the bias, Moreover, the residuals from the normal distribution conditions seem to be symmetrically distributed. However, Figure 8 clearly shows that the uniform distribution specification conditions resulted in a negative estimation bias (see the right column of Figure 8). It is also apparent that in the uniform distribution condition the residuals were large in magnitude than in the normal distribution condition.

The red line represents the mean of a particular condition.
Figure 8. Estimation Bias for Discrimination Parameter of MLV ($a_2$)

The estimation bias for the item/distractor parameters is acceptable across conditions. Overall, location parameters have smaller RMSEs than the discrimination parameters. It also can be concluded that $LVI$ item parameters ($a_1$ and $b_1$) had smaller RMSEs than the $MLV$ distractor parameters ($a_2$ and $b_2$). Based on RMSE, model estimation of the item and distractor parameters is within acceptable; however, there was limited mild negative estimation bias in the $a$ parameter in both the $LVI$ and $MLV$ model parts.

RQ 1.3: What is the degree of average absolute errors when 2PL-2PL-NLMM was used with a single $MLV$?

AAE is the absolute value of the difference between the true and the estimated probabilities, see Equation 15. For the AAE, the unit of analysis is either the probability that a response is correct, or the probability that a response is a misconception and range from 0 to 1, with values closer to zero are preferred. Table 8 displays the average absolute error statistics (AAE) for both kinds of probabilities. The first column shows AAE for the correct response probability while the second column shows the AAE for misconception response probability. AAE values were higher for both correct response probabilities and misconception response probabilities when the tests were shorter in length. In other words, the longer the test is, the lower the AAE regardless of LVI or MLV.
Table 8. Average absolute error for person parameter estimates

| Conditions   | $P(u_{ij} = 1 | \theta_j)$  | $P(u_{ij} = 0, d_{ijv} = 1 | \theta_j, \eta_{jk})$ |
|--------------|-----------------------------|-----------------------------------------------|
| 25XNormal    | 0.057                       | 0.079                                         |
| 25XUniform   | 0.057                       | 0.080                                         |
| 50XNormal    | 0.042                       | 0.060                                         |
| 50XUniform   | 0.043                       | 0.061                                         |

*Probability of a correct response.  **Probability of a misconception response.  ***Number of items X simulated sample size.

Figure 9 illustrates the distribution of the AAE values across the replications for the short test length. The first row of the figure is a single replication condition (normal prior), and the second row of the figure is another replication condition (uniform prior). The variation among the correct response AAEs (first column of Figure 9) were similar to each other while the misconception AAEs were also similar to each other (second column of Figure 9). Lastly, it was noticed that misconception AAEs evidenced slightly larger variance than the correct response AAEs.
The red line represents the mean of a particular condition.

Figure 9: Average Absolute Error for the Shorter Test Length (25 items)
Figure 10 illustrates the AAE estimates of the longer test length for the correct response and misconception response probabilities. The first row of the figure is a single replication condition (normal prior), and the second row of the figure is another replication condition (uniform prior). The left column of the figure shows AAEs for the correct response probabilities while the second column of the figure shows AAEs for the misconception response probabilities. AAEs for the correct response probabilities were similar to each other regardless of the prior. Furthermore, correct response AAEs were smaller for the long test length (Figure 10) then the shorter test length (Figure 9). Similarly, AAEs for the misconception response probabilities were also smaller when tests were longer (Figure 9 vs Figure 10). Lastly, there were less variation among the AAEs of the misconception response probabilities when longer tests were used.
The red line represents the mean of a particular condition.

Figure 10: Average Absolute Error for the Longer Test Length (50 items)
In summary, the AAE differences were minimal. For example, for a test of length 25, the AAE for the misconception probabilities was 0.079 for the normal prior distributions, and it was 0.080 for the uniform prior distributions. Similar differences can be seen for the other comparisons. Changing the test length from short to long produces smaller and more stable AAEs. For example, AAE of correct response probability for normal condition with 25 items was 0.057 while AAE for normal condition with 50 items was 0.042. In addition, the variation among the AAEs across the replications were smaller than when longer test were simulated.

RQ 1.4: What is the degree of person parameter bias when 2PL-2PL-NLMM was used with a single \( M_{LV} \)?

Table 9 is a summary of the Bias and RMSE for person trait estimates; see bias statistic, Equation 13 and RMSE statistic, Equation 14. Bias is a raw difference between a true trait score and its estimate. The expected value for the bias is zero, these values are displayed in the Table 9 are averaged over replications. RMSE is the square root of the average of the squared residuals for a single replication and the values shown in Table 9 are averaged over replications. The expected value of RMSE is zero.
Almost all of the bias values in Table 9 are very close to zero suggesting Bias is not a concern in estimating person parameters for \( LV \) and \( MLV \). Figure 11 is the summary plot for the bias estimates over replications of the short and long test. Since there is only one \( MLV \), there are two latent variables for each replication condition: one for the \( LV \) and one for the \( MLV \). Thus Figure 11 presents eight plots for the four conditions. The first two plots on the left top are for the normal prior distribution condition with a short test length. The two plots on the right top corner of the figure are for the uniform prior distribution condition with a short test length. The second row of the figure is for the long test conditions. The findings displayed on Figure 11 aligns with what was displayed in Table 9. It was noticed that the bias had a smaller variance for \( LVI \)s than the \( MLV \)s (first column of Figure 11 versus second column of Figure 11). Additionally, \( LV \) bias estimates for the normal prior distribution conditions were similar to the \( LV \) bias estimates for the uniform distribution conditions (for example the first column of Figure 11 versus the third column of Figure 11). This indicates that type of the prior distribution

<table>
<thead>
<tr>
<th>Conditions***</th>
<th>( LV )</th>
<th>( MLV )</th>
<th>( LV )</th>
<th>( MLV )</th>
</tr>
</thead>
<tbody>
<tr>
<td>25XNormal</td>
<td>-0.006</td>
<td>0.002</td>
<td>0.420</td>
<td>0.573</td>
</tr>
<tr>
<td>25XUniform</td>
<td>0.001</td>
<td>-0.002</td>
<td>0.397</td>
<td>0.553</td>
</tr>
<tr>
<td>50XNormal</td>
<td>-0.002</td>
<td>0.001</td>
<td>0.292</td>
<td>0.432</td>
</tr>
<tr>
<td>50XUniform</td>
<td>&lt;0.001</td>
<td>-0.001</td>
<td>0.293</td>
<td>0.428</td>
</tr>
</tbody>
</table>

*There is a single \( LV \) and a single \( MLV \) for each of the 50 replications. **Average of the bias and RMSE values across replications were reported. ***Number of items X simulated sample size.
is a source estimation bias for the latent variables. Further, the bias was smaller in the long test lengths (first row of the figure versus the second row of the figure).

The first and the third rows are for the normal prior distribution conditions; the second and the fourth rows are for the uniform prior distribution conditions. The column on the left is for the
LVIs; right three columns are for the MLVs. Each black line represents a replication. The red line represents the mean of all replications within a particular condition.

Figure 11. Bias for the estimates of LV1 and MLV when a short test was used

From Table 9, MLV RMSEs were consistently higher than the LV1 RMSEs, short test lengths had higher RMSEs compared to the long test lengths and changing from normal prior distribution specification to a uniform distribution specification does not seem to make a considerable impact on estimation of the latent traits. Figure 12 displays the RMSE estimates for the four conditions. The layout of Figure 12 is the same as Figure 11. Since there is only one MLV, there are two latent variables for each replication condition, LV1 and MLV. Figure 12 shows eight plots for the four conditions which are test length and the prior distribution conditions. The first two plots on the left top corner of the figure were for the normal prior distribution condition when a short test was used while the two plots on the right top corner of the figure were for the uniform prior distribution condition when a short test was used. Examination of Figure 12 suggests there is more variability among the MLVs in terms of RMSE than the LVIs. Further, the RMSEs for the LVIs did not seem to make a visible change when comparing short test to long test. However, there was a visible impact of the test length on the estimation of the MLVs, which could be seen at the second and fourth column of Figure 12. In other words, increasing the number of distractors measuring a misconception decreased the RMSE of the LVIs. In addition, the distributional assumption regarding to the priors did not seem to have a visible impact on the RMSEs of LVIs or MLVs.
*The first and the third rows are for the normal prior distribution conditions; the second and the fourth rows are for the uniform prior distribution conditions. The column on the left is for the LVIs; right three columns are for the MLVs. Each black point represents a replication. The red line represents the mean of all replications within a particular condition.*
Figure 12. *RMSE for the estimates of LVI and MLV*

RQ 1.5: What is the test information contribution of modeling a single *MLV*?

Two kinds of the test information functions were considered in this dissertation. One of them was based on modelling *LVI*, and the other one was based on modelling *MLV*. Modelling the amount of information due to *LVI* could be found in the literature. However, modelling the information due to the *MLV* seems absent in the literature, and was modeled based on Equation 17. Figure 13 shows the test information functions for two replications in the normal prior distribution condition. One for a short test (top row of the figure), and another for a long test (bottom row of the figure) replication. The left side of the Figure 13 represents information functions by modelling *LVI*s, and the right side represents the information functions by modelling *MLV*s. The amount of test information gained by modeling *LVI* was greater than the amount of test information gained by modeling *MLV* for both of the replications.
There were two latent variables in each replication. First row is an example for the 25 items with normal prior distribution specification condition. The second row is another example for the 50 items with normal prior distribution specifications. Red line: Test information function. Blue dashed line: Standard error of measurement.

Figure 13. Illustrative Test Information Functions for Normal Distribution Conditions

Figure 14 shows the test information functions for two replications under the uniform prior distribution condition. One for a short test (top row of the figure), and another for a long test (bottom row of the figure) replication. The left side of the Figure 14 illustrates information
functions by modelling \( LVIs \), and the right side represents the information functions by modelling \( MLVs \). The amount of test information gained by modeling \( LVI \) was greater than the amount of test information gained by modeling \( MLV \) for both of the replications.

There were two latent variables in each replication. First row is an example for the 25 items with uniform prior distribution specification condition. The second row is another example for the 50 items with uniform prior distribution specifications. Red line: Test information function. Blue dashed line: Standard error of measurement.

Figure 14. Illustrative Test Information Functions for Uniform Distribution Conditions
Lastly, since this section of this study examines test information functions for a 2PL-2PL-NLMM when there was one $MLV$, there were two components to be discussed: $LVI$ and, $MLV$. The previous two figures (Figure 13 and Figure 14) summarized four particular replications, one from each condition. So, as an attempt to summarize the test information functions, such as the amount of information gained by modelling each latent variable, across the replications for each condition, the peak points of the curves for each of the latent variables were displayed in Figure 15. The maximum amount of test information across the replications along the X-axis were displayed in the figure. Figure 15 displays the maximum amount of test information (peak point of the test information functions) delivered by an $LVI$ (black lines) versus by $MLV$ (red lines). For all of the conditions, the maximum of the test information functions due to $LVI$s exceeded $MLV$s. Also, a clear pattern is seen with an increase in the number of items, there is an increase the amount of test information for both uniform and normal prior distribution specification conditions.
RQ 2: What is the degree of estimation precision in a 2PL-2PL-NLMM when a single LVI and two MLVs are modeled?

The second research questions addressed in this dissertation concerned the estimation of the 2PL-2PL-NLMM model when there are two $MLV$'s, namely $MLV_1$ and $MLV_2$. This model
was evaluated in four conditions as illustrated in the design cube, Figure 3. There were two conditions for varying prior distribution specifications, and another two conditions of test length. In the shorter test (length of 25), the \( LVI \) was measured by all of the 25 items; however, the distractors of the first 12 items measured \( MLV_1 \), and the distractors of the last 13 items measured \( MLV_2 \). Table 10 illustrates the ability estimates (\( LVI, MLV_1 \), and \( MLV_2 \)) for the first six observations in the first replication for several test items. A plus (+) sign stands for a correct response, a minus (-) sign stands for a misconception response. First three columns of the table show the ability estimates, and the rest of the columns show a portion of the data layout from a single replication. Looking at Item 1 in this table, Person 1 and 5 responded with the keyed correct response, persons 2 and 3 responded incorrectly but to the keyed misconception whereas persons 4 and 6 responded incorrectly and did not endorse the keyed misconception. The rest of the distractors (A, B, C) did not measure anything specific other than being some random distractors. It can be seen that each individual has three trait estimates. The plus (+) signs measured the \( LVI \), the minus (-) signs at Item1, Item2, and Item3 measured \( MLV_1 \), and the minus (-) signs at Item14, Item15, Item16 measured \( MLV_2 \).
Table 10. Observed responses, and corresponding ability estimates for $LVI$, $MLV_1$, and $MLV_2$

<table>
<thead>
<tr>
<th>ID</th>
<th>$LVI$</th>
<th>$MLV_1$</th>
<th>$MLV_2$</th>
<th>Item1</th>
<th>Item2</th>
<th>Item3</th>
<th>Item14</th>
<th>Item15</th>
<th>Item16</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.41</td>
<td>-0.52</td>
<td>-1.47</td>
<td>+</td>
<td>-</td>
<td>C</td>
<td>B</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>-0.33</td>
<td>1.13</td>
<td>-0.13</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>-0.96</td>
<td>1.56</td>
<td>-0.33</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>0.72</td>
<td>0.14</td>
<td>-0.68</td>
<td>A</td>
<td>A</td>
<td>C</td>
<td>B</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>5</td>
<td>0.42</td>
<td>0.47</td>
<td>-0.71</td>
<td>+</td>
<td>A</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>-0.27</td>
<td>0.53</td>
<td>-0.79</td>
<td>C</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

RQ 2.1: What are the convergence rates when 2PL-2PL-NLMM was used with two $MLV$s?

Sinhaary (2003) defined convergence as “convergence occurs when the generated Markov chain converges in distribution to the posterior distribution of interest” (p. 4). The convergence was assessed based on the number of parameters that did not converge divided by the total number of parameters. Table 11 displays the convergence rates of the 2PL-2PL-NLMM when there are two $MLV$'s. The values in the table represent the percentage of the person/item parameter estimates that did not converge. It is apparent that in all four conditions, convergence was very high based on the Raftery and Lewis’s (1992) $I$. However, it was noticed that the shorter length tests had just a little bit less non-converging parameters based on the Geweke’s (1992) $Z$. Although there was very slight difference between the short test length, and the long test length, the differences (0.057 and 0.073) are ignorable.
Table 11. Percent non-Convergence Assessment*

<table>
<thead>
<tr>
<th>Conditions</th>
<th>I</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>25XNormal</td>
<td>0.0001</td>
<td>0.058</td>
</tr>
<tr>
<td>25XUniform</td>
<td>&lt;0.0001</td>
<td>0.056</td>
</tr>
<tr>
<td>50XNormal</td>
<td>&lt;0.0001</td>
<td>0.063</td>
</tr>
<tr>
<td>50XUniform</td>
<td>&lt;0.0001</td>
<td>0.062</td>
</tr>
</tbody>
</table>


Figure 16 displays the MCMC history plots for seven of the parameters. There are two discrimination parameters \((a_1, a_2)\), two difficulty parameters \((b_1, b_2)\), and three latent variables \((t_1, t_2 \text{ and } t_3)\) as examples. The plots illustrate that the chains arrived to a stationary distributions after running the 5000 iterations. Sinharay (2003) states that “if different segments of a time series plot for a parameter seems to have traversed different parts of the sample space or if there is a clear pattern in such a plot, the MCMC algorithm may not have converged” (p. 9). The trace plots in Figure 16 seems to have a stationary distribution after running the iterations. Further, if the density plot is constantly increasing or decreasing, it may indicate convergence problems (Sinharay, 2003). The plots below seem to arrive a stationary distribution.
Figure 16. *Example MCMC history plots for seven distinct parameters*

RQ 2.2: What is the degree of item parameter estimation bias when 2PL-2PL-NLMM was used with two \( MLVs \)?

Table 12 displays the RMSEs of the item and distractor parameters of \( LV1 \) and \( MLVs \).

Since there are two \( MLVs \), the total number of item and distractor parameters to be estimated remained constant however there was an additional \( MLV \) estimated. In the previous section, the short test length model had 25 items measuring an \( LV1 \), in addition to 25 distractors measuring an \( MLV \). However, this time, there were 25 items measuring an \( LV1 \), the distractors of the first 12 items measured \( MLV_1 \), and the distractors of the last 13 items measured \( MLV_2 \). Similarly, the long test length had 50 items measuring an \( LV1 \), in addition to 50 distractors
measuring two separate $MLVs$. Specifically, the distractors of the first 25 items measuring $MLV_1$, and the distractors of the last 25 items measuring $MLV_2$.

Table 12. RMSE for item parameter estimates

<table>
<thead>
<tr>
<th>Conditions*</th>
<th>$a_1$</th>
<th>$b_1$</th>
<th>$a_2$</th>
<th>$b_2$</th>
<th>$a_3$</th>
<th>$b_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25XNormal</td>
<td>0.091</td>
<td>0.066</td>
<td>0.174</td>
<td>0.102</td>
<td>0.184</td>
<td>0.106</td>
</tr>
<tr>
<td>25XUniform</td>
<td>0.160</td>
<td>0.070</td>
<td>0.223</td>
<td>0.121</td>
<td>0.182</td>
<td>0.105</td>
</tr>
<tr>
<td>50XNormal</td>
<td>0.086</td>
<td>0.065</td>
<td>0.155</td>
<td>0.104</td>
<td>0.155</td>
<td>0.101</td>
</tr>
<tr>
<td>50XUniform</td>
<td>0.159</td>
<td>0.072</td>
<td>0.171</td>
<td>0.106</td>
<td>0.189</td>
<td>0.107</td>
</tr>
</tbody>
</table>

*Number of items X simulated sample size

Table 12 shows that the RMSE values of the location parameter of the $LVI (b_1)$ were higher for the uniform distribution conditions than the normal distribution conditions. For example, the RMSE was 0.066 for the short test length with normal distribution specification condition while the RMSE was 0.070 for the short test length with uniform distribution specification condition. Similarly, RMSE was 0.065 for the long test length with normal distribution specification condition while the RMSE was 0.072 for the long test length with uniform distribution specification condition. In can be seen from Table 12 that the differences were small, occurring at the second decimal places. In addition, the test length also did not seem to be a source of negative impact on the RMSEs for the location parameter of the $LVI$s ($b_1$). For example, RMSE for the normal distribution specification with 25 items (0.066) was very close to the RMSE for the normal condition with 50 items (0.065). This same pattern can be
seen for the uniform distribution condition as the condition with uniform distributions with short test (0.070) versus the condition with uniform distribution specification with longer test length (0.072).

Figure 17 illustrates the bias (population minus estimated parameter) values of the location parameter of the \( LV \) across the replications. It is clear that there was no prominent estimation bias. Furthermore, in the uniform condition the residuals were very small. It can be concluded that in terms of both RMSE and bias, the location parameter of the \( LV \) was properly estimated.
The red line represents the mean of a particular condition.

Figure 17. Bias values for the location parameter of the LVI

The second parameter investigated was the discrimination parameter of the LVI ($a_1$). The first column of Table 12 displays the RMSE values, and Figure 18 displays the bias values by replication. The RMSE for the discrimination parameter of the LVI ($a_1$) is smallest in the 50 item test length estimated using a normal prior distribution (0.086). The second smallest RMSE occurred with a short test (25 items) also estimated using a normal prior distribution specification. In the uniform prior distribution specification, the RMSEs almost doubled in magnitude relative to the normal prior distribution conditions. For example, RMSE for normal
prior distribution specification using a long test (0.086) was almost half of RMSE in uniform prior distribution specification using a long test condition (0.159). Similarly, RMSE for normal prior distribution using a short test length (0.091) was almost half of RMSE for uniform prior distribution specification using a short test length (0.160). Figure 18 is an illustration of the bias values of the discrimination parameter of the $LVI$ across all of the four conditions. It was clear that uniform prior distribution specification conditions had many more outlying residuals than the normal prior distribution specification conditions. The uniform prior distribution specification conditions (right column of Figure 18) evidenced a negative estimation bias. In contrast, the normal prior distribution specification conditions looked symmetrical, and none of the values being excessive of $|1|$. 
The red line represents the mean of a particular condition.

Figure 18. Bias values for the discrimination parameter of LVI

The fourth column of Table 12 shows the RMSE values of the location parameter of the $MLV_1 (b_2)$. The RMSE values ranged from 0.102 to 0.121. The largest RMSE values were observed when uniform prior distribution specifications were used with a short test length (0.121). The smallest RMSE occurred with a normal prior distribution specification and a short test (0102). Figure 19 displays the estimation bias of the four replication conditions when the dimensionality of $MLVs$ was two ($MLV_1$ and $MLV_2$). As Figure 19 shows that all of the four conditions looked symmetrical, with similar amounts of variations.
The red line represents the mean of a particular condition.

Figure 19. Bias values for the location parameters of \( MLV_1 \)

The last column of Table 12 represents the RMSE values of the location parameter of \( MLV_2 \). The values are very close to each other ranging from 0.101 to 0.107. Figure 20 shows that there were minimal outliers that presented concern. The values in the last column of Table 12 are very close to the values in the fourth column of Table 12. Both of these columns represent the same kind of parameter (location) for a specific type of latent variable (\( MLV_1 \) and \( MLV_2 \)). The sizes of both values are similar to each other meaning that the location parameters of the \( MLVs \) were estimated almost at the same bias margins.
The red line represents the mean of a particular condition.

Figure 20. Bias values for the location parameters of $MLV_2$
As the third column of the Table 12 showed, the RMSE values of the discrimination parameter of $MLV_1$ was largest (0.223) with a short test length was estimated using a uniform prior distribution specification. It was at its minimum (0.155) with a longer test length estimated using a normal prior distribution specification. RMSE for the short test length with normal prior specification condition (0.174) was smaller than the RMSE for the short test length with uniform prior distribution specification condition (0.223). A similar pattern was recognized for the longer test length as the normal prior distribution specification had a smaller RMSE (0.155) than the uniform distribution specification of the long test length (0.171). It can be concluded that the normal prior distribution specifications resulted in better estimation over the uniform distribution specification. Figure 21 illustrates the estimation bias of the discrimination parameters of the $MLV_1$. The red line on each plot is the mean of the all of the parameters. There were some replications in the uniform prior distribution condition that did not yield good estimations due to having residuals of size greater $|1|$. Moreover, these residuals were mostly negative, an indication for negative estimation bias. On the other hand, normal distribution specification conditions looked like symmetrical without large residuals.
The red line represents the mean of a particular condition.

Figure 21. Bias values for the discrimination parameters of $MLV_1$

One thing that was noticed was that the RMSE values of the discrimination parameter of $MLV_1$ were systematically higher than the RMSE values of the discrimination parameter of the $LVI$, (see the first and third columns of the Table 12). The simplest explanation for this kind of pattern to emerge involves the number of test items used to estimate these parameters. There were either 25 or 50 items used to estimate the discrimination parameter of the $LVI$ while
there were either 12 or 13 items used to estimate the discrimination parameter of the $MLV_1$ and $MLV_2$, respectively.

The fifth column of the Table 12 shows the RMSEs of the discrimination parameter of the $MLV_2$. The largest RMSE (0.189) occurred with a long test estimated using a uniform prior distribution. The RMSE for the short test length with a normal prior distribution specification was 0.184, while the long test with normal prior distribution specification was 0.155. This shows that increasing the number of items has a positive impact on the discrimination parameters of the $MLV$s in terms of item estimation bias. The same result did not apply in the uniform prior distribution specification, however. For the short test length with an uniform prior distribution specification, the RMSE was 0.182 while for the long test length with uniform prior distribution specification, it was 0.189. So, for the uniform distribution specification, increasing the number of items did not produce the same kind of result for the discrimination parameter of $MLV_2$ when there are two misconceptions. Figure 22 also shows the raw form of the residuals summarized in Table 12. The uniform distribution specification conditions (right hand side plots of Figure 22) show greater variation than the normal prior specifications (left hand side plots of Figure 22). In addition, the residuals in the uniform prior distribution conditions were mostly negative, an indication of negative estimation bias. For the normal distribution conditions, the plots looked symmetrical.
The red line represents the mean of a particular condition.

Figure 22: Bias values for the discrimination parameters of MLV$_2$
Overall, item parameters of the $LVI$ had better estimation (in terms of bias and RMSE) compared to distractor parameters of the $MLV$s. The manipulation of the prior distribution specification had some impact on estimation bias on the discrimination parameters, if present - negative. However, this condition manipulation did not have any impact on the estimation of the difficulty parameters. For example, the first column of Table 12 shows that shifting from normal to uniform decreased the accuracy of the estimation for the discrimination parameter of the $LVI$. The same pattern was observed for the discrimination parameter of $MLV_1$. The same kind of impact was not observed for the difficulty parameters due to the fact that the distribution specification of the location parameters was not manipulated. In other words, prior distributions were only manipulated for the discrimination (slope) parameters, which was the reason that the discrimination parameters were influenced most from the manipulation of the prior distributions. A similar conclusion also can be made for the test length. For the difficulty parameters, the test length was not a source of variation in terms of estimation bias. However, the test lengths were a source of variation for the discrimination parameters in terms of the estimation bias.

**RQ 2.3:** What is the degree of average absolute errors when 2PL-2PL-NLMM was used with two $MLV$s?

For the average absolute error (AAE) the unit of analysis was a probability value which can range from 0 to 1. For this study, there were two separate probabilities examined. One is the probability that a response is correct, and the other one is the probability that a response is a misconception. Table 13 displays the AAE for the 2PL-2PL-NLMM when there were two $MLV$s. AAE represents the residuals between the true probabilities and their estimates. The absolute value of the response probability residuals was averaged across replications. Overall AAEs are
mostly smaller than 0.06 for the $LVI$. It was also noticed that the longer test lengths had smaller AAEs (first two rows vs last two rows of Table 13). However, the AAE values are close to 0.10 for the $MLVs$. It can be seen that longer tests had slightly smaller AAEs for $MLVs$. It can also be inferred that the prior distribution specification does not seem to have a big effect on the probability values.

Table 13. Average absolute error for person parameter estimates

| Conditions*** | $P(u_{ij} = 1|\theta_j)$* | $P(u_{ij} = 0, d_{ijv} = 1|\theta_j, \eta_{jk})$** |
|---------------|-----------------------------|--------------------------------------------------|
| 25XNormal     | 0.056                       | 0.099                                            |
| 25XUniform    | 0.057                       | 0.099                                            |
| 50XNormal     | 0.042                       | 0.080                                            |
| 50XUniform    | 0.042                       | 0.079                                            |

*Probability of a correct response. **Probability of a misconception response. ***Number of items X simulated sample size.

Figure 23 illustrates the plot of the summary statistics displayed in Table 13 for the short test lengths (25 items). The first row of the figure is for the condition where a short test was estimated using a normal prior distribution specification. The second row of the figure is for the condition where a short test was estimated using a uniform prior distribution specification. It is clear that most of the correct response probability residuals ranged around 0.05 with a very small variance but misconception response probabilities ranged around 0.10 when a normal distribution specification used for a short test (25 items, first row of Figure 23). Similarly, for the uniform distribution case, the correct response probability residuals were exactly the same as
the normal prior distribution specification. However, misconception response probabilities were around 0.10 where a short test was estimated using uniform prior distributions.
The red line represents the mean of a particular condition.

Figure 23: Average absolute error for short test conditions when MLV is 2-dimensional
Figure 24 illustrates the plot of the summary statistics displayed in Table 13 for the long test lengths (50 items). It is clear that most of the correct response probability residuals (left column of Figure 24) ranged around 0.04 with a very small variance. However, misconception response probabilities were around 0.08 with a normal prior distribution specification, and in the same range when a uniform distribution specification was used. On the other hand, correct response probability residuals were exactly same for both the normal prior distribution specification and the uniform distribution conditions.
The red line represents the mean of a particular condition.

Figure 24: Average absolute error for long test conditions when MLV is 2-dimensional
RQ 2.4: What is the degree of person parameter bias when 2PL-2PL-NLMM was used with two $MLV$s?

Table 14 is a summary of estimation accuracy of the person parameters. The range for the person parameter estimates are from -4 to +4 as in the standard normal distribution. Values close to zero are preferred. For the set of specific conditions that will be examined in this section, there were three latent variables to be examined as $LVI$, $MLV_1$, and $MLV_2$. Table 14 has both averaged bias and RMSE estimates. The RMSE values for $LVI$ ranged from 0.290 to 0.392. It is apparent that RMSEs were smaller when the tests were longer (0.392 vs 0.293 or 0.391 vs 0.290). For the $MLV_1$, RMSEs ranged from 0.553 to 0.683. The bias pattern was in favor of longer tests emerged for the $MLV_1$ as well. For the $MLV_2$, where RMSEs ranged from 0.550 to 0.684. It can be inferred that the test length had a positive impact on the RMSE of the $LVI$ trait estimate. The specification of prior distribution did not seem to make any impact on the bias of the trait estimates. For example, RMSE for $LVI$ was 0.392 when normal prior was used with a short test, while it was 0.391 when a uniform prior was used with a short test.
Table 14. Person parameter bias and RMSE*

<table>
<thead>
<tr>
<th>Conditions***</th>
<th>Bias**</th>
<th>RMSE**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(LVI)</td>
<td>(MLV_1)</td>
</tr>
<tr>
<td>25XNormal</td>
<td>0.001</td>
<td>-0.006</td>
</tr>
<tr>
<td>25XUniform</td>
<td>0.006</td>
<td>0.001</td>
</tr>
<tr>
<td>50XNormal</td>
<td>0.0001</td>
<td>0.0008</td>
</tr>
<tr>
<td>50XUniform</td>
<td>-0.007</td>
<td>-0.004</td>
</tr>
</tbody>
</table>

*There were an \(LVI\) and two \(MLV\)s that were estimated for each of the 50 replications.  
**Average of the bias and RMSE values across replications were reported in the table.  
***Number of items X simulated sample size.

Figure 25 shows the bias of the trait estimates across the replications for the short test lengths. Each row of the plot represents only one simulation condition with three latent variables as \(LVI\), \(MLV_1\), and \(MLV_2\). The first row of Figure 25 shows bias for the short test condition where a normal prior distribution specification was used while the second row of the figure shows the short test condition where a uniform prior distribution specification was used. It is clear that \(LVI\)s had smaller amount of bias compared to the \(MLV\)s. This finding aligns with what was displayed on Table 14.
The first three plots of Figure 25 shows the three latent variables \((LVI, MLV_1, MLV_2)\) that belong to the short test length where normal prior distribution specification was used for the estimation. The last three plots of Figure 25 shows the same three latent variables that belong to short test length where a uniform prior distribution specification was used.
Figure 25. *Bias for the estimates of LVI and MLV when short tests used*

Figure 26 shows the bias of the trait estimates across the replications for the long test lengths. Each row of the plot represents only one simulation condition with three latent variables as *LVI, MLV_1*, and *MLV_2*. The first row of Figure 26 shows bias for the long test condition where a normal prior distribution specification was used while the second row of the figure shows the long test condition where a uniform prior distribution specification was used. It is clear that *LVI*s had smaller amount of bias compared to the *MLV*s.
The first three plots of Figure 26 show the three latent variables ($LVI$, $MLV_1$, $MLV_2$) that belong to the long test length where normal prior distribution specification was used for the estimation. The last three plots of Figure 26 show the same three latent variables that belong to long test length where a uniform prior distribution specification was used.
Figure 26. Bias for the estimates of LVI and MLV when long tests used

Figure 27 shows the RMSE estimates for the short test lengths. The first row of the figure illustrates the short test length conditions where a normal prior distribution specification was used. The second row of the figure, however, illustrates the short test length condition where a uniform prior distribution specification was used. The figure shows that LVI s had RMSEs in a very small range, with a very small variance. On the other hand, MLVs had RMSEs that were bigger and had larger variances. RMSEs for the MLVs were similar to each other regardless of the prior distribution specification.
The left column of Figure 27 shows the three latent variables ($LVI$, $MLV_1$, $MLV_2$) that belong to the short test length where normal prior distribution specification was used for the estimation. The right column of Figure 27 shows the same three latent variables that belong to short test length where a uniform prior distribution specification was used. PP: Person parameter.
Figure 27. RMSE for the estimates of LVI and MLV when short tests used

Figure 28 shows the RMSEs for the long test conditions. The first row of the figure illustrates the latent trait RMSEs for the normal prior distribution specifications while the second row of the figure illustrates the RMSEs for the uniform distribution specifications. The findings found in Figure 27 can be compared to Figure 28. For example, LVI$s had smaller RMSEs than MLV$s. All of the MLV$s had similar amount of RMSE regardless of the prior distribution specification. One thing that was noticed was that the RMSEs of the LVI$s in Figure 27 are consistently higher from the RMSEs found in Figure 28. Similarly, the RMSEs for the MLV$s found in Figure 27 were consistently higher than the RMSEs found in Figure 28, which is an evidence to infer that the length of the test has some positive impact on the estimation of latent variables. Further, the type of prior distribution specification did not seem have some impact on the estimation bias of the latent variables. Both figures, Figure 27 and Figure 28, show that uniform distributions versus normal distributions had similar bias and RMSE estimates.
The left column of Figure 28 shows the three latent variables (LVI, MLV_1, MLV_2) that belong to the long test length where normal prior distribution specification was used for the estimation. The right column of Figure 28 shows the same three latent variables that belong to long test length where a uniform prior distribution specification was used. PP: Person parameter.
RQ 2.5: What is the test information contribution of modeling two $MLV$s?

As it was previously mentioned, there were two kinds of test information functions to be estimated for the 2PL-2PL-NLMM. One was for $LVI$ and the other one is for $MLV$. Calculating the amount of information for $LVI$ was straightforward as applications were found in typical IRT studies (De Ayala, 2009). The calculations of test information functions for the $MLV$s was derived in this study as shown in Equation 17. The following four figures (Figures 29, 30, 31, 32) illustrate the test information functions for $LVI$ and $MLV$s when there were two $MLV$s. Each plot is an example, a randomly picked dataset from each condition, in order to illustrate the test information functions. Since there are a total of three latent variables for this section ($LVI$, $MLV_1$, and $MLV_2$), there will be three plots to be displayed for each dataset.

Figure 29 presents the test information functions for a short test condition with a normal prior distribution specification. The first row of the figure is for the $LVI$, and the second row of the figure is for $MLV_1$ and $MLV_2$, respectively. It can be seen that there was more test information for modeling $LVI$ than modelling $MLV$s which is not surprising due to the study design, more items contribute to the estimation of $LVI$ than to the two $MLV$s. The maximum amount of information due to modeling $LVI$ exceeded test information due to modeling $MLV_1$ and $MLV_2$. 

Figure 28. RMSE for the estimates of $LVI$ and $MLV$ when long tests used
The LVI was measured by 25 items, MLV₁ was measured by 12 distractors (first 12 distractors), MLV₂ was measured by 13 distractors (last 13 distractors).

Figure 29. Illustrative Test Information Functions for Normal Distribution Conditions (25 items)

Figure 30 shows the test information functions for the long test condition with a normal prior distribution specification. Since there were more items measuring both LVI and MLVs, the amount of test information was greater for both LVI and MLVs, compared to the ones displayed on Figure 29.
The LVI was measured by 50 items, \( MLV_1 \) was measured by 25 distractors (first 25 distractors), \( MLV_2 \) was measured by 25 distractors (last 25 distractors).

Figure 30. *Illustrative Test Information Functions for Normal Distribution Conditions (50 items)*

Figure 31 illustrates the test information functions for a uniform prior distribution specification condition with a short test length. The first row of the figure shows information due to modeling \( LVI \), and the second row of the figure shows the test information for modeling \( MLV_1 \) and \( MLV_2 \), respectively. It is clear that \( LVI \) returned more information than the \( MLVs \).
The LVI was measured by 25 items, MLV 1 was measured by 12 distractors (first 12 distractors), MLV 2 was measured by 13 distractors (last 13 distractors).

Figure 31. Illustrative Test Information Functions for Uniform Distribution Conditions (25 items)

Figure 32 illustrates the test information functions for the uniform prior distribution specification condition with a long test length. The first row of the figure shows information due to modelling LVI, and the second row of the figure shows the information due to modeling MLV 1 and MLV 2, respectively. It is clear that LVI returned more information than the MLVs.
The \( LVI \) was measured by 50 items, \( MLV_1 \) was measured by 25 distractors (first 25 distractors), \( MLV_2 \) was measured by 25 distractors (last 25 distractors).

Figure 32. Illustrative Test Information Functions for uniform Distribution Conditions (50 items)

Since this section of this study examines test information functions for a 2PL-2PL-NLMM when there were two \( MLV \)'s, there were three components to be discussed: \( LVI \), \( MLV_1 \), and \( MLV_2 \). The previous four figures (Figure 29, Figure 30, Figure 31, and Figure 32) summarized four particular replication cases from each condition. So, as an attempt to summarize the test information functions, such as the amount of information gained by modelling each latent variable, across the replication for each condition, the peak points of the curves for each of the...
latent variables were displayed in Figure 33. The maximum amount of test information across the replications along the X-axis were displayed in the figure for each of the latent variables. The black line shows the information due to modeling the $LVI$ while blue line and red line show the test information due to modeling $MLV_1$ and $MLV_2$, respectively. It can be seen that in all of the conditions, $LVI$ always produced more information than the $MLVs$. In addition, increasing the number of items helped gain more information for $LVI$. However, the same thing cannot be said for the $MLVs$. For example, the two plots on the first row of Figure 33 shows that the increase on the number of distractors made some increase for the $MLVs$, but at very small amounts.
Black line: Information due to modeling $LVI$. Blue line: Information due to modeling $MLV_1$. Red line: Information due to modeling $MLV_2$.

Figure 33. Maximum amount of information provided by replications
RQ 3: What is the degree of estimation precision in a 2PL-2PL-NLMM when a single LVI and three MLVs are modeled?

The last part of this study examined estimation of the 2PL-2PL-NLMM with three MLVs ($MLV_1$, $MLV_2$, and $MLV_3$). This model, just like the previous two models, was evaluated in four conditions. There were two conditions of prior distribution specification, and two conditions of test length, as the Figure 3 shows. When the shorter test length was used, 25 items measured all LVI items, while 8, 8, and 9 distractors measured the 1st, 2nd, and 3rd MLVs, respectively. When a long test was used, 50 items measured a single LVI, while the $MLV_1$, $MLV_2$, and $MLV_3$ were measured by 17, 17, 16 distractors, respectively. Table 15 shows a portion of the dataset with corresponding ability estimates. In the table, a plus (+) sign stands for a correct response that measured an LVI, and a minus (-) stands for an incorrect response that measured an MLV. The rest of the distractors did not measure anything. The table presents that there are four trait estimates for each individual, one LVI, and three MLVs. All of the + signs measured the LVI, the – signs of the Item 1 and Item2 measured $MLV_1$. The – signs of Item8 and Item9 measured $MLV_2$. The – signs of Item17 and Item18 measured $MLV_3$. 
Table 15. Observed responses, and corresponding ability estimates for \( LVI, MLV_1, MLV_2, \) and \( MLV_3 \)

<table>
<thead>
<tr>
<th>ID</th>
<th>( LVI )</th>
<th>( MLV_1 )</th>
<th>( MLV_2 )</th>
<th>( MLV_3 )</th>
<th>Item1</th>
<th>Item2</th>
<th>Item8</th>
<th>Item9</th>
<th>Item17</th>
<th>Item18</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.97</td>
<td>-0.87</td>
<td>0.54</td>
<td>0.13</td>
<td>+</td>
<td>+</td>
<td>C</td>
<td>+</td>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>-1.05</td>
<td>0.54</td>
<td>0.88</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>+</td>
<td>B</td>
<td>+</td>
</tr>
<tr>
<td>3</td>
<td>0.67</td>
<td>-0.71</td>
<td>0.92</td>
<td>0.62</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>+</td>
<td>C</td>
<td>+</td>
</tr>
<tr>
<td>4</td>
<td>-0.84</td>
<td>1.24</td>
<td>-0.71</td>
<td>-1.04</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>-1.11</td>
<td>1.40</td>
<td>-0.25</td>
<td>-0.60</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>6</td>
<td>-1.06</td>
<td>0.05</td>
<td>0.94</td>
<td>-1.45</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>C</td>
</tr>
</tbody>
</table>

RQ 3.1: What are the convergence rates when 2PL-2PL-NLMM was used with two \( MLV_s \)?

Table 16 displays the average convergence rates of the 2PL-2PL-NLMM with tree \( MLV_s \).

The values in the table were averaged percentages of the non-converged parameters for a particular condition. Overall, a satisfactory level of convergence was reached. There was a slight, but ignorable difference between the small sample sizes and larger sample size conditions.
<table>
<thead>
<tr>
<th>Conditions</th>
<th>I</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>25XNormal</td>
<td>0.0006</td>
<td>0.060</td>
</tr>
<tr>
<td>25XUniform</td>
<td>0.0001</td>
<td>0.058</td>
</tr>
<tr>
<td>50XNormal</td>
<td>0.0001</td>
<td>0.065</td>
</tr>
<tr>
<td>50XUniform</td>
<td>&lt;0.0001</td>
<td>0.066</td>
</tr>
</tbody>
</table>


Figure 34 displays of the MCMC history plots for each kind of parameter to be estimated ($a_1, a_2, b_1, b_2, t_1, t_2, t_3,$ and $t_4$). This Figure illustrates that stationary distributions were reached after running the 5000 iterations for each of the parameters listed.
Figure 34. *Example MCMC history plots for eight distinct parameters*

RQ 3.2: What is the degree of item parameter estimation bias when 2PL-2PL-NLMM was used with three *MLVs*?

The RMSEs are averaged squared residuals between the true and the estimated item/distractor parameters. The range of the difficulty parameters was -4 to +4 due to being generated from a standard normal distribution. Therefore, the range for the RMSE parameter
was expected to be from 0 to 8. For the discrimination parameters, the values ranged mostly from 0 to 4.5. Therefore, the RMSE values of discrimination parameter can range from 0 to 4.5. Table 17 displays the item RMSEs and distractor RMSEs for item and distractor parameters.

Table 17 shows that the difficulty parameters of $LVI \ (b_1)$ had RMSEs ranging from 0.067 to 0.071. The RMSEs were small within all four condition, which indicated that the parameter had small estimation bias regardless of test length and prior distribution.

Table 17. RMSEs for the item parameter estimates

<table>
<thead>
<tr>
<th>Conditions</th>
<th>$LVI$</th>
<th>$MLV_1$</th>
<th>$MLV_2$</th>
<th>$MLV_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_1$</td>
<td>$b_1$</td>
<td>$a_2$</td>
<td>$b_2$</td>
</tr>
<tr>
<td>25XNormal</td>
<td>0.089</td>
<td>0.067</td>
<td>0.236</td>
<td>0.112</td>
</tr>
<tr>
<td>25XUniform</td>
<td>0.163</td>
<td>0.071</td>
<td>0.216</td>
<td>0.114</td>
</tr>
<tr>
<td>50XNormal</td>
<td>0.085</td>
<td>0.068</td>
<td>0.332</td>
<td>0.118</td>
</tr>
<tr>
<td>50XUniform</td>
<td>0.146</td>
<td>0.069</td>
<td>0.254</td>
<td>0.111</td>
</tr>
</tbody>
</table>

*Number of items X simulated sample size.

Figure 35 is the bias plot of the difficulty of $LVI$. Bias is the differences between a true parameter and its estimate. There are four plots in the Figure since there were four replication conditions. It can be seen that the bias estimates for the shorter test length has a similar range as the longer test length. Estimation bias neither as negative nor as positive does not seem to be an issue for this particular parameter ($b_1$).
The red line represents the mean of a particular condition.

Figure 35. Estimation bias of the difficulty of LVI when there were 3 MLVs

The discrimination parameters of the LVI ($a_1$), however, had RMSEs ranging from 0.085 to 0.163 as can be seen at the first column of Table 17. Test length had some impact on the size of RMSE values. Longer test length conditions had smaller RMSEs (0.089 vs 0.085 or 0.163 vs 0.146). Prior distribution specification also had some impact on the estimation of this particular parameter. It could be inferred that the normal prior distribution specifications had less bias. For example, for the shorter test length, normal prior had an RMSE of 0.089 versus uniform...
prior had an RMSE of 0.163. Similarly, for the longer test length, the normal prior had an RMSE of 0.085 versus the uniform prior had an RMSE of 0.146. Figure 36 illustrates the bias for the estimation of discrimination parameters of $LVI (a_1)$. Although the residuals summed to 0 for all four conditions, there were a number of outliers in the uniform prior distribution conditions which were mostly negative, indicative of a negative estimation bias in the uniform distribution conditions. In the normal prior distribution specification conditions only a few outliers were noted.

The red line represents the mean of a particular condition.

Figure 36. Estimation bias of the discrimination parameter of the $LVI$
The 3rd, 5th, and 7th columns of Table 17 displays the RMSE values for the discrimination parameters of $MLV$s. Since there were three separate $MLV$s, there are three discrimination RMSEs, one for each $MLV$. Discrimination parameters of the $MLV$s had similar RMSE values across conditions, ranging from 0.161 to 0.332 in Table 17. The following three Figures, Figure 37 to Figure 39, display the bias estimates of the discrimination parameters of $MLV$s. Figure 37 displays the estimation bias for the discrimination parameter of $MLV_1 (a_2)$. There were some residuals for both short test length and the long test length. The size of the residuals was larger for the longer test lengths. The type of the prior distribution specification (uniform or normal) did not seem to have a consistent impact on estimation bias. For example, for the short test length with normal prior distribution, there were some parameters that were positively estimated, whereas large residuals were negative in the uniform prior distribution specifications. For the longer test lengths, both uniform and normal prior specifications showed negative estimation bias since most of the residuals were negative.
The red line represents the mean of a particular condition.

Figure 37. Estimation bias for the discrimination parameter of \( MLV_1 (a_2) \)

Figure 38 displays the estimation bias for the discrimination parameter of the \( MLV_2 (a_3) \). The shorter test length did not show any estimation bias but the longer test lengths had some. The normal distribution specification with the longer test length had positive estimation bias while the longer test length with uniform prior distribution had negative estimation bias.
The red line represents the mean of a particular condition.

Figure 38. Estimation bias for the discrimination parameter of $MLV_2 (a_3)$

Figure 39 shows the bias for the discrimination parameter of the $MLV_3 (a_4)$. From these Figures, the residuals were larger for the longer test length than the shorter test length. The
The red line represents the mean of a particular condition.
Figure 39. *Estimation bias for the discrimination parameter of MLV$_3$ ($a_4$)*

The 4$^{th}$, 6$^{th}$, and 8$^{th}$ columns of Table 17 present the RMSE values for the difficulty parameters of $MLVs$. The range of the values displayed on the table was from 0.105 to 0.119. RMSE values were similar in magnitude across conditions. The following three Figures, Figure 40 to Figure 42, illustrates the estimation bias for the difficulty parameter of $MLVs$. The red lines on the Figures is the mean of the all of the bias values of that particular condition. Figure 40, for example, shows that for all of the four conditions, the estimation bias of the difficulty parameter of $MLV_1$ ($b_2$) were symmetrical around zero meaning that there is no estimation bias for this particular parameter. Bias values were slightly larger in magnitude when longer tests were used.
The red line represents the mean of a particular condition.

Figure 40. Estimation bias for the difficulty parameter of $MLV_1 (b_2) \) 

Similarly, Figure 41, shows the bias for the difficulty parameter of $MLV_2 (b_3)$. The values depicted on the plots are similar to the previous Figure. Shorter tests had similar estimation bias as the longer test lengths. The type of prior distribution specification did not make any impact on the estimation bias for this particular parameter.
The red line represents the mean of a particular condition.

Figure 41. Estimation bias for the difficulty parameter of $MLV_2 (b_3)$
Lastly, Figure 42 shows the bias for the difficulty parameter of $MLV_3 (b_4)$. The bias values seemed to be equal across the different test lengths. There were not any large outliers that could require some attention. The impact of the type of prior distribution on the estimation bias was minimal.
The red line represents the mean of a particular condition.

Figure 42. Estimation bias for the difficulty parameter of MLV_3 (b_4)
It can be inferred that $LVI$ parameters had smaller RMSEs and biases than the $MLV$ parameters. Although prior distribution specification has some impact on the estimation of the discrimination parameters ($a_1, a_2, a_3, a_4$) such impact was minimal for the difficulty parameters ($b_1, b_2, b_3, b_4$). The impact of test length on RMSE of item/distractor parameters was minimal.

RQ 3.3: What is the degree of average absolute errors when 2PL-2PL-NLMM was used with three $MLVs$?

Table 18 is a summary of the average absolute error (AAE) statistics for person parameter estimates. AAE is the absolute value of the difference between a true probability and its estimate. The range for AAEs is from 0 to 1. Smaller values of AAE are preferred. In other words, AAE is an indicator for the probability residuals. Table 18 shows the probability of a correct response was smaller for longer test lengths but the difference was very small. For the $MLVs$, the difference between the short test versus long test was also minimal in terms of AAE. It can be concluded that the test length does not seem to have a big impact on both of the scales ($LVI$ and $MLV$) in terms of AAE. Further, AAEs were always smaller for $LVI$s than the $MLVs$. 

131
Table 18. Average absolute error for person parameter estimates

| Conditions*** | $P(u_{ij} = 1 | \theta_j)$* | $P(u_{ij} = 0, d_{ijv} = 1 | \theta_j, \eta_{jk})$** |
|--------------|-----------------------------|--------------------------------------------------|
| 25XNormal    | 0.056                       | 0.110                                            |
| 25XUniform   | 0.057                       | 0.111                                            |
| 50XNormal    | 0.042                       | 0.093                                            |
| 50XUniform   | 0.043                       | 0.092                                            |

*Probability of a correct response. **Probability of a misconception response. ***Number of items X simulated sample size.

Figure 43 presents AAE for the short test length conditions. Since there were two probability values considered for each replication, the two plots on the first row of the Figure represent a single replication condition. Similarly, the second row represents another replication condition. It is clear that the AAEs for the correct response probabilities are consistently smaller than the AAEs for the misconception response probabilities (first column of the Figure versus the second column). It could also be said from the Figure that misconception response probability AAEs had larger variances. In addition, the type of prior distribution specification does not seem to have a visible effect on AAEs (first row of the Figure versus the second row). AAE for the correct response probability of normal prior specification conditions had similar pattern as the uniform prior distribution conditions. Similar pattern emerged for the AAE for the misconception probability when comparing normal prior specification to the uniform prior specification.
The red line represents the mean of a particular condition.

Figure 43. Average absolute error for short test conditions when MLV is 3-dimensional
Figure 44 shows the AAEs for the longer test lengths. The AAEs for the long tests had the same pattern as the short test lengths. The impact of prior distribution specifications was negligible for both short and long test conditions.
The red line represents the mean of a particular condition.

Figure 44. Average absolute error for long test conditions when MLV is 3-dimensional
RQ 3.4: What is the degree of person parameter bias when 2PL-2PL-NLMM was used with three $MLV_s$?

The bias and RMSE for the person ability estimates were displayed in Table 19. The bias statistic was calculated based on Equation 13. It is clear that the bias statistics were all very close to 0. This is a partial evidence that the latent traits did not have estimation bias for any of the latent variables. The last four columns of Table 19 show the RMSE values for the estimations of $LVI$ and $MLV_s$. The RMSEs for the $LVI$ ranged from 0.294 to 0.394 while RMSEs for the $MLV_s$ ranged from 0.624 to 0.770. It is clear that $MLV_s$ had larger RMSEs than the $LVI_s$. In addition, the test length also had an effect on RMSE of trait estimates. Comparing the first row of Table 19 to the third row, it can be seen that test length had changed the size of the RMSE values for all of the latent variables. For instance, the RMSE for the normal prior distribution for the short test was 0.394, and 0.294 for the long test length. In addition, the type of prior distribution did not have any effect on the estimation of latent variables (first row versus second row or third row versus fourth row).
Table 19. Person parameter bias and RMSE*

<table>
<thead>
<tr>
<th>Conditions ***</th>
<th>Bias**</th>
<th>RMSE**</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$LVI$</td>
<td>$MLV_1$</td>
</tr>
<tr>
<td>25XNormal</td>
<td>-0.006</td>
<td>0.001</td>
</tr>
<tr>
<td>25XUniform</td>
<td>0.004</td>
<td>-0.004</td>
</tr>
<tr>
<td>50XNormal</td>
<td>0.008</td>
<td>-0.0003</td>
</tr>
<tr>
<td>50XUniform</td>
<td>-0.004</td>
<td>0.003</td>
</tr>
</tbody>
</table>

*There is a single $LVI$ and a single $MLV$ was estimated for each of the 50 replications. **Average of the bias and RMSE values across replications were reported in the table. ***Number of items X simulated sample size.

Figure 45 illustrates the bias for the short test conditions when the prior was specified as normal. The bias for $LVI$ was always smaller than the bias of $MLV$s. The plots on the Figure were symmetrical which was another partial evidence that there was no estimation bias for these latent traits. However, the sizes of the residuals were severe especially for the $MLV$s.
Figure 45 shows the three latent variables (LV1, MLV₁, MLV₂, MLV₃) that belong to the short test length where normal prior distribution specification was used for the estimation.

Figure 45. Bias for the estimates of LV1 and MLV when short tests were used (normal prior specification)

Figure 46 shows the bias for the short test conditions when the prior was specified as uniform. The size of the bias values was similar to Figure 45. Therefore, it can be concluded that the impact of the type of prior distribution specification on the latent trait estimation was minimal.
Figure 46 shows the three latent variables \((LVI, MLV_1, MLV_2, MLV_3)\) that belong to the short test length where uniform prior distribution specification was used for the estimation.
Figure 46. Bias for the estimates of $LVI$ and $MLV$ when short tests were used (uniform prior specification)

Figure 47 shows the estimation bias for the long test conditions where a normal prior was used. Because the test was longer, the sizes of the biases displayed in the Figure were smaller than the ones on Figure 45. As a result, one can argue that longer tests had smaller latent trait estimation bias when normal priors were used.
Figure 47 shows the three latent variables (LVI, MLV₁, MLV₂, MLV₃) that belong to the long test length where normal prior distribution specification was used for the estimation.

Figure 47. Bias for the estimates of LVI and MLV when long tests were used (normal prior)
Figure 48 shows the bias for the latent traits when long test lengths were used with uniform prior distribution specifications. The plots on the Figure show that $LVI$ had smaller bias than the $MLVs$, which was consistent with all of the previous replication conditions. The plots displayed on the Figure 46 had larger bias than the plots displayed on Figure 48. This indicates that longer tests had less estimation bias for latent traits when uniform prior distribution were used.
Figure 48 shows the three latent variables (LVI, MLV₁, MLV₂, MLV₃) that belong to the long test length where uniform prior distribution specification was used for the estimation.

Figure 48. Bias for the estimates of LVI and MLV when long tests were used (uniform prior)
Figure 49 shows the RMSE of the traits across the replications when short tests were used. Since there were three $MLVs$, there were a total of four latent variables estimated for each replication. Therefore, the Figure has eight latent variables where the first row displays the short tests when normal priors were used. The second row of the Figure shows the latent variables when uniform distributions were used. From the Figure, it was difficult to differentiate the $LVI$ of the normal prior from the $LVI$ of the uniform prior. They were similar to each other. Similarly, $MLV$ variables were also similar to each other. As a result, it could be argued that the type of prior specification did not influence the estimation of latent variables.
The first two rows of Figure 49 show the three latent variables (LVI, MLV₁, MLV₂, MLV₃) that belong to the short test length where normal prior distribution specification was used for the estimation. The second two rows of Figure 49 show the same three latent variables that belong to short test length where a uniform prior distribution specification was used. PP: Person parameter.
Figure 49. *RMSE for the estimates of LVI and MLV when short tests used*

Figure 50 displays the RMSE values across the replications when long test lengths were used. The first row of the Figure shows that RMSEs when normal prior specifications were used. The second row of the Figure shows the RMSEs when uniform prior specifications were used. One thing that was noticed in this Figure was that *LVI*s had smaller RMSEs than the *MLV*s. Additionally, the type of the prior specification does not seem to have some impact on the estimation of the latent variables. Lastly, Comparing Figure 49 to Figure 50, one could see that the longer test length had consistently smaller RMSEs than the short test length. In other words, adding more items will help more precise estimations for any type of latent variable.
The first two rows of Figure 50 show the three latent variables (LVI, MLV1, MLV2, MLV3) that belong to the long test length where normal prior distribution specification was used for the estimation. The second two rows of Figure 50 shows the same three latent variables that belong
to long test length where a uniform prior distribution specification was used. PP: Person parameter.

Figure 50. RMSE for the estimates of LV1 and MLV when long tests used

RQ 3.5: What is the test information contribution of modeling three MLVs?

The following Figures (Figure 51 to Figure 54) illustrateative test information functions.

Each plot illustrates a different condition. For example, Figure 51 is for the condition that has short test length (25 items) and normal prior distribution specification. In Figure 51, there are four graphs, the one on the upper left corner is the information function for LV1, and the other three are the information functions for MLV1, MLV2 and MLV3, respectively. It can be concluded that the raw amount of test information gained from modeling LV1 covers a wider range than the modelling of MLVs for this particular replication. It is meaningful because LV1 was measured by 25 items while MLVs were measured by 8, 8, and 9 items. Further, the peak point of the information function of LV1 is greater than MLVs.
The LVI was measured by 25 items, MLV₁ was measured by 8 distractors, MLV₂ was measured by 8 distractors, and MLV₃ was measured by 9 distractors.

Figure 51. *Illustrative Test Information Functions for Normal Distribution Conditions (25 items)*

Figure 52 represent the information functions for the simulation condition where a long test length was estimated using normal prior distribution specification. Unlike the previous Figure (Figure 51), the longer test length produced more information, which is natural since the
test information function depends on the sum of the information provided by each item. As there are more items, it is natural to have more information being produced. It is difficult to make the same conclusion for the misconceptions. Although the number of distractors used in Figure 51 were twice as the ones used in Figure 52 (8, 8, 9 vs 17, 17, 16), it is difficult to see the double change in the amount of test information being produced.

The \( LVI \) was measured by 50 items, \( MLV_1 \) was measured by 17 distractors, \( MLV_2 \) was measured by 17 distractors, and \( MLV_3 \) was measured by 16 distractors.

Figure 52. Illustrative Test Information Functions for Normal Distribution Conditions (50 items)
Figure 53 shows the test information functions for the short test length where a uniform distribution specification was used. The same pattern seen in the previous two Figures (Figure 51 and Figure 52) is also seen in Figure 53. LVI provided more information than each of the MLVs. MLVs produced approximately similar amounts of information.

The LVI was measured by 25 items, MLV₁ was measured by 8 distractors, MLV₂ was measured by 8 distractors, and MLV₃ was measured by 9 distractors.

Figure 53. Illustrative Test Information Functions for Uniform Distribution Conditions (25 items)
Figure 54 displays the amount of information being produced by the long test length where a uniform distribution specification was used. For all of the latent variables, the amount of test information increased to some degree.

The $LVI$ was measured by 50 items, $MLV_1$ was measured by 17 distractors, $MLV_2$ was measured by 17 distractors, and $MLV_3$ was measured by 16 distractors.

Figure 54. Illustrative Test Information Functions for Uniform Distribution Conditions (50 items)
The previous four Figures were illustrative examples that were picked from each replication condition but they lacked a whole picture perspective of this part of the simulation. Therefore, as an attempt to summarize the entire replication conditions, the peak point of each information function was calculated, and displayed in Figure 55. The black lines show the information provided by \( LVI \), and green, red and blues lines represent the \( MLV_1, MLV_2, \) and \( MLV_3 \), respectively. For all of the four conditions, the peak point of the information curves was higher for \( LVI \) than the \( MLV \). The test length had visible impact on the information produced. Longer tests produced more information than the shorter tests. It was difficult to make the same conclusion for the \( MLV \).
Figure 55. Maximum amount of information by replications
CHAPTER V

CONCLUSIONS, DISCUSSION, AND RECOMMENDATIONS

The purpose of this dissertation was to estimate person location parameters, item difficulty parameters, item discrimination parameters, distractor location parameters, and distractor discrimination parameters in an IRT model that includes both $LVI$ and $MLV$. In particular, this study had three main objectives: 1) Estimate misconceptions as a single continuous latent variable in addition to a latent variable of interest; 2) estimate 1, 2, and 3 misconceptions as continuous latent variables in addition to a latent variable of interest, and 3) quantify the information on both the misconception latent variables ($MLV$) and the latent variable of interests ($LVI$). To achieve these aims, an IRT model was specified (Equation 6) and the research questions were answered via a simulation study of a 25 or 50 item MC response test, 3000 examinees with 50 replications. Simulated data were modeled via a Bayesian estimation algorithm with two prior probability specifications (uniform or normal distribution). Fifteen separate research questions (RQs) were proposed. The first, second and third RQs investigated the addition of one, two, and three misconception latent variables respectively in a model already containing an $LVI$. Within each RQ, sub questions focused on model convergence levels (RQ 1.1, RQ 2.1, and RQ 3.1); item parameter estimation bias (RQ 1.2, RQ 2.2, RQ 3.2); person parameter estimation bias (RQ 1.3, RQ 2.3, RQ 3.3); degree of average absolute error (RQ 1.4, RQ 2.4, and RQ 3.4); and if misconception test information adds to the test information (RQ 1.5, RQ 2.5, and RQ 3.5).
Conclusions

Geweke’s (1992) Z, and Rafter and Lewis’s (1992) I statistics showed that all models converged to stationary distributions with values close to 0.05 or smaller indicating overall model convergence among the three primary models (RQ 1.1, RQ 2.1, and RQ 3.1). Neither prior distribution specification nor the test length had any impact on the convergence rates nor did the number of misconception latent variables have any appreciable effect on convergence rates. This evidence indicates the Bayesian MCMC algorithms estimated the model parameters satisfactorily.

The degree of item parameter bias (RQ 1.2, RQ 2.2, and RQ 3.2) was examined via RMSE. Results of the RMSE values for the parameters of correct responses were similar to each other. Discrimination parameters had slightly higher RMSEs than the difficulty parameters. Adding additional misconception latent variables from one to three did not have any impact on the RMSE of the parameters of correct responses. This is partly due to the fact that the relationship between the $LV_I$ and the $MLV$ was zero across all conditions (see Appendix C for bivariate correlations). The parameters associated with the $MLV$, however, did not have constant RMSE estimates across the simulation conditions. For instance, the difficulty parameter RMSEs became larger as the number of $MLV$s increased from one to three. The RMSE ranged from 0.101 to 0.107 when there was a single $MLV$. For the two $MLV$ model, RMSE ranged from 0.101 to 0.121, and in the three $MLV$ model RMSE ranged from 0.105 to 0.119. This is contrasted in the RMSE values for the discrimination parameters which did not evidence the same increases as the number of $MLV$s increased. For the single $MLV$ model, RMSEs ranged from 0.144 to 0.198, in the two $MLV$ model, they ranged from 0.155 to 0.223.
For the three $MLV$ model RMSEs ranged from 0.161 to 0.332. Two explanations are plausible for why the discrimination parameters evidenced larger RMSE values. First, is related to the prior distribution specification of the discrimination parameters. Originally, the discrimination parameters were generated from a lognormal distribution with mean 0 and standard deviation of 0.5 (Table 4). In the estimation part of the simulation, the prior distribution of the discrimination parameters was manipulated as shown in Table 4. This manipulation had some impact on the estimation of the discrimination parameter of the distractors measuring misconceptions. The second plausible explanation is related to the conditional probability specification between a correct and an incorrect response. As Figure 1 illustrates, a response can be a misconception only if it is an incorrect response, and a response can be incorrect only if it is not a correct response. For example, if an item had a difficulty of 0 and also was contained a distractor measuring a misconception with a difficulty of 0 (50% of the time the misconception distractor was selected), the percentage that the distractor was selected can be calculated by multiplying the correct response percentage (0.50) by the misconception selection percentage (0.50), which is equal to 0.25. Thus, the variance of the distractor, 0.25 is multiplied by 0.75 (1.00-0.25), is 0.188 whereas the variance of the correct response, 0.50 multiplied by 0.50 (1-0.50), is 0.25. In other words, although the discrimination and the difficulty parameters of $LVI$ and $MLV$ had similar population parameter distributions, and similar prior distribution specifications, the misconception response probabilities were always bounded by the correct response probabilities, which leads to a larger discrepancy (RMSE) in the estimation of the distractor parameters of $ML$. 

The degree of person parameter bias (RQ 1.3, RQ 2.3, and RQ 3.3) was examined via RMSE values and a bias statistic. Bias statistics were approximately zero for both \( LVI \) and \( MLV \) regardless of the number of \( MLV \)s suggesting partial evidence that person parameters were estimated without bias. However, person parameter RMSEs were not same across models. For \( LVI \), the RMSEs increased as a function of test length regardless of how many misconception latent variables were modeled (0.39, 0.29 for short and long test length). For the \( MLV \), RMSE values increased as the number of \( MLV \) increased from one to three. For the one misconception latent variable condition, the RMSE was around 0.55 for the short test conditions while it was around 0.42 for the long test conditions. The RMSE was around 0.68 for the short test condition while it was around 0.55 for the long test conditions when there were two misconception latent variables. When there were three misconception latent variables, the RMSE was around 0.76 for the short tests, while it was around 0.62 for the long test conditions. Consistent with test theory, longer tests have smaller RMSEs and adding \( MLV \) increases bias regardless of test length.

Average absolute errors were investigated for the three \( MLV \) models: RQ 1.4, RQ 2.4, and RQ 3.4. The AAE for the correct response probability was either 0.05 (short test) or 0.04 (long test) regardless of the number of misconceptions used and the prior distribution. However, for the misconception response probabilities, the AAE was 0.06 (long test) and 0.08 (short test) when there was a single \( MLV \). When there were two \( MLV \)s, AAE increased to 0.08 (long test) or 0.10 (short test) and again for three \( MLV \)s; 0.09 (long test) and 0.11 (short test). It was clear that increasing the number of misconception variables increases the AAE of the misconception response probabilities. Test length has also impacted AAE, such that short test
evidenced higher AAE values that longer test lengths. However, prior distribution specification was minimal and did not seem to affect AAE.

The additional test information collected by modeling misconceptions was addressed in RQ 1.5, RQ 2.5, and RQ 3.5 and estimated by Equation 17. Figures 15, 33, and 55 presented the peak point of the TIFs by simulation replication for 1, 2, and 3 MLVs. From these figures, it is clear that the amount of information estimated from modelling misconceptions was always less than the amount of information estimated from modeling the LVIs. These figures also provide evidence that increasing the number of distractors measuring MLVs also decreases specific MLV test information but the test overall is providing information on more latent variables overall. A likely explanation for the drop in MVL information as MVLs are added is most likely due to the decreasing number of distractors committed to measuring a specific MVL, but this is off set by the increase in MVL measured.

To summarize, the item and distractor parameters \{\theta, \eta, \alpha, \beta, \lambda, \zeta\} showed varying degree of estimation precision. Correct responses parameters \{\alpha \text{ and } \beta\} were similar across the replication conditions but the parameters of the misconception distractors \{\lambda \text{ and } \zeta\} evidenced lower estimation precision as the number of MLVs increased from one to three. Overall, prior distribution had a minimal effect on item parameters for the correct responses and the difficulty parameter of the misconception responses. However, prior distribution effected the discrimination parameter of misconception responses substantially increasing these estimates as MVLs increased. Similarly, the same conclusion can be made for the person parameters. LVIs showed similar amounts of bias and RMSE across the conditions; however, the MVLs had varying levels of RMSEs and as the number of MVLs increased from one to
three, the RMSE of the misconception latent variables increased. The same conclusion summarizes the AAE findings. The AAE for the correct response probabilities were constant across conditions regardless of prior distribution specification and the number of misconception latent variables. However, AAE for the misconception response probability did not change by the prior distribution specification, changing the number of $MV_L$ from one to three increased the AAE. Lastly, the modeling of $ML_V$s helped gain additional test information in all of the conditions. Increasing the number of $ML_V$s dropped the amount of test information per $ML_V$; however, the total amount of test information provided remained at similar levels.

Discussion

Misconceptions are student misunderstandings that manipulate a student’s way of thinking in a systematic way (Confery, 1990; Khazanov, 2008). As such, many qualitative and quantitative studies focusing on misconceptions have focused on either understanding or explaining misconceptions, or trying to remedy them through some form of intervention (instruction, curriculum etc.). Although the majority of misconception studies found in the literature were qualitative, there were a number of quantitative studies too. Those misconception studies that employed a quantitative approach used instruments that purport to measure misconceptions across a range of purposes; diagnostically to see if a student holds one or more misconceptions, for placement pre curriculum, or for the evaluation of classroom instruction trying to reduce or eliminate misconceptions.

Overall, this study employed a complex item response model for modeling student misconceptions as continuous latent variables while modeling a latent variable of interest. A
quantitative framework was built to help researchers to scale the misconception levels of students into continuous metrics. Previous attempts to model student misconceptions using modern test theory included option characteristic curves (Sadler, 1998) which used nominal response model (Bock, 1972), and a model where categorical misconception latent variables were estimated in addition to simultaneous estimation of a latent variable of interest (Bradshaw & Templin, 2014). Sadler’s (1998) approach did not attempt to score misconceptions, rather produce option characteristics curves to visually inspect distractors measuring student misconceptions.

Researchers have applied different item response models to quantify misconceptions. In efforts to quantify misconceptions, two studies deserved critical attention; one was Saddler (1998) study, the other one was Bradshaw and Templin (2014) study. Saddler’s (1998) study used Bock’s (1972) nominal response model to understand misconceptions. This model was able to produce option characteristics curves to see how a distractor measuring a misconception looks like with respect to the underlying latent trait (e.g. ability) although it did not quantify misconceptions. Bradshaw and Templin (2014) attempted to model misconceptions with both item response theory and cognitive diagnostic model applications. This model threatened misconceptions as categorical latent variable. The authors argued that this model could be useful when there is a need to deliver feedback to students. The feedback would include a score on a continuous IRT metric indicating how much a latent trait of interest possessed by a student, in addition to a dichotomous score {0,1} that indicates whether a student possesses a misconception or not.
Providing a student with feedback related to a misconception s/he has is an important piece of instructional information that can complement feedback related to the LVI. However, categorical feedback of a misconception incomplete in itself since it does not effectively communicate the strength or preponderance of the misconception held by the student. Past misconception assessment practices have attempted to solve this limitation by developing misconception inventories from the perspective of CTT. This study offers a psychometric framework which could be used for developing a misconception instruments that are capable of placing learners on both continuums of the LVI and the MLV. The continuous misconception scores derived from the estimation of the model proposed could be used for assessing the quality of the inventories that were developed to remedy student misconceptions.

Attempts to quantify misconceptions as continuous latent variables while quantifying the latent ability of interest simultaneously have primarily been via CTT models. There is some, although limited, attention in the literature addressing misconception modeling from the IRT perspective. This may be in part due to the forced-choice nature of the response. As previously illustrated, modeling a misconception can only begin after an incorrect response choice was made by the student. Thus the misconception response if conditional on the probability of the first response. Suh and Bolt (2010) recently showed how the nested logit model could be employed to model such choice behavior. To address the research question posed in this study, two things needed to occur: first, a misconception needed to be modeled as a continuous latent variable and 2) the modeling needed to follow choice behavior. It was therefore natural to extend the nested logit model to allow for estimation of continuous person and item parameter in the lower model in Suh and Bolt’s nested logit model.
By using the model proposed in this dissertation, researchers could develop assessments that better inform test users and test examinees by taking advantage of the parameter estimation specific to the IRT model proposed, e.g., item location, item discrimination, distractor location, and distractor discrimination. For example, the discrimination parameters (item or distractor) may be used to indicate the quality of the items/distractors. Items/distractors with discriminations negative or zero may be considered for revision or they can be removed from the inventory. Visual tools such as item/distractor characteristic curves could be used to display item and distractors together which eventually may be used to make the decision at which location of the latent continuum need most misconception treatment/intervention. Further, test information functions are excellent tools which show the range in which a test is most useful. Test information functions and the standard error of measurement could be used to guide the instrument development process of misconception assessments. For example, the test items could be developed in a way that test information function could be maximized at the location of LVI/MLV that needs most attention. The model developed in this study opens up many possibilities for researchers to consider the reality that multiple misconceptions do exist and that they can be estimated with adequate precision.

The framework provided by this study will produce continuous student ability scores in addition to continuous student misconception scores. Continuous quantification of both the LVI and the MVL could serve a very useful function in studies where researchers are focused on estimating the outcomes of an intervention (e.g. instruction, curriculum). Through analysis of the changed within a student from misconception latent variable processing to correct latent
variable processing, teachers will be able to more critically evaluate the intra-individual transition from incorrect to correct processing.

Limitations

This study evaluated two prior distribution specifications when estimating both item ($\alpha, \beta, \zeta, \text{and } \lambda$), and person ($\eta$ and $\theta$) parameters. The specification of the misconception category priors might limit the findings of this study since the specification of the priors arbitrarily set to either normal or uniform while the true values were assumed to be from a lognormal distribution. This arbitrariness is due to the fact that there is no real evidence with regard to empirical range of the discrimination parameter category that measures an MLV in the literature. In addition, the priors were manipulated only for the discrimination parameters as displayed in Table 4. This is a limitation because it was assumed that the prior specification for the item/distractor location parameters as well as ability/misconception parameters were the same as the original population parameters. If the prior distributions are not properly specified, the AAE and the RMSEs may have different size (larger) and direction (negative or positive) depending on the departure from the true/empirical distributions. Although, the initial (true) misconception category discrimination parameter values used in this study and the two arbitrarily selected prior distributions cover the empirical range of a test item found in typical IRT applications. Using true/prior item discrimination distributions for the misconception category true/prior distributions is another limitation because there was no investigation related to how location/discrimination parameters could be distributed in this simulated instrument.
Another limitation of this study is the lack of a real test dataset which postulates to measure both a LVI and one or more MLVs. Given a real dataset, a prior distribution can be postulated from the empirical evidence. Unfortunately, past misconception assessment practices have not provided this information. Although the model may converge and estimation occur as in this study, having a real dataset will not ensure that the estimation is free of bias. Parameter estimation bias may be different in an actual test relative to a simulated test.

Lastly, this study used a Bayesian estimation algorithm to estimate item/distractor/person parameters. Future research might consider other estimation algorithms such as maximum likelihood estimation. Bayesian estimation is useful for small sample situations, which actually fit to the misconception assessment due to having small (single classroom size of 30) to medium size datasets (classrooms totaling up to 200). However, Bayesian estimation is biased towards the prior selection, and the prior specification needs to be based on empirical evidence as expressed above.

Recommendations for Future research

The most important recommendation is an actual test data be collected from a test measuring one or more MLVs in order to evaluate the contribution to understanding student misconceptions relative to the LVI. For example, are the estimates for the LVI derived from the 2PL-2PL-NLMM model consistent with those derived from a more typical 2PL model?

Second, in this study, the prior distributions of the discrimination parameters were manipulated. Both of the prior distributions (item and distractor) covered the empirical range of possible discrimination values but future research may investigate the impact of using different prior distribution forms on the estimation of discrimination parameters. For instance,
the variance of the true misconception discrimination parameters was fixed at 0.5 with a lognormal distribution, and the prior distribution specifications had variance of 0.3 for the normal distribution prior specification conditions. Different variances could be considered and investigated such as smaller variance as 0.2, or 0.1.

Third, the test length used in this study was either 25 or 50 items. Other test lengths should be considered in the future research applications. Similarly, the examinee sample size was fixed at 2000, so this should be varied in future research applications to see how the changes in sample size impact the estimation precision of this model.

Fourth, this study examined misconceptions using a 2PL IRT model in both lower and upper hierarchy of Equation 6. Future research may examine other combinations of item response models such as a 1PL model on the lower hierarchy, and a 2PL IRT model in the upper hierarchy of Equation 6 or vice versa.

Fifth, there could be instances where the MLVs might correlate. If it is the case, then multidimensional item response modeling approach might fit better to modeling misconceptions. Future research might take this possibility into consideration.

Sixth, this study did not fully investigate methods for combining the item information functions for the LVI and MVL. This study demonstrated that an item information function for the MVL can be developed in a similar manner to that of the LVI. But how these competing information functions may be combined is left for future research.

Lastly, the test design simulated in this study kept a strict one to one item to misconception ratio at 1:1. Thus future simulations could examine the impact of increasing this ration to many to one on estimation precision.
REFERENCES


170


APPENDICES

Appendix A

R codes for data generation and estimation

#2D (t1,t2,t3; 1:12,13:25) Bugs 25

setwd("/Users/macbookpro/Desktop/simulation/2D/25/normal")

##utilities

library(R2WinBUGS)
library(R2jags)
library(coda)
library(boot)
library(R2jags)
#two dimentional model

one <- function() {
  for (i in 1:N) {
    for (j in 1:J) {
      r[i,j]~dcat(p[i,j,1:3])
    }
  }
}

for (i in 1:N) {
  for (j in 1:12) {
    p[i,j,3]<-exp(a1[j]*t1[i]+b1[j])/(1+exp(a1[j]*t1[i]+b1[j]))
    num[i,j,2]<-exp(a2[j]*t2[i]+b2[j])/(1+exp(a2[j]*t2[i]+b2[j]))
  }
}
\begin{verbatim}
p[i,j,2] <- (1-p[i,j,3])*num[i,j,2]
p[i,j,1] <- (1-(p[i,j,2]+p[i,j,3]))
}
}
for (i in 1:N) {
    for (j in 13:25) {
        p[i,j,3] <- exp(a1[j]*t1[i]+b1[j])/(1+exp(a1[j]*t1[i]+b1[j]))
        num[i,j,2] <- exp(a2[j]*t3[i]+b2[j])/(1+exp(a2[j]*t3[i]+b2[j]))
        p[i,j,2] <- (1-p[i,j,3])*num[i,j,2]
        p[i,j,1] <- (1-(p[i,j,2]+p[i,j,3]))
    }
}
}

for (i in 1:N){
    t1[i] ~ dnorm(0,1)
    t2[i] ~ dnorm(0,1)
    t3[i] ~ dnorm(0,1)
}
}
for (j in 1:J) {
    a1[j] ~ dnorm(1.5,0.3)
    b1[j] ~ dnorm(0.01,1)
}
\end{verbatim}
for (j in 1:12) {
    a2[j]~dnorm(1.5,0.3)
    b2[j]~dnorm(0.5,1)
}

for (j in 13:25) {
    a2[j]~dnorm(1.5,0.3)
    b2[j]~dnorm(0.5,1)
}

write.model(one, con = "twoD.bug", digits = 5)

n.exams=2000
n.items=25
test.diff=.001
misc.diff=.5
rep=50
library(MASS)
seeds=sample(1:10000, rep, replace = F)
for (k in 41:50) {
    set.seed(seeds[k])
    prob0 = matrix(rep(NA,n.exams*n.items),nrow=n.exams,ncol=n.items)
    resp = matrix(rep(NA,n.exams*n.items),nrow=n.exams,ncol=n.items)
    oneD = matrix(rep(NA,n.exams*n.items),nrow=n.exams,ncol=n.items)
    prob1 = matrix(rep(NA,n.exams*n.items),nrow=n.exams,ncol=n.items)
    beta1 = rnorm(n.items,test.diff,1)
    beta2 = c(rnorm(12,misc.diff,1),rnorm(13,misc.diff,1))
    alpha1 =rlnorm(n.items,0.01,.5)
    alpha1[alpha1>4.5]=4.5
    alpha2=c(rlnorm(12, 0.01, .5), rlnorm(13, 0.01,.5))
    alpha2[alpha2>4.5]=4.5
    Sigma2 <- matrix(c(1,0,0,
                       0,1,0,
                       0,1,0
                      ),3,3)
    x=mvrnorm(n.exams, mu=c(0,0,0), Sigma = Sigma2)
    teta1 = x[,1]
    teta2 = x[,2]
    teta3 = x[,3]
    for (i in 1:n.exams){
        for (j in 1:n.items){
            ...
        }
    }
}
prob0[i,j]<-exp(alpha1[j]*teta1[i]+beta1[j])/(1+exp(alpha1[j]*teta1[i]+beta1[j]))

resp[i,j] = rbinom(1,1,prob0[i,j])

}

for (i in 1:n.exams){
  for (j in 1:12){
    if (resp[i,j] == 1) {
      oneD[i,j]<-3
    }
    else if (resp[i,j] == 0)
    {
      prob1[i,j]<- exp(alpha2[j]*teta2[i]+beta2[j])/(1+exp(alpha2[j]*teta2[i]+beta2[j]))
      oneD[i,j]<- rbinom(1,1,prob1[i,j])
    }
  }
}

for (i in 1:n.exams){
  for (j in 13:25){
    if (resp[i,j] == 1) {
      oneD[i,j]<-3
    }
    else if (resp[i,j] == 0)
    {
      prob1[i,j]<- exp(alpha2[j]*teta3[i]+beta2[j])/(1+exp(alpha2[j]*teta3[i]+beta2[j]))
      oneD[i,j]<- rbinom(1,1,prob1[i,j])
    }
  }
}
N=nrow(oneD)
J=ncol(oneD)
r= oneD
r[r==1]=2
r[r==0]=1

# Save generated item response file
write.csv(r, file = paste("data", k,".csv", sep = "))
data=list("N", "J", "r")

# R2jugs
rjags=jags(data=data, inits = NULL, n.chains = 1, n.iter = 5500,
n.thin = 1, model.file = "twoD.bug", n.burnin = 500, DIC = T,
parameters.to.save=c("a1","b1","a2","b2","t1","t2","t3"))

## for mcmc
mcmc=as.mcmc(rjags)

# Convergence
raftery=raftery.diag(mcmc)

I.percentage=length(raftery[[1]]$resmatrix[,4][raftery[[1]]$resmatrix[,4]>5])/length(raftery[[1]]
$resmatrix[,4])
geweke=geweke.diag(mcmc)
geweke1=data.frame(geweke[[1]]$z)
gewk=(length(geweke1[geweke1<=1.96])+length(geweke1[geweke1>=1.96]))/length(geweke1$geweke..1...z)

#AAE

# average absolute error for P(X=3)

p.obs=matrix(NA, nrow = N, ncol = J )
p.est=matrix(NA, nrow = N, ncol = J )

# average absolute error for P(X=2)

p.obs.m=matrix(NA, nrow = N, ncol = J )
p.est.m=matrix(NA, nrow = N, ncol = J )

for (i in 1:N) {
  for (j in 1:12) {
    # average absolute error for P(X=3)
    p.est[i,j]=exp(rjags$BUGSoutput$mean$a1[j]*rjags$BUGSoutput$mean$t1[i]+rjags$BUGSoutput$mean$b1[j])/(1+exp(rjags$BUGSoutput$mean$a1[j]*rjags$BUGSoutput$mean$t1[i]+rjags$BUGSoutput$mean$b1[j]))
    p.obs[i,j]=exp(alpha1[j]*teta1[i]+beta1[j])/(1+exp(alpha1[j]*teta1[i]+beta1[j]))
    # average absolute error for P(X=2)
    p.est.m[i,j]=exp(rjags$BUGSoutput$mean$a2[j]*rjags$BUGSoutput$mean$t2[i]+rjags$BUGSoutput$mean$b2[j])/(1+exp(rjags$BUGSoutput$mean$a2[j]*rjags$BUGSoutput$mean$t2[i]+rjags$BUGSoutput$mean$b2[j]))
  }
}
for (i in 1:N) {
    for (j in 13:25) {
        # average absolute error for P(X=3)
        p.est[i,j] = exp(rjags$BUGSoutput$mean$a1[j]*rjags$BUGSoutput$mean$t1[i]+rjags$BUGSoutput$mean$b1[j])/(1+exp(rjags$BUGSoutput$mean$a1[j]*rjags$BUGSoutput$mean$t1[i]+rjags$BUGSoutput$mean$b1[j]))
        p.obs[i,j] = exp(alpha1[j]*teta1[i]+beta1[j])/(1+exp(alpha1[j]*teta1[i]+beta1[j]))
        # average absolute error for P(X=2)
        p.est.m[i,j] = exp(rjags$BUGSoutput$mean$a2[j]*rjags$BUGSoutput$mean$t3[i]+rjags$BUGSoutput$mean$b2[j])/(1+exp(rjags$BUGSoutput$mean$a2[j]*rjags$BUGSoutput$mean$t3[i]+rjags$BUGSoutput$mean$b2[j]))
        p.obs.m[i,j] = exp(alpha2[j]*teta3[i]+beta2[j])/(1+exp(alpha2[j]*teta3[i]+beta2[j]))
    }
}
obs.prob.theta=data.frame(p.obs)
write.csv(obs.prob.theta, file = paste("obs.prob.theta", k,".csv", sep = ","))

obs.prob.misc=data.frame(p.obs.m)
write.csv(obs.prob.misc, file = paste("obs.prob.misc", k,".csv", sep = ","))

est.prob.theta=data.frame(p.est)
write.csv(est.prob.theta, file = paste("est.prob.theta", k,".csv", sep = ","))

est.prob.misc=data.frame(p.est.m)
write.csv(est.prob.misc, file = paste("est.prob.misc", k,".csv", sep = ","))

o.a1=alpha1
o.a2=alpha2
o.b1=beta1
o.b2=beta2

o.t1=teta1
o.t2=teta2
o.t3=teta3

a1=rjags$BUGSoutput$mean$a1
a2=rjags$BUGSoutput$mean$a2
b1=rjags$BUGSoutput$mean$b1
b2=rjags$BUGSoutput$mean$b2
t1=rjags$BUGSoutput$mean$t1


t2=rjags$BUGSoutput$mean$t2


t3=rjags$BUGSoutput$mean$t3


ppars=data.frame(t1,t2,t3,o.t1,o.t2,o.t3)

write.csv(ppars, file = paste("ppars", k,".csv", sep = ""))


pars=data.frame(a1,a2,b1,b2,o.a1,o.a2,o.b1,o.b2)

write.csv(pars, file = paste("ipars", k,".csv", sep = ""))


conv=data.frame( I.percentage,gewk)

write.csv(conv, file = paste("model.summary", k,".csv", sep = ""))


}
Appendix B

Item characteristic curves as functions of LVI and $MLV$s for the probabilities of correct choice, and the distractors measuring a misconception

Figure 56. Item/distractor characteristics curves as functions of latent variables when there was an LVI and an MLV
Figure 57. LVI is 1D; MLV is 2D

Item/distractor characteristics curves as functions of latent variables when there was an LVI and two MLVs
Figure 58. *Item/distractor characteristics curves as functions of latent variables when there was an LVI and three MLVs*
Appendix C

Correlations among the latent variables for all conditions

Normal distribution with 25 items

Normal distribution with 50 items

Uniform distribution with 25 items

Uniform distribution with 50 items
Figure 59. Trait correlations when there was a single MLV
Figure 60. *Trait correlations for short test forms when there were two MLVs (Normal prior)*

Figure 61. *Trait correlations for long test forms when there were two MLVs (Normal prior)*
Figure 62. Trait correlations for short test forms when there were two MLVs (Uniform prior)
Figure 63. Trait correlations for long test forms when there were two MLVs (Uniform prior)
Figure 64. Trait correlations for short test forms when there were three MLVs (Normal prior)
Figure 65. Trait correlations for long test forms when there were three MLVs (Normal prior)
Figure 66. Trait correlations for short test forms when there were three MLVs (Uniform prior)
Figure 67. Trait correlations for long test forms when there were three MLVs (Uniform prior)