Using the Score-based Testlet Method to Handle Local Item Dependence

Author: Wei Tao

Persistent link: http://hdl.handle.net/2345/1363

This work is posted on eScholarship@BC, Boston College University Libraries.

Boston College Electronic Thesis or Dissertation, 2008

Copyright is held by the author, with all rights reserved, unless otherwise noted.
BOSTON COLLEGE
Lynch School of Education

Department of
Educational Research, Measurement, and Evaluation

USING THE SCORE-BASED TESTLET METHOD TO HANDLE LOCAL ITEM DEPENDENCE

Dissertation

by

WEI TAO

submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

May 2008
Abstract of Student Research

USING THE SCORE-BASED TESTLET METHOD TO HANDLE LOCAL ITEM DEPENDENCE

Dissertation by: Wei Tao

Dissertation Chair: Prof. Larry H. Ludlow

Item Response Theory (IRT) is a contemporary measurement technique which has been used widely to model testing data and survey data. To apply IRT models, several assumptions have to be satisfied. Local item independence is a key assumption directly related to the estimation process. Many studies have been conducted to examine the impact of local item dependence (LID) on test statistics and parameter estimates in large-scale assessments. However, in the healthcare field where IRT is experiencing greater popularity, few studies have been conducted to study LID specifically.

LID in the healthcare field bears some unique characteristics which deserve separate analysis. In health care surveys, it is common to see several items that are phrased in a similar structure or items that have a hierarchical order of difficulties. Therefore, a Guttman scaling pattern, or a deterministic response pattern, is observed among those items. The purposes of this study are to detect whether the Guttman scaling pattern among a subset of items exhibit local dependence, whether such dependence has any impact on test statistics, and whether these effects differ when different IRT models are employed. The score-based approach - forming locally dependent dichotomous items into a polytomous testlet - is used to accommodate LID.

Results from this dissertation suggest that the Guttman scaling pattern among a
subset of items does exhibit moderate- to high-degree of LID. However, the impact of this special LID is minimal on internal reliability estimates and on the unidimensional data structure. Regardless of which models are employed, the dichotomously-scored person ability estimates are highly correlated with the polytomously-scored person ability estimates. However, the impact of this special LID on test information differs between Rasch models and non-Rasch models. Specifically, when only Rasch models are involved, test information derived from the LID-laden data is underestimated for non-extreme scores; whereas, when non-Rasch models are used, the opposite finding is reached –that is, LID tends to overestimate test information.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>ACKNOWLEDGEMENTS</th>
<th>vi</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF TABLES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vii</td>
</tr>
</tbody>
</table>

## CHAPTER ONE: INTRODUCTION .................................................................1
- Measurement Theories ................................................................. 2
- Statement of the Problem .............................................................. 7
- Purposes and Research Questions .................................................. 14
- Significance of the Study .............................................................. 15
- Summary ....................................................................................... 16

## CHAPTER TWO: REVIEW OF LITERATURE ...............................................18
- Definitions of Local Item Independence ....................................... 19
- Causes of Local Item Dependence .................................................. 20
- Types of Local Item Dependence .................................................... 24
  - Positive versus Negative Dependence ...................................... 24
  - Order versus Combination Dependence .................................... 25
  - Underlying versus Surface Dependence .................................... 25
- Methods of Assessing Local Item Dependence ............................... 28
  - Dimensionality Assessment ..................................................... 28
    - Exploratory Factor Analysis .............................................. 29
    - Categorical Confirmatory Factor Analysis ............................ 31
    - Full-information Item Factor Analysis ................................. 33
    - Bi-factor Item Factor Analysis .......................................... 34
    - NOHARM Residual Analysis ................................................. 35
- Inter-item Covariance Analysis ..................................................... 36
- Conditional Inter-item Correlation Analysis .................................. 36
  - Yen’s Q3 Statistic ................................................................. 37
  - Inter-item Raw Score Correlation ............................................ 41
- Fit-statistic LID Measures ($G^2$ and $X^2$) .................................. 42
  - The Null Distributions of $G^2$ and $X^2$ Measures .................... 45
# The Power of G² and X² Measures

- Consequences of Local Item Dependence: 48
- Impact of LID on Item Parameter Estimates: 48
- Impact of LID on Item Characteristic Curve: 50
- Impact of LID on Internal Reliability: 51
- Impact of LID on Test Information: 53
- Impact of LID on Ability Estimates: 55

## Approaches to Managing Local Item Dependence

- The Score-Based Approach: 58
- The Item-Based Approach: 64

## Testlet Response Theory Model

- Full-information Bi-factor Model: 69

## Item-interaction Model

## Approaches Employed in Health Care

- Gaps in the Literature: 75
  - Inadequacy of Research on Order Dependence: 76
  - Inadequacy of Research on 1-Parameter IRT Models: 77

## Summary

### CHAPTER THREE: RESEARCH DESIGN

- Data Description: 81
- Instrument: 81
- Population and Sample: 82
- Hypotheses: 83
- Selection of an Approach to Manage LID: 86
- IRT Models: 88
- Simulated Data: 91
- Analyses:
  - Analysis 1: Exploratory Analyses: 93
  - Analysis 2: LID Assessment: 98
  - Analysis 3: Dimensionality Assessment: 104
  - Analysis 4: Internal Reliability Assessment: 104
  - Analysis 5: Impact of LID on Test Information Function: 105
  - Analysis 6: Impact of LID on Score Estimates: 110
- Summary: 111
CHAPTER FOUR: RESULTS .......................................................................................................................... 112
  Analysis 1: Exploratory Analyses .......................................................................................................... 113
  Analysis 2: LID Assessment .................................................................................................................. 137
  Analysis 3: Dimensionality Assessment ............................................................................................. 142
  Analysis 4: Internal Reliability Assessment ....................................................................................... 148
  Analysis 5: Impact of LID on Test Information Function .................................................................... 150
  Analysis 6: Impact of LID on Score Estimates .................................................................................... 173
  Summary .............................................................................................................................................. 183

CHAPTER FIVE: CONCLUSIONS .................................................................................................................. 184
  Summary of Findings ............................................................................................................................ 187
  Discussions ......................................................................................................................................... 192
  Implications ......................................................................................................................................... 198
  Limitations .......................................................................................................................................... 203
  Future Research Suggestions ............................................................................................................... 204
  Summary .............................................................................................................................................. 206

REFERENCES ............................................................................................................................................ 207

APPENDIX A: THE PEDI – MOBILITY DOMAIN ..................................................................................... 217
ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who have supported me during this doctoral work. First of all, I want to thank the Health and Disability Research Institute at Boston University Medical Center for allowing me to conduct the research on their data and for giving me ever important practical training in the measurement field. I also want to thank Lynch School of Education at Boston College for awarding me the Dissertation Development Grant which financially supported me at the beginning of my dissertation.

I am very grateful for having an exceptional doctoral committee. I especially wish to thank Dr. Larry Ludlow, my advisor and dissertation chair, for his continuous support over the past few years as well as his stimulating suggestions and encouragement during all the research. My gratitude goes to all my readers, Dr. Henry Braun, Dr. Tzur Karelitz, and Dr. Steve Haley, for their insightful thoughts into the research design, the analyses, and the writing style, and for their time and commitment, which by all means made this dissertation a much better piece of work.

I owe a special note of gratitude to Dr. Pengsheng Ni who was always available and willing to help me out of any technical obstacle I encountered. I also want to thank
Mrs. Tracy McMahon and Ms. Diane Chen for all their assistance in polishing the writing.

My gratitude goes to all my friends whose encouragements have accompanied me through the long journey. I would like to give my special thanks to my family, my eternal cheerleader whose patient love enabled me to complete this work. I especially thank my husband, my best friend, for the moral support that he has given me all these years.
LIST OF TABLES

Table 3.1: Content Coverage in the PEDI Mobility Domain ...........................................82
Table 3.2: Dataset Names ................................................................................................93
Table 3.3: IRT Models Fitted to the Real Data.................................................................96
Table 3.4: Sets of Contrasts between Dichotomous and Polytomous data .....................106
Table 4.1: Guttman Scaling Analysis, by Testlet ..............................................................114
Table 4.2: CTT Analysis on the Dichotomous Raw Data ..................................................119
Table 4.3: CTT Analysis on the Polytomous Raw Data ....................................................120
Table 4.4: Item Parameter Estimates and Fit Statistics – Dichotomous Rasch Model ..........125
Table 4.5: Step Parameter Estimates and Fit Statistics – PCM ........................................126
Table 4.6: Thurstone Threshold Estimates – PCM .............................................................127
Table 4.7: Item Parameter Estimates and Fit Statistics - 1PL Model ...............................131
Table 4.8: Thurstone Threshold Estimates and Fit Statistics – GRM with Equal Slopes .................................................................132
Table 4.9: Item Parameter Estimates and Fit Statistics - 2PL Model ...............................135
Table 4.10: Thurstone Threshold Estimates and Fit Statistics – GRM with Varying Slopes ...................................................................................136
Table 4.11: Within-Content $Q_3$ values, Derived from the Rasch Model.......................140
Table 4.12: Within-Content $Q_3$ values, Derived from the 2PL Model ............................141
Table 4.13: Eigenvalues from Factor Analysis on the Dichotomous Data .......................145
Table 4.14: Eigenvalues from Factor Analysis on the Polytomous Data ...........................147
LIST OF FIGURES

Figure 3.1: $Q_3$ Matrix Structure Partitioned by Content Sub-matrixes ............................................. 101
Figure 3.2: Calculation of the Average Within-content $Q_3$ ................................................................. 101
Figure 3.3: Calculation of the Average Between-content $Q_3$ ............................................................... 102
Figure 4.1: Dichotomous Raw Data: Distributions of Item Easiness and Item-Total Score Correlations ........................................................................................................... 118
Figure 4.2: Polytomous Raw Data: Distributions of Adjusted Testlet Averages and Testlet-Total Score Correlations ......................................................................................... 118
Figure 4.3: Scatter Plot of Item Easiness and Item-Total Score Correlations ................................. 118
Figure 4.4: Scatter Plot of Item Difficulty Estimates from Dichotomous Rasch Model versus Item Step Estimates from PCM .......................................................... 126
Figure 4.5: Scatter Plot of Item Difficulty Estimates from Dichotomous Rasch Model versus Thurstone Threshold Estimates from PCM .......................................................... 127
Figure 4.6: Scatter Plot of Item Difficulty Estimates from 1PL Model versus Thurstone Threshold Estimates from GRM with Equal Slopes ......................................................... 132
Figure 4.7: Comparison of Thurstone Threshold Estimates between the PCM and the GRM with Equal Slopes ............................................................................................... 133
Figure 4.8: Scatter Plot of Item Difficulty Estimates from 2PL Model versus Thurstone Threshold Estimates from GRM with Varying Slopes ......................................................... 136
Figure 4.9: Distribution of Within- and Between-$Q_3$, Derived from the Rasch Model .......................................................... 140
Figure 4.10: Distribution of Within- and Between-$Q_3$, Derived from the 2PL Model .......................................................... 141
Figure 4.11: Scree Plot of the Dichotomous Data ......................................................................................... 146
Figure 4.12: Scree Plot of the Polytomous Data ............................................................................................. 147
Figure 4.13: Alpha Coefficients for the Dichotomous and the Polytomous Data .......................................................... 149
Figure 4.14: Scatter Plots of Item Difficulty Estimates against Thurstone Threshold Estimates, before and after Person Score Anchoring ......................................................... 152
Figure 4.15: Total Test Information Comparison: Contrast 1 – Rasch Models ........................................ 159
Figure 4.16: Total Test Information Comparison: Contrast 2 – 1PL vs. GRM with Equal Slopes ......... 159
Figure 4.17: Total Test Information Comparison: Contrast 3 – 2PL vs. GRM with varying Slopes ............ 159
Figure 4.18: Test Information Derived from Dichotomous and Polytomous
Figure 4.19: Test Information Derived from Dichotomous and Polytomous Scoring Schemes, by Testlets - Contrast 1: Rasch Models

Figure 4.20: Test Information Derived from Dichotomous and Polytomous Scoring Schemes, by Testlets - Contrast 2: 1PL vs. GRM with Equal Slopes

Figure 4.21: Score Estimates from the Dichotomous and the Polytomous Scoring Methods

Figure 4.22: Standard Error of Estimates from the Dichotomous and the Polytomous Scoring Methods - Contrast 1: Rasch Models

Figure 4.23: Standard Error of Estimates from the Dichotomous and the Polytomous Scoring Methods - Contrast 2: 1PL vs. GRM-Equal Slopes

Figure 4.24: Standard Error of Estimates from the Dichotomous and the Polytomous Scoring Methods - Contrast 3: 2PL vs. GRM with Varying Slopes
Chapter One: Introduction

This chapter starts with an introduction to two commonly used measurement theories - Classical Test Theory (CTT) and Item Response Theory (IRT). In the past two decades, Item Response Theory has experienced great popularity in practice because of its capability of addressing several problems encountered in CTT. However, some strong assumptions have to be satisfied in order to generate valid results using IRT models. Local item independence is a key assumption directly related to the estimation process. Previous research examining the impact of violating the local item independence assumption will be reviewed briefly in this chapter.

This study intends to examine one type of local item dependence common in health surveys but inadequately studied previously. The purpose statement, specific research questions to be answered, and significance of the current study are presented in this chapter.
Measurement Theories

Two theories dominate the field of measurement, Classical Test Theory (CTT) and Item Response Theory (IRT). Classical Test Theory, also called Classical True-Score Theory, has been widely used for many decades since Spearman (1904, 1913) first laid the foundation for the theory. It is a simple and quite useful model that describes how errors of measurement can influence observed scores (Allen & Yen, 1979). The essence of Spearman’s model was that any observed test score could be envisioned as the composite of two hypothetical components – a true score and a random error component, expressed in the form of \( X = T + E \), where \( X \) represents the observed score, \( T \) the true score and \( E \) the random error score (Crocker & Algina, 1986).

Advantages of CTT models are that they are based on relatively weak assumptions (i.e., they are easy to meet in real test data) and they are well-known and have a long track record (Hambleton & Jones, 1993). The major disadvantage is that item parameter estimates (item difficulty) are dependent on examinees and person parameter estimates (raw score) are dependent on the test. These dependencies limit the utility of the person and item statistics in practical test development work and complicate analyses (Hambleton & Jones, 1993).

Facing the disadvantages of Classical Test Theory, researchers started to work on a theory which could generate invariant item and person estimates independent of the test
forms and the testing sample. Item Response Theory (IRT), with a former name of Latent Trait Theory, was introduced systematically by Rasch (1960), Birnbaum (1968), and Lord and Novick (1968), and has the potential to address the disadvantages of Classical Test Theory.

Item Response Theory is a collection of statistical models that specify the relationship between examinees’ testing performances and the unobservable latent factors through certain mathematical functions. Several synonyms of the latent “factors” which have been used in the literature include traits, abilities, skills, proficiencies, constructs and attributes. These terms will be used interchangeably in this dissertation. Despite these different name tags, they share similar characteristics in that they all refer to an unobservable, latent “thing” which cannot be measured directly but can only be inferred from observations (Crocker & Algina, 1986). There can be one ability or many abilities underlying the test performance; and there are many ways (i.e., models) to specify the relationship between item responses and the underlying abilities (Hambleton, Swaminathan, & Rogers, 1991).

Item Response Theory gives us great flexibility in both handling different item formats and modeling the ability-performance relationships. For example, we can construct unidimensional or multidimensional IRT models based on the number of abilities underlying the performance; we can construct dichotomous IRT models to
analyze binary items (i.e., multiple-choice questions) or polytomous IRT models to analyze items with more than two response options (i.e., constructed response items allowing for partial credit); and we can specify different mathematical functions to describe the ability-performance relationship. The two most common mathematical forms used to model such a relationship are the normal ogive function and the logistic function. No one has shown that either one fits mental test data significantly better than the other (Lord, 1980). Regular IRT models using either form are equivalent except for the scaling difference. But the logistic models are used more often in current studies due to some computational advantages over the normal ogive models.

IRT models can also be distinguished according to the number of item parameters specified in the model. The most often used dichotomous models are the one-, two- and three-parameter models. The one-parameter model contains only one item parameter -- item difficulty. All items in a test differ in nothing else but the difficulty level. In the two-parameter model, items differ in the difficulty level and in the extent to which they discriminate among persons with very similar abilities. This second item parameter is called item discrimination, or the slope parameter. And finally, the three-parameter model introduces an additional pseudo-guessing parameter meaning that even zero-ability examinees have a higher than zero probability of getting a correct answer to multiple-choice items, due to guessing.
Compared with CTT models, IRT models have several advantages such as their capability of: 1) generating sample-independent item calibrations and test-independent person ability estimates, 2) generating scores at the interval level, 3) generating measurement errors at different ability levels, and 4) placing item and person parameter estimates on the same metric. Major disadvantages of the IRT models are that 1) the concept of the theory is less straightforward than the classical test theory; 2) estimation of IRT parameters requires higher computer performance, availability of special software, and larger sample sizes; and 3) several key assumptions (i.e., local item independence, dimensionality, monotonicity, and continuity of the latent traits) have to be satisfied.

Satisfaction of the assumptions is extremely important for generating valid parameter estimates. IRT models are sometimes referred to as strong stochastic models due to the fact that strong assumptions have to be met; whereas the CTT model is sometimes referred to as a weak stochastic model. Satisfaction of key assumptions can also ensure a good fit between the proposed model and the observed data.

There are four basic assumptions common to all IRT models: 1) local item independence, 2) continuity of the latent trait, 3) monotonicity, and 4) dimensionality. The continuum assumption means that the measurement level of the underlying latent trait(s) should be continuous, rather than discrete. The monotonicity assumption refers to the positive relationship between the underlying trait of the examinees and their test
performance. Examinees with a higher latent trait level will have a higher probability of answering an item correctly.

The dimensionality assumption says that the number of latent abilities (dimensions) have to be specified correctly before any model is established. For unidimensional IRT models, it is assumed that there is one underlying dimension that influences the expected responses of test takers. This single dimension could be a single psychological trait or a fixed composite of several traits. For multidimensional IRT, the number of latent traits (dimensions) influencing test performances has to be specified correctly before parameter estimation.

The local item independence assumption is closely associated with the dimensionality assumption. Local item independence means that “when the abilities influencing test performance are held constant, examinees’ responses to any pair of items are statistically independent” (Hambleton, Swaminathan, & Rogers, 1991, p.10). In the case of a unidimensional IRT model, the complete latent space contains only one factor, or one ability, which determines the expected response pattern exclusively. Once we hold this factor constant, what is left is completely random and thus no other factor or ability can be used to predict examinees’ responses. Similarly, in the case of multidimensional IRT, two or more abilities define the entire latent space. Examinees with the same ability levels tend to answer items in the same way, allowing for random errors. Therefore, when
unidimensionality is satisfied in a unidimensional IRT, or when dimensions for multidimensional IRT are correctly specified, local item independence will be satisfied. Lord (1980) pointed out that in this sense, the two concepts (dimensionality vs. local item independence) are equivalent. Hambleton, Swaminathan, and Rogers (1991) also noted that “local independence will be obtained whenever the complete latent space has been specified; that is, when all the ability dimensions influencing performance have been taken into account” (p.10). Local item dependence (LID) is said to be present when items are not locally independent.

Statement of the Problem

Although IRT originated and has been widely applied in an educational and psychological testing context, its advantages have also been appreciated in other fields, e.g., health care. To date, IRT methods, largely the one- and two-parameter models, have been used in a wide variety of applications in health care, ranging from 1) identification of differential item functioning (DIF), to 2) test equating, to 3) computerized adaptive testing (McHorney & Monahan, 2004). Item Response Theory has become a mainstream measurement theory in the health care field.

In order to generate precise and accurate parameter estimates from various IRT models, important assumptions introduced from the previous section must be satisfied. Local item independence (LII) is a key assumption because it is directly related to the
parameter estimation process. To be specific, the formation of the likelihood function, which is used to find the most likely estimates of item and ability parameters, relies upon the local independence assumption explicitly.

During the past two decades, extensive research was conducted to examine LID related problems. Yen (1984, 1993), for example, identified many situations in which LID is likely to occur. Yen noted that local dependence can stem from content factors, speededness, passage-based items, fatigue, and practice effects. Some research proposed methods to assess the existence of LID (Chen & Thissen, 1997; Holland & Rosenbaum, 1986; McLeod, Swygert, & Thissen, 2001; Yen, 1984, 1993). Other research identified consequences of LID for parameter estimation in IRT (Masters, 1988; Reese, 1995, 1999; Yen, 1993). The common findings from these studies are that moderate- to high-level LID may lead to an overestimation of the slope parameter, information value, and test reliability. Facing the problems of LID, some researchers have attempted to build new models to account for LID so that it might be allowed to occur (Gibbons & Hedeker, 1992; Hoskens & De Boeck, 1997; Jannarone, 1986; Tuerlinckx & De Boeck, 1998; Wainer, Bradlow, & Wang 2007; Wilson & Adams, 1995). Other researchers (Rosenbaum, 1988; Reese, 1995), however, argue that most of the LID models are not suitable for practical use and suggest combining interdependent items into one super-item. These super-items would then be treated as polytomous items for further measurement.
Despite the abundance of LID studies, almost all of them are limited to the educational field examining standardized achievement tests. In the health sciences where IRT has been experiencing greater popularity, few LID studies to date have been conducted.

One major reason for overlooking the examination of LID problems is probably due to the fact that the dimensionality assumption and the local item independence assumption are closely related and to some degree, satisfaction of one assumption implies the satisfaction of the other.

It is true that checking the sustainability of the dimensionality assumption may inform the absence of local item dependence. However, it should be noted that any dimensionality test is a probability test and thus no guarantee can be made as to whether the conclusion is absolutely right or wrong. Moreover, among all the available dimensionality checking methods, no agreement has been made as to which is the best (Hattie, 1985). In other words, the conclusion of any dimensionality test is always a matter of degree rather than a yes/no statement. Since no instrument could strictly satisfy the dimensionality assumption as required in the IRT method, what we are testing is actually the degree to which the violation is small enough to be insensitive in the parameter estimation process. Although certain items may be interdependent in the “true”
situation, the small magnitude of such dependence may allow them to pass the
dimensionality test. The noise generated from these tolerable violations, though not
detected by the dimensionality test, will ultimately accumulate in the model-fit statistics.
The larger the deviation, the poorer the model fits the data. Therefore, even when the
dimensionality assumption is essentially met, locating items showing LID and finding
ways to handle the problem will have the incremental benefits of improving the model fit
and parameter estimation.

Since Item Response Theory can be regarded as a generic measurement theory,
findings from the LID studies conducted in education provide suggestions for the health
care field. However, the measurement of health outcomes has its own unique features and
challenges, which requires additional research.

First, IRT models might be different due to the difference in the underlying traits
being measured. Standardized tests measure cognitive skills which are made manifest by
the correct responses given to items. Health surveys administered to patients intend to
measure general physical functioning or disease-specific status, which are made manifest
by their capabilities of, or limitation in, accomplishing the tasks specified in each item.
The majority of items in standardized tests are in a multiple-choice format and are scored
dichotomously. For this type of item, the three-parameter IRT model, which takes into
account the pseudo-guessing influence, is a reasonable choice. In health surveys,
pseudo-guessing is no longer considered. The one- or two-parameter IRT models are more appropriate. Previous LID studies in education mainly deal with the three-parameter IRT model, while in health care, findings from one- or two-parameter models are more relevant.

The second major difference between the two fields is that the causes and types of local dependence might be different. Hoskens and De Boeck (1997) have distinguished combination dependence from order dependence. Combination dependence refers to the situation in which items are attached to the same stimuli, such as items attached to the same reading passage in a reading comprehension exam. Order dependence means that responses to early items affect the responses to subsequent items. Order dependency can occur between adjacent items or items that are located farther apart in the order. The distance between (directly) dependent items can be considered as the dependency lag. An example of order dependency is item-chaining in a math test, in which one task is divided into several steps. The success of later steps depends on the success of previous steps.

In standardized tests, combination dependence is the most often observed cause of all types and, thus, almost all previous studies examined this type of dependence. In health surveys, however, items have order dependence. For example, in the MOS SF-36\(^1\) functional scale (Ware & Sherbourne, 1992), one item asks “can you climb one flight of

---

\(^1\) Medical Outcomes Study, 36-item Short-Form Health Survey
stairs” and another asks “can you climb several flights of stairs?” If individuals are able to climb several flights of stairs, they must be able to climb one flight of stairs.

There are other situations in which order dependence is not as obvious as the above example, but contents of the items do indicate a hierarchical order of item difficulties among those items. In the Mobility domain of the Pediatric Evaluation of Disability Inventory (PEDI) developed by Haley, Coster, Ludlow, Haltiwanger, and Andrellos (1992), four dichotomous items relate to a child’s bed mobility or transfer level: 1) can the child rise to sitting position in bed or crib; 2) can the child come to sit at edge of bed and lie down from sitting at edge of bed; 3) can the child get in and out of own bed; and 4) can the child get in and out of own bed, not needing own arms. These four items are arranged hierarchically, ordered by increasing levels of difficulty. The success of performing activities stated in later items would, therefore, indicate the success of performing activities in previous items.

As shown, items assumed to be locally dependent in health surveys have a deterministic pattern of responses. In other words, response patterns of these items follow a Guttman scale (Guttman, 1950). Essentially, if items follow a Guttman scale, the total score of these items decides the response pattern. Taking, as an example, the four items in the PEDI survey from the above paragraph, a sum score of 1 would indicate a child is able to accomplish item 1; a sum score of 2 would indicate a child is able to accomplish
item 1 and 2; a sum score of 3 would indicate a child is able to accomplish item 1, 2 and 3, and so on and so forth. The deterministic pattern is one type of order dependence. For clarification, the term *deterministic order dependence* will be used to denote this type of order dependence.

The deterministic response pattern is not rare in scales measuring physical functioning, and keeping these items is desirable for the following reasons. First, items like these may help us examine the consistency of persons’ responses. Second, they help remove ceiling effects. In standardized tests measuring cognitive skills, there are always more ways to generate a large number of items covering a wider range of difficulty levels. For example, items in a math test could be written in pure mathematical formula or integrating actual situations; and difficulty levels of items could be increased by involving more steps or abstract variables. However, in health surveys, it is difficult to create a large number of discrete items with various item stems. The physical activities which could be asked about are not as diversified as the cognitive activities in educational tests, and the way an item could be written is limited to phrases like “are you able to…”, “how limited are you in…”, or “how difficult is it to…”. Therefore, increasing item difficulty is practically accomplished by increasing the demand or challenge of tasks stated in individual items.

The deterministic response pattern is necessary and commonly presented in health
surveys but inadequately studied in past literature. Based on findings from LID research in education, which will be discussed in the next chapter, it is hypothesized that the deterministic response patterns might cause items to be locally dependent. But more research is needed to verify this hypothesis and to investigate whether this type of dependence has any undesirable impact on the psychometric properties of the instrument.

**Purposes and Research Questions**

Grounded in previous LID research, to be discussed in detail in Chapter Two, the purposes of this study are to (1) investigate whether the deterministic response structure appeared among a subset of items, in patient-reported outcome measures, exhibit local dependence; (2) examine the impact of the deterministic order dependence on test statistics and on parameter estimates; and (3) determine whether these effects differ between the one-parameter (Rasch) model and the two-parameter IRT model.

Five research questions will be addressed:

(1) Does the deterministic response pattern among a subset of items cause items to be locally dependent?

(2) Does the deterministic response pattern among a subset of items influence the dimensionality assessment of test data?

(3) Does the deterministic response pattern have any impact on test reliability?

(4) Does the deterministic response pattern have any impact on test information
function and whether the impact differs when different IRT models are fitted?

(5) Does the deterministic response pattern have any impact on person score estimates and whether the impact differs when different IRT models are fitted?

**Significance of the Study**

Checking key assumptions of IRT models is necessary for constructing any reliable and valid instrument. Local item independence is a key assumption to be met for generating sample-free item calibration and item-free sample calibration for any instrument when IRT is employed. Previous LID studies (Reese, 1995, 1999; Yen, 1984, 1993; Zenisky, Hambleton, & Sireci, 2002) have shown that LID at certain degrees will influence item parameter estimates, as well as person ability estimates. The precision of person ability estimates in the medical field is especially important, because those estimates may influence the evaluation of patients’ functioning or their rehabilitation progress, which in turn decides the subsequent treatment plan.

Moreover, computerized adaptive testing (CAT) is being used more frequently due to its ability to reduce the testing burden substantially while remaining at the same precision level (Wainer et al., 2000). When an instrument is to be developed into a CAT version, checking key assumptions is especially important in the calibration stage. This is because all the item parameter estimates from the calibration stage will be relied upon to generate CAT scores for subsequent CAT administrations. When the item parameter
estimation is biased in the first stage, precision of person ability estimates in any CAT will become problematic.

This study will be the first systematic study that investigates major causes of LID existing in health surveys. IRT is experiencing fast growth in the health care field. However, among the studies examining psychometric properties of the outcome measures, local item dependence seems to be inadequate. There are numerous studies which have examined reliability, validity and other characteristics of IRT models. But there is a lack of research in the local item dependence problem. Though researchers may have checked the dimensionality assumption, which to some degree implies that the local item independence assumption may be sustainable, no study has investigated this issue directly and systematically.

Checking LID is a crucial step in developing a reliable and valid instrument. Controlling the quality of the instrument is a key interest of both the instrument developers and its users.

Summary

Item Response Theory is a modern measurement technique which has been widely used in analyzing standardized tests in education and surveys in health care. It has several key advantages over the traditional Classical Test Theory in that IRT models are able to generate sample-independent item parameters and test-independent person ability
estimates, and they are the building block of Computerized Adaptive Testing.

In order to realize these advantages, certain strong assumptions have to be met. Local item independence is a key assumption which directly determines the correct formation of the likelihood function used for parameter estimation. Violation of this assumption may have a negative effect on test statistics and parameter estimates. In health care, where IRT has become the mainstream measurement technique, examination of the local item independence assumption has been inadequate. The purposes of this study are to verify the existence of hypothesized dependence in health surveys and explore the impact of such dependence on psychometric properties of the instrument. A complete review of previous research examining the local item dependence problem is presented in Chapter Two.
Problems of local item dependence (LID) have been studied systematically during the past two decades after Goldstein (1980) stated that “there seems to have been little systematic attempt to carry out suitable experiments or to study the consequences for estimation and inference procedures when [the local item independence assumption] is violated” (p.289). Since then, a wealth of research has been conducted to investigate various LID related problems.

This chapter reviews major LID studies conducted after 1980. The studies covered in this review only involve those examining local item dependence among dichotomously scored items and those employing parametric unidimensional IRT models. This chapter is divided into three main sections. First, a formal definition of local item independence is presented. Second, research findings are organized into five sections: (i) situations where LID is most likely to occur, (ii) types of LID, (iii) consequences of LID to parameter estimates when conventional IRT models are incorrectly applied, (iv) methods of assessing LID, and (v) approaches to managing LID. The advantages and disadvantages of each method, as well as comparisons among methods, will be discussed at length in each section. The final section of this chapter identifies gaps in the previous literature and argues for the necessity of conducting this current study.
Chapter Two: Review of Literature

Definition of Local Item Independence

Lord (1980) presented the mathematical definition of local item independence as

\[
P(U_1, U_2, \ldots, U_n | \theta) = P(U_1 | \theta)P(U_2 | \theta) \cdots P(U_n | \theta) = \prod_{i=1}^{n} P(U_i | \theta),
\]

where \( \theta \) is the complete set of abilities assumed to influence the performance of an examinee on the test, \( U_i \) is the response of a randomly chosen examinee to item \( i \) \((i=1,2,\ldots,n)\), and \( P(U_i | \theta) \) denotes the probability of the response of a randomly chosen examinee with ability \( \theta \).

The property of local independence explains that for a given examinee (or all examinees at the same given ability level) the probability of a response pattern on a set of items is equal to the product of probabilities associated with the examinee’s responses to the individual items. Since local item independence refers to item independence after taking into account, or conditional on, the ability level, it is also called the assumption of conditional independence.

McDonald (1981) refers to Lord’s (1980) definition as the strong principle of local independence, and presented the weak principle of local independence as

\[
\rho(U_j, U_k | \theta) = 0, \quad (j \neq k),
\]

where \( \rho \) represents correlation. The weak principle of local independence indicates that the correlation of examinees’ responses to any two different items is zero, after taking
into account their ability level. In other words, the residual correlation of examinees’ responses on any two items is zero when the local item independence assumption holds. The weak principle of local item independence is also referred to as an operational definition for local item independence because it is the one that is often used to examine the existence of local dependence. Local item dependence (LID) occurs when neither of these equations can be established.

**Causes of Local Item Dependence**

Yen (1993) has comprehensively described some contextual conditions in which LID is most likely to occur. Similar to Lord (1980) and Hambleton, Swaminathan, and Rogers (1991), she stated that “the basic principle involved in producing LID is that there is an extra dimension (factor) introduced that consistently affects the performance of some students on some items to a greater extent than others” (p.188). Wainer and Thissen (1996) called this extra dimension a *random dimension* to indicate that it is a noise dimension not intended to be measured by the test, as opposed to the *fixed dimension* which is intended to be measured and estimated using multidimensional IRT models. Gibbons and Hedeker (1992) used *secondary factor* to refer to the extra dimension and *primary factor* to refer to the main dimension intended to be measured by the test.

Yen (1993) noted that when the effect of this unintended extra factor was constant, meaning that it influenced all students and/or all items the same way, no LID was
produced. When a constant effect occurs, a unidimensional IRT model can still be applicable, except that the estimated underlying ability for the examinees is the geometric sum of both the primary and additional dimensions. Nevertheless, when the effect of the extra factor works differently for certain students and/or for certain items, LID will be present among those items. Many contextual circumstances could lead to the occurrence of LID. Yen (1993) summarized the following situations in which LID is most likely to be present.

*External assistance, interference, or practice.* When a student receives extra assistance or practice on some test items, they may perform better than expected on these items than other students at the same proficiency level receiving no extra assistance. The assistance or practice effect is the additional factor.

*Speededness.* In a speeded test, when some students do not reach the last few items in the test and these items are scored as incorrect answers, or when examinees randomly guess the answers for un-reached items, their ability estimates may be underestimated. Local dependence appears for the unreached items.

*Fatigue.* Fatigue of examinees at the end of a test may make them feel that the last few items are more difficult than they would be if they appeared at the beginning of the test. The extra factor of fatigue will influence their responses to the last few items, as well as the estimates on their abilities. Similar to LID caused by speededness, local
dependence will appear among the last few items.

*Item or response format.* Responses to items can vary, involving either selected or constructed responses. Constructed responses can vary in terms of length or type, as when a student can respond by writing a story, drawing a picture, or building a model. These variations in response can all produce LID.

*Passage dependence.* Items attached to the same stimuli, (for example, items attached to the same passage in a reading comprehension test), may show inter-dependence. This is due to the fact that some students may have more knowledge or are more interested in certain topics stated in a passage and thus have a higher probability of answering those items correctly.

*Item chaining or explanation of previous answer.* Item chaining means that the answer of a later item depends on a correct answer from the previous item. Examples of this type of items include explanation to a previous problem or a series of multi-step mathematical questions.

*Scoring rubrics or raters.* Essay items or items in a performance assessment can be scored by a variety of rubrics or raters. Therefore, items scored by the same rater or using the same rubrics tend to show certain dependence.

In addition to the theoretical summary of LID causes by Yen (1993), empirical studies have also been conducted to confirm the existence of LID in various situations.
Of all the causes stated by Yen, the *passage dependence* is the most often and adequately studied cause in the context of large scale assessments. Lee (2004), Sireci, Thissen, and Wainer (1991), Thissen, Steinberg, and Mooney (1989), and Yen (1993) have examined whether items attached to the same reading passage in reading comprehension exams or language arts assessments showed interdependence. Also, Keller, Swaminathan, and Sireci (2003) investigated whether items based on a common auditing scenario in the Uniform Certified Professional Accountants (CPA) Exam showed local dependence. Finally, Zenisky, Hambleton, and Sireci (2002) studied whether items attached to the same stimuli in the test sections of Verbal Reasoning, Biological Sciences, and Physical Sciences in the Medical College Admission Test (MCAT) showed local dependence. The existence of LID caused by *passage dependence* was found in all these studies. Major consequences of LID, including an overestimation of internal reliability, item slopes and item information, will be discussed later in this Chapter.

In addition to *passage dependence*, some researchers have also examined other situations in which LID may occur, including speededness, item-type and multi-step tasks. Thompson and Pommerich (1996) investigated LID caused by *speededness*. They confirmed that LID was present among the last 20 items in a national standardized achievement assessment.

Lee (2004) examined not only whether items attached to the same reading passage
showed local dependence, but also whether the difference in item types caused LID in a reading comprehension test. Lee classified all items into four types based on the nature of the question asked in the test – main topic inference, recall of details, inference of details, and prediction of content. No item-type-related LID was confirmed in this study.

Ferrara, Huynh, and Michaels (1999) studied LID caused by multi-step items - items reflecting sequential steps in solving one task. They found that items tend to elicit responses that are locally dependent if examinees are required to answer, explain, defend, or use information generated from previous items.

**Types of Local Item Dependence**

In order to better understand the nature of LID and properly manage LID in test data, researchers have classified or modeled LID in several different ways.

**Positive versus Negative LID**

Yen (1984) distinguished positive LID from negative LID. Positive LID refers to a positive relationship among items after taking into account the primary factor. Since LID is present when an extra factor is introduced, positive LID indicates a positive correlation among the locally dependent items, as well as between the observed item responses and the underlying secondary factor. Similarly, negative LID indicates a negative relationship among items when taking the additional factor into consideration. It is rare to see negative LID in the context of educational assessment. Up until now, only
positive LID has been detected in empirical studies.

An extreme example for LID arises when a perfect association happens between two items after taking into consideration the primary factor. Chen and Thissen (1997) pointed out that, theoretically, this happens when the test has identical items, and “a pair of items is said to be perfectly locally dependent if every test taker responds identically to both items” (p.266).

**Order versus Combination Dependence**

Hoskens and De Boeck (1997) distinguished two general types of LID and labeled them as order dependence and combination dependence. Order dependence refers to a response to early items that affects the response to subsequent items. Order dependence can occur between adjacent items or items that are located farther apart. The distance between dependent items can be considered the lag of dependence. Examples of order dependence are an item-chaining effect in a mathematics test, or a series of items involving multi-step solutions to one task. The success of later items depends on the success of previous items.

Combination dependence refers to the interdependence among items related to the same stimuli. Examples of combination dependence include LID caused by passage dependence, item response format, fatigue and scoring rubrics/rater effect.

**Underlying Local Dependence (ULD) versus Surface Local Dependence (SLD)**
In examining the sensitivity of several pair-wise LID assessment measures, Chen and Thissen (1997) simulated LID through two types of local dependence models, the Underlying Local Dependence (ULD) and the Surface Local Dependence (SLD).

The ULD model assumes that “there is a separate trait that is common to each set of locally dependent items but is not common to the rest of the items in the test” (p.271). They used a hypothetical test composed of 4 items to structure this model:

<table>
<thead>
<tr>
<th>Factor loadings</th>
<th>$\theta_1$</th>
<th>$\theta_2$</th>
<th>$\theta_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>$w_{t11}$</td>
<td>$w_{t12}$</td>
<td>0</td>
</tr>
<tr>
<td>Item 2</td>
<td>$w_{t21}$</td>
<td>$w_{t22}$</td>
<td>0</td>
</tr>
<tr>
<td>Item 3</td>
<td>0</td>
<td>0</td>
<td>$w_{t33}$</td>
</tr>
<tr>
<td>Item 4</td>
<td>0</td>
<td>0</td>
<td>$w_{t43}$</td>
</tr>
</tbody>
</table>

In this model, all four items load on the primary factor ($\theta_1$). Item 1 and Item 2 load on one secondary factor ($\theta_2$), and Item 3 and Item 4 load on the other secondary factor ($\theta_3$). Therefore, Items 1 and 2 are locally dependent, as are Items 3 and 4. This model mimics the scenario in which LID is caused by combination dependence - items attached to the same stimuli, as in a reading comprehension test with a set of items following each passage.

The Surface Local Dependence (SLD) may “arise on a long test where some of the items at the end are omitted, or in a test with similar items” (Chen & Thissen, 1997, p. 272). The idea is that a pair of items is so similar that examinees tend to give identical answers to both items. Giving a correct response to a dependent item is decided jointly by
two components – the probability that a test taker gives exactly the same answer to the similar item and the probability determined by a certain IRT model fitted to the data.

Let $\pi_{LD}$ denote the probability that the test taker will respond to the second item in the same way as to the first item without regard to $\theta$. The SLD model is illustrated as follows.

With probability of $\pi_{LD}$:

$$\text{Response to Item 2} = \begin{cases} 1, & \text{if } X_1 = 1 \\ 0, & \text{if } X_2 = 0. \end{cases}$$

With probability $1 - \pi_{LD}$:

$$\text{Response to Item 2} = \begin{cases} 1, & \text{with } P(X_2 = 1|\theta) \\ 0, & \text{with } P(X_2 = 0|\theta). \end{cases}$$

In the above formula, $P(X_2=1|\theta)$ and $P(X_2=0|\theta)$ are obtained from the item response function of a certain IRT model fitted to the data. This model indicates that with a probability of $\pi_{LD}$, an examinee will give identical responses to both items; and with a probability of $(1-\pi_{LD})$, the response follows an IRT model.

The ULD model has become a theoretical model frequently used by recent studies in exploring ways to detect the existence of LID, as well as to develop new models in handling LID (Gibbons & Hedeker, 1992; Gibbons et al., 2007; Wainer, Bradlow, & Wang, 2007). Whereas the SLD model does not seem to have gained currency since Chen and Thissen’s study.
Methods of Assessing Local Item Dependence

A number of methods have been used to detect potential LID present in the test data. Some research suggests checking the dimensionality structure of the test data through factor analysis or principal component analysis at the item level (Fraser, 1988; Gibbons & Hedeker, 1992; McLeod, Swygert, & Thissen, 2001; Smith & Smith, 2006; Thissen, Steinberg, & Mooney, 1989). Other research (Holland & Rosenbaum, 1986; Nandakumar & Stout, 1993; Stout, 1987, 1990) suggests testing inter-item covariance. Still others propose using conditional inter-item correlations (Ferrara, Huynh, & Baghi, 1997; Ferrara, Huynh, & Michaels, 1999; Yen, 1984, 1993); and some have developed specific pair-wise dependence measures in the form of a fit statistic (Chen & Thissen, 1997). Descriptions of each method and comparisons among some measures are summarized as follows.

Dimensionality Assessment

Since local item independence and dimensionality are two closely related assumptions (Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980; Yen, 1993), satisfaction of one may imply satisfaction of the other. In other words, local item dependence is assumed to be absent when we can prove that unidimensionality holds, or the number of dimensions in multidimensional IRT models is correctly specified.

Methods to assess the dimensionality structure of item-level data in the current
literature include traditional exploratory factor analysis or principal component analysis, confirmatory factor analysis (Haley, Raczek, Coster, Dumas, & Fragala-Pinkham, 2005; Hays, Liu, Spritzer, & Cella, 2007; Hill et al., 2007), and multidimensional approaches (Gibbons & Hedeker, 1992; Gibbons et al., 2007).

**Exploratory Factor Analysis**

Exploratory factor analysis (EFA), or principal component analysis (PCA), conducted on raw scores or residual scores is often used to detect underlying latent traits or factor structure. Similar to traditional exploratory factor analysis, item factor analysis - so called because it is the traditional linear factor analysis (FA) conducted on item level test data (McLeod, Swygert, & Thissen, 2001), may be used to explore the interrelationship among individual items. Nevertheless, McLeod, Swygert, and Thissen (2001) pointed out several caveats when employing exploratory factor analysis to item level data.

The default setting for conducting exploratory factor analysis in most statistical packages begins with the Pearson product-moment correlations, or the covariance, among variables. One assumption of the Pearson correlation is that all variables are measured continuously. This assumption cannot be met when test items are scored dichotomously. McDonald and Ahlawat observed that the Pearson product moment coefficients tend to decrease as the items analyzed become less similar in difficulty. Therefore, traditional
factor analysis on the matrix of Pearson coefficients may produce spurious difficulty factors for the items (as cited in McLeod et al., 2001). The Smith and Smith (2006) study also observed an extra difficulty factor when applying the general principal component analysis in SPSS using the Pearson product-moment correlation as the input matrix.

To solve this problem, some researchers suggest using the tetrachoric correlation in place of the Pearson correlation matrix. The tetrachoric correlation is used on variables whose underlying distribution is normal but whose observed values are dichotomized. But there are some computational problems one might encounter when using the tetrachoric correlation matrix (McLeod et al., 2001):

Tetrachoric matrices for item-level data are often not positive definite, a property needed for many modern factor analysis procedures; in addition, when correlation values approach plus and minus one, estimates of the tetrachoric correlation become difficult to compute. (p. 197)

However, Brown and Benedetti (as cited in Budescu, Cohen, & Ben-Simon, 1997) indicate that these computational problems “are rare if the sample size is large and the magnitude of the correlations is moderate; if it does occur, the zero can be replaced by a small positive value” (p.233).

Although using the tetrachoric correlation does not require satisfaction of normal distribution of the variables, it still does not solve the problem of generating extra factors.
Hambleton and Rovinelli (as cited in McLeod et al., 2001) indicate that linear factor analyses using tetrachoric correlations have been found to indicate more factors than are actually present in the data.

To solve this problem, various versions of parallel analysis (Budescu et al., 1997; Drasgow & Lissak, 1983; Horn, 1965) can be used. Parallel analysis is conducted such that the eigenvalues in the real dataset are compared to the eigenvalues from a simulated dataset of the same size. Only factors with eigenvalues exceeding the values obtained from the corresponding random dataset are retained. Parallel analysis is not only applicable for factor analysis, but also for principal component analysis. Zwick and Velicer (1986) state that this procedure, which works well for both PC and FA, has been identified as one of the most accurate approaches to estimating the number of components.

**Categorical Confirmatory Factor Analysis**

Categorical confirmatory factor analysis is “a factor analytic approach that accounts for the non-normality of categorical data that renders traditional confirmatory factor analysis methods inappropriate” (Hill et al., 2007, p.40). Residual correlations, modification index, and various fit statistics are used to evaluate model fit. Good fit of a model is indicated by a non-significant $\chi^2$, the non-normed fit index (NNFI) greater than 0.92, the root mean square residual (RMSR) near zero, the root mean square error of
approximation (RMSEA) less than 0.05, and small magnitude of the modification indexes (MI) for the error covariance (Hill et al., 2007). Commercial programs such as PRELIS/LISREL and Mplus are available for model estimation. In the health care field, categorical confirmatory factor analysis has been widely used to examine the dimensionality structure of datasets prior to applying any IRT models (Haley et al., 2005; Hays et al., 2007; Hill et al., 2007).

To evaluate potential locally dependent items using categorical confirmatory factor analysis, researchers have examined the residual correlations among items in the one-factor model (Haley et al., 2005; Hays et al., 2007). “High residual correlations (greater than 0.2) will be flagged and considered as possible LD [local dependence]” (Reeve et al., 2007). Haley et al. (2005) flagged item pairs as exhibiting LID with the magnitude of residual correlations larger than ±0.2 in the Mobility domain of the Pediatric Evaluation of Disability Inventory; Hays et al. (2007) identified two item pairs with residual correlations larger than 0.14 and 0.2 respectively, in a 15-item physical functioning domain of the health-related quality of life (HRQOL) instrument.

Other than the residual correlations, the modification indexes (MI) can also serve as a statistic to detect local dependence (Reeve et al., 2007). Large MIs between items on the same domain might be an indication of local dependence. The size of MIs “should be considered in regard to the other modification indexes…and in regard to the magnitude of
the $\chi^2$ statistics” (Hill et al., 2007, p.41) – that is, whether the modification index is abnormally large compared with others in this model, or will the model fit substantially improve if this parameter was allowed to vary. Unfortunately, no specific standard is available to judge how large a MI should be in comparison with other MIs to flag an item pair for LID.

Categorical confirmative factor analysis remains an area of active development. The weight least square (WLS) estimation method is currently the statistically optimal method because it provides an appropriate statistical estimate with categorical data. But WLS suffers a numerical problem when sample size is small and the number of variables is large (Hill et al., 2007). Alternative approaches include diagonally weighted least square (DWLS) and unweighted least squares (ULS). The ULS method is conservative and less desirable (Hill et al., 2007).

**Full-Information Item Factor Analysis**

*Full-information item factor analysis*, also called the unrestricted full-information item factor analysis, was first proposed by Bock and Gibbons (1988). Strictly speaking, this method is not a factor analysis because it does not use inter-item correlation coefficients. Rather, it uses item response data directly and resembles the multidimensional IRT models for dichotomous items (McLeod et al., 2001). Thus, full-information item factor analysis not only estimates factor loadings but also intercept
values which are analogous to the pseudo-guessing item parameter in a 3-parameter IRT model. The number of dimensions is determined in advance, and fit statistics are examined and compared among models with different numbers of specified dimensions. The model with the best fit is retained. Full-information item factor analysis avoids spurious difficulty factors and other problems associated with factor analysis of correlation coefficients (McLeod et al., 2001).

**Full-Information bi-factor Item Factor Analysis**

A major drawback of the full-information item factor analysis is the substantial computational burden when the number of underlying dimensions is large. In order to solve this problem, and to reflect the fundamental causes of LID, Gibbons and Hedeker (1992) proposed using the full-information bi-factor item factor analysis - abbreviated as the bi-factor model in later texts - to model LID in dichotomous data. Gibbons et al. (2007) further extended the method to include situations in which items are scored polytomously. The bi-factor analysis constrains each item to load only on the primary dimension and one secondary dimension. For four items, the bi-factor pattern matrix might be:

\[
\alpha = \begin{bmatrix}
\alpha_{11} & \alpha_{12} & 0 \\
\alpha_{21} & \alpha_{22} & 0 \\
0 & \alpha_{33} \\
0 & \alpha_{43}
\end{bmatrix}
\]
In the matrix, rows correspond to items and columns to factors. Column 1 is the primary factor (F1) and columns 2 and 3 are two secondary factors (F2 and F3). Each item has a non-zero loading on the primary factor (F1), and a non-zero loading on one of secondary factors. Both Items 1 and 2 load on the F2; and both Items 3 and 4 load on F3. Thus, local dependence exists between Items 1 and 2, as well as between Items 3 and 4. Note that this structure is exactly the same as the Underlying Local Dependence (ULD) model in the Chen and Thissen (1997) study, as discussed earlier in this chapter. Chen and Thissen laid out this theoretic framework to model the fundamental cause of LID, while Gibbons applied this model to accommodate LID in dataset.

The bi-factor model provides a more parsimonious factor solution than the unrestricted full-information item factor analysis. It permits analysis of models with larger numbers of domains and allows conditional dependence among subsets of items. The computer program TESTFACT (du Toit, 2003) is available to conduct the full-information item factor analysis and the bi-factor analysis.

NOHARM Residual Analysis

NOHARM is a software package used to fit unidimensional and multidimensional IRT models (Fraser, 1988). NOHARM residual analysis identifies LID by analyzing residual item covariance patterns after fitting a multidimensional IRT with a predetermined number of dimensions. When certain items are locally dependent, the
average residual item covariance of these items is expected to be larger than the average residual item covariance of locally independent items. Using the NOHARM residual analysis, Thompson and Pommerich (1996) detected LID caused by speededness among the last 20 items of their data set. However, since the sampling distribution of the test statistics is unknown, it is unable to conduct hypothesis testing.

**Inter-item Covariance Analysis**

Holland and Rosenbaum (1986) have proven that any pair of item covariance is non-negative, given independence conditional on a latent attribute and a monotonic increasing function of the remaining item responses. Therefore, conditional independence can be inferred by testing for non-negative covariance between any two items. They used a contingency coefficient such as the Mantel-Haenszel odds-ratios as test statistics. A separate statistical test is made for each item comparison. Although theoretically appealing, their procedure is laborious (Meara, Robin, & Sireci, 2000), i.e., given $n$ items, there are a total of $[n(n-1)]/2$ comparisons to be made.

**Conditional Inter-item Correlation Analysis**

As stated previously, the operational definition for local item independence is:

$$\rho(U_i, U_j | \theta) = 0, \text{ where } i \neq j,$$

indicating that conditional on the ability ($\theta$) level, or after controlling the underlying trait level, no correlation will be observed between any pair of items. Based on this definition,
two methods are proposed to detect LID by examining pair-wise item correlations after conditioning on the $\theta$ level. Yen (1984, 1993) proposed using the $Q_3$ statistic – inter-item correlations of residual scores, and Ferrara, Huynh, and Baghi (1997) proposed using inter-item correlations of raw scores.

**Yen’s $Q_3$ statistic**

The $Q_3$ statistics proposed by Yen (1984, 1993) are inter-item correlations after taking into account the latent trait. To obtain $Q_3$ between items $i$ and $j$ ($Q_{3ij}$), a proper IRT model is first fitted to the test data. Item parameter and person proficiency estimates are computed. Then the point estimates of examinees’ proficiency levels, together with all the item parameter estimates, are used in calculating the probability of giving a correct response by each examinee on items $i$ and $j$. Next, the residual scores between each examinee’s observed response and the computed probability of giving a correct response on items $i$ and $j$ ($d_i$ and $d_j$) are obtained. $Q_{3ij}$ is the Pearson correlation of the residual scores between items $i$ and $j$, $Q_{3ij} = \text{Corr} (d_i, d_j)$, across all examinees. For a test composed of $n$ items, there are $n^*(n-1)/2$ possible pairs of items, and thus $n^*(n-1)/2$ unique $Q_3$ values. For an entire test, $Q_3$ can be represented by the lower triangle of the residual correlation matrix and each entry ($Q_{3ij}$) in the matrix denotes the specific residual correlation value between items $i$ and $j$.

Yen (1993) states that “if item scores are locally dependent, then, in factor
analysis terminology, they will have non zero residual correlations after removal of the first factor” (p.188). Similarly, when items are locally independent, residual correlations, or $Q_{3ij}$, should be zero for any two items. In reality, though, even when items are locally independent, the expected value of $Q_{3ij}$ is not exactly zero, but slightly negative. This is because the observed scores are used explicitly for calculating both the expected scores and residual scores, which is known as the part-whole contamination (Yen, 1984). Yen (1993) has shown that the expected value of $Q_3$, when the local independence assumption holds, is approximately $-1/(n-1)$, where $n$ is the total number of items in a test.

For diagnostic purposes, item pairs showing potential LID can be flagged by comparing the observed $Q_3$ with the expected $Q_3$. Large differences indicate the possible existence of LID. Alternatively, Fitzgeralnds (as cited in Chen & Thissen, 1997) proposed that when using $Q_3$ to screen items for local dependence, an absolute value of 0.2 can be used as the uniform critical value. However, Chen and Thissen note that using this criterion in a test with a realistic length of 40-80 items will underestimate the type I error rates substantially, and result in very low power for $Q_3$ in comparison to $G^2$ and $X^2$ indexes, two LID measures discussed later in this chapter.

For the purpose of hypothesis testing, Yen (1984) and Shen (1997) suggest using the Fisher’s $z$ transformation of $Q_3$ which is approximately normally distributed with a mean of zero and a variance of $1/(n-3)$. However, after examining the null distribution of
When local item independence holds, Chen and Thissen (1997) point out that it is not appropriate to use the Fisher’s z transformation for hypothesis testing. It should be noted that the $Q_3$ used in the Chen and Thissen study is not exactly the same as the $Q_3$ in Yen’s (1984, 1993) study. In Yen’s study, the ability ($\theta$) estimate is the point estimate using the mode of the likelihood function without specifying any population distribution of $\theta$. The $Q_3$ used in Chen and Thissen’s study is the mode of the marginal maximum likelihood function assuming the population distribution of $\theta$ is normal with a mean of zero and a standard deviation of one.

Chen and Thissen state that using the Fisher’s z transformation with a variance of $1/(n-3)$ would be correct, if the data being correlated arose from a bivariate Gaussian distribution. However, their simulation study shows that the residual scores used in calculating $Q_3$ are not bivariate Gaussian and the sampling distribution of the standardized Fisher’s z transformation of $Q_3$ is “certainly bell shaped, but its approximation by the standard normal is not particularly good, especially in the (crucial) tails” (p.284).

Since the sampling distribution of $Q_3$ is unknown and the normal approximation of the Fisher’s z is not proper, it is impossible to conduct hypothesis testing using the traditional parametric method. In order to solve this problem, Habing (2001) has proposed that a parametric bootstrap sampling distribution could be used for $Q_3$ and other
statistics having no proper sampling distributions. Habing suggested three steps in this method: 1) fit an appropriate IRT model to the observed data and estimate item parameters and proficiency scores; 2) use the specified model and parameter estimates obtained from the first step to generate a fixed number (B) of simulated sample datasets and calculate the sample statistic (i.e., average $Q_3$ for the entire test) for each simulated dataset; and 3) construct the Bootstrap sampling distribution of the $Q_3$ statistics from the B simulated datasets. Then compare the observed $Q_3$ against the Bootstrap sampling distribution of $Q_3$ and compute the percentage of $Q_3$’s as extreme as or more extreme than the observed statistic. This is the estimated $p$-value for the hypothesis test.

Although Habing’s method makes it possible to conduct hypothesis testing on LID measured by $Q_3$ and other indexes, Chen and Thissen (1997) suggest that assessment of LID should not be determined exclusively by hypothesis testing. Rather, these pair-wise LID indexes should be used more for diagnostics purposes because “any meaningful interpretation of these measures requires skills and experience in IRT analysis and close examination of the item content” (p.288). Up until now, $Q_3$ has been successfully used in many empirical studies for diagnostic purposes (Keller, Swaminathan, & Sireci, 2003; Lee, 2004; Yen, 1993; Zenisky, Hambleton, & Sireci, 2002).
Chapter Two: Review of Literature

Inter-item Raw Score Correlation

Unlike Yen (1984), who used the inter-item residual correlations to detect LID, Ferrara, Huynh, and Baghi (1997), and Ferrara, Huynh, and Michaels (1999) used the inter-item raw score correlations to detect local dependence. To implement their method, examinees are sorted into several homogeneous score groups based on their raw scores and items hypothesized to be inter-dependent are grouped into clusters. For each score group, inter-item correlation matrices for items within the same cluster (within-cluster), as well as between different clusters (between-cluster), are constructed. It is assumed that within-cluster items are locally dependent, whereas between-cluster items are not. Therefore, the between-cluster correlations are used to mimic a situation in which LID is not expected. The frequency distribution of all between-cluster correlations is plotted and a cut point is determined. Next, the within-cluster correlations are averaged across all the score groups and compared to this cut point. In Ferrara, Huynh, and Baghi’s (1997) study, a cut point of 0.1 is used because all the between-cluster correlations are below this value. In Ferrara, Huynh, and Michaels’s (1999), a cut point of 0.03 is used since this value represents an empirical $p$-value of 0.05. Item clusters with average correlation values larger than the cut point are identified as exhibiting possible local dependence.

Both Yen’s $Q_3$ statistic and the inter-item raw score correlation method use the conditional inter-item correlation to detect local dependence. The difference between
these two methods lies in how the ability level is “conditioned”, or controlled. For Yen’s $Q_3$ statistic, the ability level is conditioned by partialing out the deterministic part of the responses that can be predicted by the specified IRT model; whereas for the inter-item raw score correlation method, the ability level is conditioned by grouping examinees having similar total raw scores. The raw score correlation method does not impose any IRT models on the data and thus avoids undesirable consequences (which could be caused by mis-specifying an IRT model), but grouping examinees is still arbitrary.

Ferrara, Huynh, and Michael (1999) have pointed out that the inter-item raw score correlation method is “similar in theoretical underpinnings to Yen’s $Q_3$ statistics” and they “have [been] shown to produce highly similar results” (p.127).

**LID Measures in the Form of Fit Statistics**

Deviation from important assumptions of IRT will result in a poor fit between the proposed model and the observed data. Therefore, fit indexes could also be used to examine the existence of LID. However, there are many factors that can affect the fit of a model (Yen, 1984), e.g., misspecification of models or violation of one or more assumptions. Certain fit indices are only sensitive to certain circumstances. Yen’s $Q_3$ statistic is actually developed from a series of $Q$ fit statistics ($Q_1$, $Q_2$ and signed $Q_2$) and it has been found to be sensitive to violation of the dimensionality assumption (Yen, 1984).

The $G^2$ and $X^2$ LID measures to be discussed in this section are also forms of a fit
statistic. They were initially developed to test for association within $2 \times 2$ tables, and are used for examining differences between observed frequencies and expected frequencies. Chen and Thissen (1997) studied the distribution and sensitivity of $G^2$ and $X^2$ in detecting local dependence.

In order to calculate $G^2$ and $X^2$, the observed frequencies and expected frequencies of examinees’ response patterns to items $i$ and $j$ are first constructed as follows:

<table>
<thead>
<tr>
<th>Observed Frequency</th>
<th>Expected Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item $j$</td>
<td>Item $j$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$O_{11}$</td>
<td>$E_{11}$</td>
</tr>
<tr>
<td>$O_{12}$</td>
<td>$E_{12}$</td>
</tr>
<tr>
<td>$O_{21}$</td>
<td>$E_{21}$</td>
</tr>
<tr>
<td>$O_{22}$</td>
<td>$E_{22}$</td>
</tr>
</tbody>
</table>

In the “Observed frequency” table, “0” stands for an incorrect answer and “1” for a correct answer. $O_{11}$ is the observed number of examinees who answer both items incorrectly; $O_{12}$ is the observed number of examinees who answer item $i$ incorrectly but item $j$ correctly; and the same naming convention applies for $O_{21}$ and $O_{22}$. The table “Expected Frequency” represents similar information as the “Observed Frequency” table, except that it now shows the number of examinees who are expected to give certain response patterns. The expected frequencies come from the IRT model using marginal maximum likelihood (MML) estimation, assuming the underlying population distribution
of the ability level is normal with mean of 0 and standard deviation of 1.

If the IRT model fits the dataset well enough, the observed frequency should be very similar to the expected frequency allowing random fluctuation. However, if the two tables differ more than expected, “that is [an] evidence that the IRT model induces more or less dependence than is observed” (Chen & Thissen, 1997, p.268). Using the notation in the above frequency table, $G^2$ and $X^2$ are defined as

$$G^2 = -2 \sum_{i=1}^{2} \sum_{j=1}^{2} O_{ij} \ln \left( \frac{E_{ij}}{O_{ij}} \right),$$

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})}{E_{ij}}.$$  

The purpose of Chen and Thissen’s study was to examine the null distributions of these two LID measures when local item independence holds, as well as the power of detecting dependent items when LID is present. They simulated datasets reflecting two types of local item dependence models – the Underlying Local Dependence (ULD) and the Surface Local Dependence (SLD). The ULD model mimics the situation in which local dependence is most likely to be caused by items related to the same stimuli; and the SLD model simulates situations in which LID is most likely to be caused by not-reaching items at the end of a speeded test. Both 2PL and 3PL IRT models were used to analyze the data. Chen and Thissen compared the sampling distributions under the null condition of local independence and the power of detecting LID items between $X^2$ and $G^2$. 

Chapter Two: Review of Literature

Null Distributions of $X^2$ and $G^2$

Under the null condition of local item independence, there is almost no difference between $X^2$ and $G^2$ in the 2PL model. However, the mean and standard deviation of the $X^2$ distribution are slightly smaller than that of the $G^2$ distribution in the 3PL model.

Both $X^2$ and $G^2$ LID indexes appear to be distributed very nearly as a $\chi^2$ distribution with degrees of freedom slightly less than one in the null condition. The critical values of 3.84 and 6.63, which correspond respectively to type I error rates of 0.05 and 0.01 from a $\chi^2$ distribution with one degree of freedom, are slightly conservative.

Simulation results from the Chen and Thissen study suggest that the true 0.05 critical values are around 3.38 and 3.84 for $X^2$ and $G^2$, respectively, and the true 0.01 critical values are about 5.65 and 6.07 for $X^2$ and $G^2$, respectively. When used for diagnostic purposes rather than hypothesis testing, “W.-H Chen has recommended that item pairs with $G^2$ values greater than 10.0 be flagged for possible local dependence” (as cited in Thompson & Pommerich, 1996, p.3).

Power of $X^2$ and $G^2$

Chen and Thissen also compared the power of $G^2$ and $X^2$ when LID is present in the dataset. In general, the power of detecting LID is always slightly higher in the 2PL model than in the 3PL model. For datasets simulated under the Underlying Local Dependence (ULD) model, the power of both $G^2$ and $X^2$ is around 0.30 - 0.40, when the
nominal alpha is set at 0.05 and when the dimensionality is obvious. When weak dimensionality exists in the dataset, the power is lower. For data simulated under the \textit{Surface Local Dependence} (SLD) model, the power for both measures is never less than 0.86 for a nominal alpha of 0.05 or 0.01. In general, \( G^2 \) is slightly more powerful than the \( X^2 \) index.

Chen and Thissen also compared the power of \( G^2 \) and \( X^2 \) with the power of \( Q_3 \). They found that under the \textit{Underlying Local Dependence} (ULD) model, \( Q_3 \) outperforms \( X^2 \) and \( G^2 \); but under the \textit{Surface Local Dependence} (SLD) model, the power of the three statistics to detect LID item pairs can be considered equal.

Of all the fit-statistic LID assessment measures discussed above, the \( Q_3 \) statistic is most often and most successfully used in studies (Keller, Swaminathan, & Sireci, 2003; Lee, 2004; Reese, 1995; Zenisky, Hambleton, & Sireci 2002). The popularity of \( Q_3 \) is probably due to its following advantages.

First, \( X^2 \) and \( G^2 \) measures follow a chi-square distribution under the null condition, and their magnitudes are sensitive to sample sizes (Thompson & Pommerich, 1996); whereas \( Q_3 \) does not have this problem.

Second, as Yen (1984) has shown, \( Q_3 \) is an IRT fit index for item pairs that is sensitive to LID-related item misfit. Various item fit statistics can be used as an overall
indicator of the degree of departure from the local independence or unidimensionality assumption. However, misfit could be caused not only by the violation of IRT assumptions, but also mis-specification of the IRT model. $Q_3$ is specifically targeted at the misfit for item pairs caused by LID.

Third, the definition and computation of $Q_3$ examines whether the correlation between two items is zero after partialing out the ability level. This definition is a direct reflection of the weak principle of local independence, or the operational definition of local item independence proposed by McDonald (1981). The computation of $Q_3$ makes it easy to understand and readily interpret in terms of LID.

Lastly, although the $Q_3$ statistic is the residual correlation between two items, the use of $Q_3$ is not limited to describing relationships only between two items. Arithmetic averages of $Q_3$ values among more than two items can be obtained to evaluate inter-dependence among multiple items. For example, items hypothesized to show inter-dependence can be grouped and calculated for the within-group average of $Q_3$, or the between-group average $Q_3$. If the group factor has caused LID, the within-group average $Q_3$ is expected to be higher than the between-group average $Q_3$. The flexibility in grouping items and computing average within- or between-group $Q_3$ easily allows for the examination of different potential causes of LID easily.
**Consequences of Local Item Dependence**

Thompson and Pommerich (1996) indicate that “identifying the existence and sources of local dependence is not sufficient to preclude the use of a unidimensional IRT model; it is necessary to directly examine the effect of local dependence on test characteristics” (p.5).

Previous studies have systematically examined the impact of LID on item parameter estimates, item characteristic curve, test reliability, test information/precision level of the estimates, and on proficiency estimates. Conceptual explanations of the consequences of LID and some simulation study results are summarized in the following section. Findings on consequences to parameter estimates when different models are used to accommodate LID will be discussed later in this chapter.

**Impact of LID on Item Parameter Estimates**

Reese (1995) conducted a series of simulation analyses to examine how LID of various degrees would influence item parameter estimates. She simulated four datasets with zero, low, medium and high LID levels based on $Q_3$ values. Each simulated test data had 40 items grouped into 4 clusters, with each cluster having 10 items. In the zero-LID data, all the items were independent of each other; in the low-LID data, the average within-cluster $Q_3$ was 0.01 for all the four clusters, and the average between-cluster $Q_3$ values were all less than 0.01; in the mid-LID data, average within-cluster $Q_3$ were 0.05,
0.02, 0.02 and 0.03 for the four clusters, respectively, and all the average between-cluster $Q_3$ were less than 0.02; and in the high-LID data, the average within-clusters $Q_3$ was 0.3 for the four clusters, and all the average between-cluster $Q_3$ were less than 0.3. The 3PL IRT model was used to analyze each dataset.

The results showed that for the discrimination parameter, or the $a$-parameter, low to mid-LID levels seem to have a minimal impact on the estimates, while high-LID tend to overestimate the $a$-parameter. Specifically, the correlation of the $a$-parameter estimates with the true parameters for the zero through medium LID level data are 0.91, 0.91 and 0.88, respectively, but the correlation drops to a very low value of 0.26 for the high-LID level data.

For the difficulty parameter, or the $b$-parameter, the estimates decrease slightly as LID level increases, but the correlations of the $b$-parameter estimates with the true parameters are high for all LID-level datasets.

For the pseudo-guessing parameter, or the $c$-parameter, the estimates also decrease as LID level increases. The correlations of the $c$-parameter estimates with the true parameters are 0.88, 0.80 and 0.80 for the zero, low and medium LID levels, and drops to 0.23 in the high-LID data.

Reese's results have suggested that high-LID level (with average within-cluster $Q_3$ larger than 0.3) has an adverse effect on the $a$- and $c$-parameter. Specifically, an
overestimation of the \( a \)-parameter and an underestimation of the \( c \)-parameter are observed in the study. The impact of LID on the \( b \)-parameter is minimal.

Tuerlinckx and De Boeck (2001) have further shown that positive LID tends to overestimate item discrimination and negative LID tends to underestimate item discrimination. Reese (1995) modeled positive LID (all within- and between-cluster \( Q_3 \) averages are positive) in each data. Therefore, it was not surprising to find an overestimation of the discrimination parameter in her study.

**Impact of LID on Item Characteristic Curve**

Item characteristic curve (ICC) is generated jointly from \( a \)-, \( b \)-, and \( c \)-parameters. Various combinations of these three parameters can produce similar or dissimilar ICCs for certain ranges on the ability scale. Studying the effect of LID on only one item parameter without considering other item parameters is less meaningful.

Reese (1995) examined the impact of LID on ICCs. She found that for zero- and low-LID levels, ICCs generated from the estimated parameters were very similar to ICCs generated from the true parameters. For the medium-LID level, a slight crossing of ICCs began to emerge, with the estimated ICCs slightly steeper than the true ICCs in the middle range of the ability, and the lower asymptote for the estimated ICCs slightly below that of the true ICCs. For the high-LID level, the ICCs diverges slightly more and in the same direction as the medium LID condition.
Impact of LID on Internal Reliability

Reliability is defined as the correlation between the true score (T) and the observed score (X). Operationally, it is expressed as the ratio of true score variance over observed score variance:

$$\rho_{TX}^2 = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2 - \sigma_e^2}{\sigma_X^2},$$

where $$\sigma_T^2$$ is the true score variance, $$\sigma_X^2$$ is the total raw score variance, and $$\sigma_e^2$$ is the error variance, ($$\sigma_e$$ is the standard error of measurement [SEM]).

Two reliability coefficients have been used in previous studies when investigating the impact of LID on reliability. They are the Cronbach alpha coefficient in Classical Test Theory and the marginal reliability coefficient in Item Response Theory. Since in Classical Test Theory, $$\sigma_e^2$$ is constant over the person ability continuum, the alpha coefficient remains unchanged for the same test regardless of the difference in proficiency estimates. The alpha coefficient is calculated as

$$\alpha = \frac{k}{k-1} \left(1 - \frac{\sum_i \sigma_i^2}{\sigma_X^2}\right),$$

where $$\sigma_i^2$$ is the variance of item $$i$$ and $$\sigma_X^2$$ is the variance of the total raw scores.

In Item Response Theory, $$\sigma_e^2$$ is a function of proficiency ($$\theta$$) estimates, which is defined as the inverse of the test information function. In order to provide a reliability
index that is comparable to the traditional reliability coefficient, Sireci, Thissen, and Wainer (1991) used the averaged error variance ($\bar{\sigma}_e^2$) in place of $\sigma_e^2$:

$$\bar{\sigma}_e^2 = \int \sigma_e^2 g(\theta) d\theta,$$

where $\sigma_e^2$ is the expected value of the error variance associated with the expected a posteriori estimate at $\theta$, and $g(\theta)$ is the population distribution of proficiency. The IRT marginal reliability is then computed as:

$$\rho_{Xt}^2 = \frac{\sigma_X^2 - \bar{\sigma}_e^2}{\sigma_X^2}.$$

Despite the difference in mathematical computations of the above two reliability coefficients, “experience shows that IRT marginal reliability is usually within about 0.02 of coefficient $\alpha$” (Wainer & Thissen, 1996, p.26).

Theoretical analyses, as well as empirical studies, reveal that presence of LID in a dataset may result in an overestimation of test reliability (Keller, Swaminathan, & Sireci, 2003; Reese, 1999; Sireci, Thissen, & Wainer, 1991; Thompson & Pommerich, 1996; Wainer & Thissen, 1996; Zenisky, Hambleton, & Sireci, 2002). In theory, when two or more items show local dependence, they may have a stronger correlation with each other, and thus a stronger correlation with the total test score. Test reliability will be overestimated by the excess correlation among these items. Therefore, the true reliability is lower than what is computed.
Reese (1999) examined how LID of various degrees would influence the alpha coefficient. Following the same simulation procedure of the study conducted by Reese (1995), which was described in the previous section, she generated four datasets with zero, low, medium and high LID levels based on $Q_3$ values. The results show that low and medium LID levels caused some overestimation of the coefficient $\alpha$, whereas a high LID level results in a strong overestimation of the reliability index. Specifically, the alpha coefficient increased from 0.91 in the zero-LID data to 0.92, 0.93, and 0.97 in the low-, mid-, and high-LID data, respectively. When the estimated coefficient alpha was compared with the true reliability (ratio of the true score variance over the observed score variance) in each dataset, the overestimation was more dramatic. The true reliability was 0.88, 0.80 and 0.41 in the low-, mid-, and high-LID level data respectively, whereas the coefficient alpha in each case was 0.92, 0.93 and 0.97. In the zero-LID data, the coefficient alpha of 0.92 was almost the same as the true reliability of 0.93.

**Impact of LID on Test Information Function**

In Item Response Theory, the test information function plays an important role in describing the precision level of examinees’ proficiency/ability estimates. It is the reciprocal of the error variance. Mathematically, test information is expressed as the sum of item information, $T(\theta) = \sum_i I_i(\theta)$, where $I_i(\theta)$ is the item information for item $i$ and defined as:
where \( P_i(\theta) \) is the probability that an examinee with a proficiency level of \( \theta \) will answer item \( i \) correctly, \( Q_i(\theta) \) is the probability that an examinee with a proficiency level of \( \theta \) will answer item \( i \) incorrectly, and \( P_i'(\theta) \) is the first derivative of \( P_i(\theta) \).

Through mathematical derivation (Baker, 2004), the item information function of dichotomous items for the 1-, 2-, and 3-parameter logistic IRT model can be written as:

1PL model: \( I_i(\theta) = P_i(\theta)Q_i(\theta) \),

2-PL model: \( I_i(\theta) = a_i^2 P_i(\theta)Q_i(\theta) \), and

3-PL model: \( I_i(\theta) = a_i^2 \left[ \frac{Q_i(\theta)}{P_i(\theta)} \right] \left[ \frac{P_i(\theta) - c_i}{1 - c_i} \right] \)

where \( a_i \) is the discrimination parameter for item \( i \) and \( c_i \) is the pseudo-guessing parameter for item \( i \).

As shown, item information is related to the \( a \)-parameter, and so is the test information. When other things are equal and \( \theta \) is not at the extreme ends of the proficiency continuum, the higher the discrimination parameter, the higher the item information. Since LID tends to overestimate the discrimination parameter, it also tends to overestimate the item and test information.

The simulation study conducted by Reese (1995) suggested that the test information computed from parameters estimated in the high-LID level test data could be
3 to 4 times higher than the test information derived from the true parameters for a certain range of the ability distribution. More discussion about inflation of test information will be discussed later in the chapter when describing different models are used to handle LID.

**Impact of LID on Ability Estimates**

Two simulation studies identified how LID with various degrees in test data can influence score estimates (Reese, 1995) and percentile ranks (Reese, 1999).

Reese (1995) compared three score-distributions for each of four simulated datasets with zero, low, mid and high-LID level. The three score distributions were (a) the true score distribution, (b) the estimated true score distribution derived using the item and ability parameters estimated from the simulated data, and (c) the observed raw score distribution. In the zero-, low- and mid-LID datasets, the three score-distributions were very similar to each other. However, dramatic differences were presented in the high-LID dataset, in which the observed and estimated true score distributions were still very similar to one another, but they lost their normality and did not resemble the true score distribution. The scores of low ability test takers were underestimated and the scores of high ability test takers were overestimated. In addition, the Spearman correlation of the score rank orders between the true scores and the estimated true scores (as well as between the true scores and the observed scores), were only 0.63 and 0.64, respectively,
in the high-LID dataset; whereas, these correlations were at least 0.90 for the datasets with zero- to mid-level LID.

Following the same simulation procedure, Reese (1999) examined whether various degrees of LID had an impact on percentile ranks of the ability estimates. She showed that the bias of the estimated percentile ranks (bias is computed as the true percentile ranks minus the estimated percentile ranks), was minimal for data with zero- to mid-LID levels. Whereas in the high-LID level test, she indicates that “the underestimation of low scores”, as detected in Reese (1995), “results in higher percentile ranks for lower scoring test takers and the overestimation of high scores results in lower percentile ranks for higher scoring test takers” (Reese, 1999, p.10). Therefore, “when actual test scores are of interest, high LID results in an unfair advantage for high scoring test takers and an unfair disadvantage for low scoring test takers. However, the opposite is true when percentile ranks are considered” (Reese, 1999, p.11).

**Approaches to Managing Local Item Dependence**

Although LID is undesirable, there are good reasons for including items that are interdependent on an assessment. Many real world tasks require solving related problems or solving a single problem in stepwise fashion. Including context-dependent items on a test may increase construct validity. Therefore, “the challenge for the test developer is not the elimination of item dependencies, but rather, how to properly model such
dependencies” (Zenisky et al., 2002, p.292).

One major approach to managing potential LID is to consider the response patterns to sets of interdependent items as a basic unit instead of responses to single items. Several terms have been used to denote such item sets - *subtests* (Andrich, 1985), *testlets* (Wainer & Kiely, 1987), *item bundles* (Rosenbaum, 1988), and *superitems* (Wilson & Adams, 1995). The term *testlet* will be used throughout this paper to indicate a set of items assumed to be locally dependent. Though given a generic name, distinctions should be made between two completely different methods in modeling the testlet, namely, the *score-based approach* and the *item-based approach* (Wilson & Adams, 1995).

The *score-based approach*, also called the *testlet-as-polytomous-item* model (DeMars, 2006), is more commonly used on dichotomously scored items. This approach accommodates LID by combining locally dependent items into one testlet, the score of which is simply the sum of all the items it comprises. Then, polytomous IRT models are employed to analyze the “newly” assembled test composed of testlets. Advantages of this method are that the concept is relatively easy to understand, and the estimation of parameters can be achieved by commercially available software like PARSCALE or MULTILOG (du Toit, 2003). However, changing from the dichotomous scoring scheme to polytomous scoring scheme overlooks differences in response patterns at the item-level, and alters the interpretation of item parameters. Moreover, combining dichotomous items
into testlets results in loss of test information even when all items are locally independent. Details about this problem are discussed in the next section.

On the other hand, the item-based approach accommodates LID by including additional interaction parameters in the conventional 1-, 2-, and 3-parameter IRT models. The explicit modeling guarantees that item-level information is retained. However, this method involves establishing more complicated IRT models and more item parameters to be estimated. New software packages or programming syntax may also be needed to estimate the parameters.

In practice, selection of certain methods in dealing with LID is decided by the cause(s) of the dependence, the availability of software and knowledge in applying new models, and the impact of the approach on item and person parameter estimates.

The Score-Based Approach

As mentioned above, combining inter-dependent dichotomous items into one testlet and applying polytomous IRT models to the newly assembled test are easy to understand in theory, as well as to implement in practice. Note that this method does not change examinees’ total raw scores, but does change examinees’ response patterns. In fact, the score-based testlet approach accommodates LID by changing the scoring scheme from dichotomous IRT models to polytomous IRT models, the process of which allows elimination of local dependence among items in the same testlet, while still assuming
local independence across testlets.

This approach to LID has been commonly used and thoroughly studied by many researchers (Keller, Swaminathan, & Sireci, 2003; Lee, 2004; Sireci, Thissen, & Wainer, 1991; Thissen, Steinberg, & Mooney, 1989; Wainer, 1995; Yen, 1993; Zenisky, Hambleton, & Sireci, 2002) in situations where LID is caused by items attached to the same stimuli. All these studies have analyzed test data in two ways: (a) treating the original dichotomous items as if they were locally independent and fitting standard dichotomous IRT models, and (b) combining locally dependent items into testlets and fitting polytomous IRT models to the “new” test.

Since most of these studies dealt with large scale assessment data, which require consideration of the slope and pseudo-guessing item parameters, the 3-parameter logistic (3PL) model was applied to analyze the dichotomously scored tests. The Generalized Partial Credit Model (GPCM) developed by Muraki (1992) and the Graded Response Model (GRM) developed by Samejima (1969) were the two commonly used models to analyze polytomous test with ordered responses. The choice of the polytomous model is not a difficult one since there is little evidence demonstrating that different polytomous IRT models produce substantially different measurement outcomes when applied to the same data (Cook & Dodd, 1999; Ostini & Nering, 2006).

Next, test reliability, test information, and ability (score) estimates were compared
between the two scoring schemes when LID was present.

*Test Reliability*

When studying the reliability of the two scoring schemes, Wainer and Thissen (1996) distinguished item reliability from testlet reliability. Item reliability provides the internal consistency estimate of a test based on individual items, while testlet reliability provides the internal consistency estimate of a test composed of testlets. Though the total raw score for each examinee remains the same regardless of the scoring method, the testlet reliability is expected to be lower than item reliability even when all the dichotomous items are independent of each other. This is due to the fact that there are fewer “items” in the polytomous test. However, when items show local dependence in the testlet, the reliability drop would be much higher than expected, which, in turn, is a sign of LID.

Zenisky, Hambleton, and Sireci (2002) have shown that when LID is absent in a test, combining items into testlets will lead to an expected reliability drop of 2 – 3% (computed as \(\frac{\text{item reliability} - \text{testlet reliability}}{\text{item reliability}}\times100\%\)); whereas for tests containing LID items, the decrease is about 5%. More severe drops in reliability estimates (over 10%) were detected in studies conducted by Wainer (1995), and Sireci, Thissen, and Wainer (1991). It should be noted that the magnitude of the reliability drop is affected by many factors, such as the degree of local dependence, the number of
dependent items relative to the number of independent items in the test, and the number of items in each testlet. When using reliability as an indicator of the existence of LID, simulated data have to be generated as a reference point for comparison.

Test Information

Test information is a concept closely related to reliability, both of which are measures related to measurement error. Theoretically, like reliability, test information is expected to decrease as the scoring scheme changes from a dichotomous model to a polytomous model, even when all items are independent of each other. This expected loss of information can be attributed to the fact that the item-level response patterns are ignored in testlet scoring. Therefore, the fact that a testlet yields less information is not necessarily evidence of LID (Keller et al., 2003). When the score-based approach is adopted to handle LID, the drop of information is partly due to the change of the scoring scheme and partly due to the existence of LID. Only when the drop of test information is considerably larger than the expected drop, can we claim a potential existence of LID.

Several empirical studies (Keller et al., 2003; Thompson & Pommerich, 1996; Yen, 1993; Zenisky et al., 2002) have also confirmed the theoretical descriptions made above. The loss of test information in real test data where LID is present is much higher than the loss of test information in a simulated data where LID is absent. In other words, test information will be overestimated if no action is taken to control LID.
It should be noted that the loss of information when forming items into testlets for independent items is true for 2- and 3-PL models, but not necessarily true for Rasch models (i.e., change the dichotomous Rasch model to the Partial Credit Model). This is because Rasch models do not estimate the discrimination parameter and use the total score as the sufficient statistic. Keller et al. (2003) have indicated that “an item that is constructed as a polytomous item for partial credit scoring will most often yield more information than the same item scored dichotomously” (p.212).

Huynh (1994) has shown that when items are locally independent, equivalence exists between a partial credit item and a set of independent Rasch binary items. Huynh (1996) further demonstrates that, in theory, each Rasch partial credit item can be expressed as the sum of independent Rasch binary items. Wilson (1988), based on the same theory as Huynh, proposed to detect LID by comparing the observed item response category characteristic curve with the generated category characteristic curve under the local independence assumption. Large differences between these curves would indicate the existence of LID. However, Wilson’s study is based on a small sample size of 30 examinees and 5 testlets. Few empirical studies have been conducted to investigate the impact of forming locally dependent items into testlets on test information when the Rasch models are used.
Chapter Two: Review of Literature

Score Estimates

Zenisky et al. (2002) compared the ability estimates between the two scoring schemes and found that the two scoring methods generated highly correlated but not identical results. The correlation between the two methods is at least 0.97, but some disparities in individual scores between the two scoring schemes were almost one half of one standard deviation, a difference of which may “have significant impact on an examinee’s percentile rank, performance classification, and other important uses of the scores” (Zenisky et al., 2002, p.304).

Unfortunately, a clear assessment of the impact of LID on ability estimates using the score-based approach is difficult to make because changing the scoring method to handle LID involves changing IRT models, as well as the associated number of parameters estimated. As Zenisky et al. (2002) have pointed out, “all of these changes can influence the ability estimates, as can model misfit and any differential impact of item parameter estimation due to use of one model of another” (p.302). Clearly, more research in this regard is needed with simulated data, so that more controls can be placed on the nature of the data.

In summary, the advantages of the score-based approach are that it is easy to interpret and to implement in practice, fewer parameters are estimated which leads to a more parsimonious representation, and no additional resources are required to analyze the
polytomous model. However, use of this approach is only practical for dichotomous dependent items and will result in loss of item-level information should the testlet score not be a sufficient statistic. It is also not appropriate when the test is administered adaptively, i.e., in computerized adaptive testing (Wainer, Bradlow, & Du, 2000).

**The Item-Based Testlet Approach**

As the name itself indicates, using the *item-based testlet approach* to accommodate LID retains item-level information. Local dependence is controlled by building explicit models based on standard IRT models. Several models following this approach have been developed, such as the Testlet Response Theory model (Wainer, Bradlow, and Wang, 2007), the full-information bi-factor item factor analysis (Gibbons & Hedeker, 1992; Gibbon et al., 2007), the item-interaction model (Hoskens & De Boeck, 1997; Tuerlinckx & De Boeck, 1998, 1999), the random coefficient multinomial log model (Wilson & Adams, 1995), and the conjunctive IRT model (Jannarone, 1997). In this section, the first three models (which seem to have gained popularity in the current literature) will be discussed in detail.

**Testlet Response Theory Model**

Development of the Testlet Response Theory (TRT) model is marked by several key papers. Bradlow, Wainer, and Wang (1999) studied the performance of the 2PL testlet model using simulated data, Wainer and Wang (2000) further employed the 3PL testlet
model to fit real test data, and finally, Wang, Bradlow, and Wainer (2002) extended from
modeling dichotomous data to include situations in which a test is composed partially, or
completely, of polytomous items and/or testlets. Recently, the entirety of this theory and
its applications have been published in the book titled “Testlet Response Theory and its
Applications” written by Wainer, Bradlow, and Wang (2007). Parallel to the development
of the theoretic framework of the TRT model, the corresponding computer program -
SCORIGHT (Wang, Bradlow, & Wainer, 2004) has also been developed to estimate item
and person parameters using a Gibbs Markov Chain Monte Carlo (MCMC) procedure.

The TRT Model (Wainer, Bradlow, & Wang, 2007) accommodates LID by
incorporating a person-testlet interaction effect for items nested within the same testlet.
Taking dichotomous items as an example, this interaction term is added to the standard
2PL or 3PL IRT model. In this way, items within the same testlet can still be scored
dichotomously, but the local dependence among items within a testlet is taken into
account. For example, the item response function for the conventional 3PL model is

\[ P(y_{ij} = 1) = c_j + (1 - c_j) \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]}, \]

where \( y_{ij} \) is the score for item \( j \) received by examinee \( i \)

\( P(y_{ij}=1) \) is the probability that examinee \( i \) answers item \( j \) correctly,

\( \theta_i \) is the trait value of examinee \( i \),
$a_j$ is the discrimination parameter of the item $j$.

$b_j$ is the difficulty parameter of the item $j$.

$c_j$ is the pseudo-guessing parameter of item $j$.

The new testlet model is defined as

$$P(y_{ij} = 1) = c_j + (1 - c_j) \frac{\exp[a_j(\theta_i - b_j - \gamma_{id(j)})]}{1 + \exp[a_j(\theta_i - b_j - \gamma_{id(j)})]},$$

where parameters $\theta_i$, $a_j$, $b_j$, and $c_j$ retain the same meaning as in the standard 3PL model.

The new element $\gamma_{id(j)}$ in the logit part denotes the extra person/testlet interaction effect, where subscript $i$ indicates examinee and subscript $d(j)$ indicates the testlet to which item $j$ belongs. Items belonging to the same testlet share the same amount of interaction effect.

In fact, the element $\gamma_{id(j)}$ can be understood as person $i$’s ability due to the presence of an extra dimension associated with testlet $j$ (Li, Bolt, & Fu, 2006). This can be shown through mathematical transformations on the logit element in the item response function of TRT (Thissen, 2006):

$$z_{ij} = a_j(\theta_i - b_j - \gamma_{id(j)}),$$

distributing the slope parameter:

$$z_{ij} = a_j \theta_i - a_j b_j - a_j \gamma_{id(j)},$$

re-labeling $\theta_i$ as $\theta_{1i}$, $\gamma_{id(j)}$ as $\theta_{id(j)}$, and let $d_j = a_j b_j$,

$$z_{ij} = a_j \theta_{1i} - d_j - a_j \theta_{id(j)} = a_j (\theta_{1i} - \theta_{id(j)}) - d_j.$$

Specification of the TRT model is embedded in a larger Bayesian hierarchical
framework by asserting Gaussian population distributions for the ability and item parameters and treating them as random effects. Specifically, the model assumes independent priors given by:

\[ \theta \sim N(0,1), \text{ to identify the model,} \]
\[ \log(a_j) \sim N(\mu_a, \sigma^2_a), \]
\[ b_j \sim N(\mu_b, \sigma^2_b), \]
\[ c_j \sim N(\mu_c, \sigma^2_c), \text{ and} \]
\[ \gamma_{id(j)} \sim N(0, \sigma^2_{d(j)}). \]

As shown, \( \gamma_{id(j)} \) is assumed to be normally distributed with a mean of 0 and a variance of \( \sigma^2_{d(j)} \). When the variance of \( \gamma_{id(j)} \) is 0, the model reduces to the conventional 3PL model, and when \( c_j=0 \) and the variance of \( \gamma_{id(j)} \) is 0, it reduces to the conventional 2PL model. Therefore, the standard 2PL and 3PL models are nested within the TRT model.

Bradlow et al. (1999) conducted a series of simulation studies using the 2PL testlet model and illustrated that both the item parameter estimates (discrimination and difficulty) and the ability estimates are less biased and more highly correlated with the true parameters, than if the LID-laden data were analyzed using the standard 2PL model. Consistent with the findings from previous studies, Bradlow also demonstrated that treating LID-laden data as if they were independent tends to result in overestimation of
test information and thus the precision level of the examinee proficiency.

Compared to the *score-based approach*, major advantages of the TRT model include that it allows for detection, explicit modeling, and assessment of the degree of item conditional dependence; it calculates the parameters of the model and the standard errors accurately after the items have been administered; it retains the dichotomous scoring scheme of dichotomous items; and it interprets item parameters (i.e., the slope and difficulty parameters) the same way as they are in the standard IRT models.

Moreover, using the computer software SCORIGHT, researchers are able to incorporate covariates in the analysis to go beyond the IRT results and help to understand why some examinees responded the way they did. In this sense, the program conducts an IRT analysis and regression analysis simultaneously.

However, complexity of the model also induces heavier computational burdens than the conventional IRT models. For a model consisting of 20 testlets, with 2PL, 3PL and polytomous items, and covariates for both item and person parameters, the entire estimation process takes about 2 and half hours for 4000 iterations using SCORIGHT (Wang, Bradlow, & Wainer, 2004). The actual running time could be longer or shorter depending on the speed of the computer processors, the complexity of the model, the size of the dataset, and the number of iterations specified.
The Full-Information Bi-Factor Item Factor Analysis

The Full-information bi-factor item factor analysis, hereafter abbreviated as the bi-factor model, was introduced previously in the section of Methods of Assessing LID. This approach handles LID from a multidimensional perspective and retains item-level information. The model specifies that each item is correlated with two factors (or dimensions) - a primary factor with which all items are correlated and a secondary factor with which only items common to the same stimuli are correlated. The primary factor and all the secondary factors are orthogonal to each other. Gibbons and Hedeker (1992) illustrated the theoretic framework and application of this model for dichotomously scored items, and Gibbons et al. (2007) further extended the model to analyze polytomously scored items.

Each item estimated using the bi-factor model has non-zero factor loadings on the primary factor and one of the secondary factors, zero factor loadings on other secondary factors, and one intercept. The factor loadings and intercepts can be transformed mathematically into the item discrimination parameters and item difficulty parameters, respectively, in the multidimensional IRT models (McLeod, Swygert, & Thissen, 2001).

The interesting thing here is that, although the TRT model and the bi-factor model were developed separately, the TRT model is actually a restricted version of the bi-factor model (DeMars, 2006). In the TRT model, $\gamma_{id(j)}$ can be interpreted as “a random shift in
examinee’s ability due to the presence of a secondary dimension associated with the passage” (Li, Bolt, & Fu, 2006). The item discrimination parameter associated with this secondary ability \( \gamma_{id(j)} \) is fixed to be equal to the item discrimination parameter associated with the primary ability in the TRT model; whereas in the bi-factor model, the factor loadings, or the item discrimination parameters, of the primary factor are independent of the factor loadings of the secondary factors. Therefore, the TRT model is nested within the bi-factor model.

Estimation of the bi-factor model on dichotomously scored items can be performed in the commercially available computer software - TESTFACT (du Toit, 2003), which uses marginal maximum likelihood to estimate item parameters and uses either the expected a posterior (EAP) method or maximum a posterior (MAP) method to estimate proficiency scores.

So far, four different models could be used to analyze dichotomously scored test with LID items: (a) ignoring that local dependence existed in the data and fitting the conventional 1-, 2-, or 3-PL models; (b) applying the scored-based approach to control LID; (c) applying the testlet model by inserting a person-testlet interaction effect; and (d) applying the bi-factor model. It has been previously shown that relationships exist among the three item-level models (a), (c), (d). Specifically, the conventional IRT model is nested within the testlet model, and the testlet model is nested within the bi-factor model.
Given the above four models, an interesting and practical question to ask is what are the differences and similarities among these models when they are used to analyze test data with LID. To investigate this question, DeMars (2006) conducted a comprehensive simulation study. She generated datasets using each of these four models, then each generated dataset was analyzed by each of the four models. Test reliability, item and proficiency estimates were compared and contrasted.

It should be noted that the testlet model used in this study was not exactly the same as the TRT model described in Wainer et al. (2007). Rather, it was an alternative testlet model proposed by Li et al. (2006). This testlet model differs from Wainer et al. (2007) in that it removes the constraint of equal slopes between the primary and secondary ability by allowing the within-testlet slope to be proportional to the primary slope.

Her study found that when the focus is on estimates of proficiency scores and not item parameters, any of the models will perform satisfactorily, though estimated reliability will be inflated for the standard dichotomous IRT model if items within testlets are not independent. However, the choice of the model had a bigger impact on the item parameter estimates. A comparison of item parameter estimation is restricted to the three item-level models because item parameters in the score-based model are not comparable to those of item-level models. Since the standard IRT model is nested with the testlet
model, which is nested within the bi-factor model, the standard IRT model is the least complex and the bi-factor model is the most complex. DeMars finds that using the bi-factor model or testlet model to analyze test data without local dependence does not lead to any bias, but does lead to a slightly higher root mean square error (RMSE), defined as the square root of the average squared difference between the estimated and true parameters; whereas using the standard IRT model to estimate local dependent items results in negatively biased slopes (overestimation of the slopes) and an increased RMSE. Generally, evidence favors the use of the more parsimonious testlet-effect model over the bi-factor model. But the bi-factor model can be estimated using commercially available software.

*Item-interaction Model*

The Testlet Response Theory (TRT) proposed by Wainer et al. (2007) models LID by inserting a person-testlet interaction term in the logit part [i.e., $a_j(\theta_i-b_j-\gamma[id])$] of the standard IRT model. An alternative to the person-testlet interaction is using the item-item, or inter-item, interaction effect to model LID. Research using the inter-item interaction is represented by five major studies conducted chronologically by Hoskens and De Boeck (1997), Tuerlincks and De Boeck (1998, 1999, & 2001), and Smits, De Boeck, and Hoskens (2003). This type of LID model is referred to as the item-interaction model in this paper.
As the name itself indicates, this model specifies an interaction effect among items to control LID. The item interaction term could be classified based on two factors - the interaction type and interaction modus (Hoskens & De Boeck, 1997, 1998). With respect to the interaction type, there is a distinction between combination and order interaction. Combination interaction refers to a situation in which two or more items are attached to the same stimuli, while order interaction refers to a situation in which a carry-on effect exists among a series of items. With respect to the interaction modus, there is a distinction between constant and dimension dependent interaction. Constant interaction specifies that the interaction effect is constant in magnitude and direction for all persons regardless of their proficiency level, while dimension-dependent interaction specifies that the strength and direction of the interaction depends on the position of the person on the latent trait.

By cross-tabulating two levels of the two factors, Hoskens and De Boeck (1997) examined four types of item-interaction models: combination dependence with constant effect, combination dependence with dimension-dependent effect, order dependence with constant effect, and order dependence with dimension-dependent effect. These four models differ in the formulation of their item interaction term. Tuerlinckx and De Boeck (1998) further proposed a generalized mathematical form for all types of item-interaction models.
The item-interaction model examines all possible response patterns in a test. For example, if a testlet contain \( n \) dichotomous items, there are \( 2^n \) different response patterns. Note that the number of response patterns increases fast as \( n \) increases, which tends to be a disadvantage for this type of model. Estimation of parameters could be carried out using loglinear models in a standard statistical package such as SPSS or SAS. Due to the complexity of this model, a detailed mathematical formulation is omitted here, but a thorough discussion can be found in Tuerlinckx and De Boeck (1998).

This model has great flexibility in modeling various types of item-interactions. Since the unit of analysis is response patterns, it is virtually able to model any type of item interactions. Unfortunately, this theoretic flexibility sometimes cannot outbalance its practical inconvenience. As pointed out previously, the major disadvantage is that only a limited number of items could be analyzed. For example, a test composed of 20 dichotomous items would have \( 2^{20} = 1,048,576 \) response patterns, which creates a huge computational burden.

**Approaches Employed in the Health Care Field**

Up to this date, the common practice in the health care field when handling LID items, is that when local dependence was suspected or detected, parameters were estimated with and without the problematic items, and comparisons were made to see whether results were robust to item dependencies (Hays et al., 2007; Hill et al., 2007;
Reeve et al., 2007). When item slopes were not very different between the model that included the locally dependent items and the model in which no LID items were included, results were robust and LID items could stay in the model. Even when substantial difference occurred, it was not necessarily to delete LID items. Hill et al. (2007) pointed out that, in CAT administration, when LID was identified for an item pair, it was not necessarily to exclude either one from the item bank. “Often, 1 item will be informative for examinees in 1 area of the latent trait, and the other will be similarly informative in another region” (Hill et al., 2007, p.46). In this case, LID items must be calibrated separately and linked. CAT programmers would then need to specify that these items should not be administered to the same examinee. In the case of fixed form surveys, locally dependent items should be avoided to appear in the same form.

**Gaps in the Literature**

As shown from the above reviews, LID studies during the past two decades have both breadth and depth. Since IRT is widely applied in educational assessments, most of these studies investigated LID specific to the context of large scale testing. As a generic measurement theory and technique, IRT is not limited to the educational context. In fact, its application has been recognized and used in other fields such as the medical field where health surveys are given to patients to evaluate their general physical functioning (McHorney & Monahan, 2004). It is true that results from LID studies conducted on
educational assessments can be generalized to other fields, but specific attention is also needed to address the characteristics of other fields.

The study proposed in this paper examines the application of IRT to health surveys. After reviewing the available LID studies, two inadequacies are identified in addressing LID common to health surveys.

**Inadequacy of Research on Order Dependence**

Hosken and De Boeck (1997) have classified local dependence into combination dependence and order dependence. In existing LID studies, almost all of them deal with combination dependence, be it passage dependence, speededness dependence, or item type dependence. This is not surprising because combination dependence is more often observed in large scale assessment in an educational context. Ferrara, Huynh, and Michaels (1999) examined LID caused by a multi-step item cluster which can be classified as order dependence. However, their study only verified the existence of LID caused by multi-step items, instead of taking actions to control the LID. The order dependence models proposed by Hosken and De Boeck (1997), and Tuerlinckx and De Boeck (1998) have some practical inconvenience and thus are rarely used in empirical studies.

As introduced in Chapter 1, the local dependence detected in health surveys is probably caused by the deterministic response patterns among certain items in health
surveys. It is hypothesized that the deterministic pattern might cause order dependence. But such a hypothesis needs to be verified by more research.

In the case of deterministic order dependence, the score-based testlet approach is the most appropriate way to manage LID. Since the testlet score is a sufficient statistic to determine item response patterns within each testlet, forming a testlet will not result in item-level information loss. However, questions such as to what extent the deterministic order dependence will influence the dimensionality assessment of the original data and parameter estimates, remain unanswered.

**Inadequacy of Research on the One-Parameter (Rasch) Model**

Most LID studies consider the item discrimination and pseudo-guessing parameters. The pseudo-guessing parameter is definitely not appropriate for health surveys and the discrimination parameter may, or may not, be needed depending on the data itself.

Previous studies have revealed that the item discrimination parameter tends to be overestimated in the presence of LID, and so does the test information. When item discrimination is not needed to model the test, this might not be true. Keller et al. (2003) pointed out that more often than not the Partial Credit Model will generate more test information than the sum of binary items. Again, no empirical study has ever examined this statement.
This proposed study intends to address the above two inadequacies in the existing LID literature. First, it will examine whether the deterministic order dependence detected in the dichotomous data influences dimensionality assessment, test reliability and test precision level as indicated by the test information function; and secondly, it will examine the extent to which item and ability parameter estimates are influenced by LID when a one parameter (Rasch) model is fitted to the data.

**Summary**

This chapter reviewed research examining problems related to violations of the local item independence assumption in IRT applications. These studies covered a broad range of topics such as the causes of LID, the consequences of LID, the methods of assessing the existence of LID, and approaches to managing LID. It has been shown that the fundamental cause of LID is the introduction of additional dimensions. The consequence of violating the local item independence assumption may lead to an overestimation of test reliability, item discrimination parameter and test information values. The impact of LID on item difficulty parameter and person proficiency estimates is minimal. Dimensionality tests involving various factor analyses, or specifically developed measures, such as Yen’s $Q_3$, $G^2$ and $X^2$, could be used to detect the existence of LID.

Two major approaches are commonly used to accommodate LID, namely, the
The score-based approach and the item-based approach. The score-based approach manages LID by changing the scoring scheme from a dichotomous model to a polytomous model, and the item-based approach manages LID by establishing explicit models. The score-based method is applicable for a broad range of situations but suffers from the drawback of losing item-level information. On the other hand, the item-based method retains item-level information but requires more additional resources in the parameter estimation.

Finally, two inadequacies of the current LID literature are identified in IRT applications to health surveys - the lack of research on order dependence and research on the one parameter (Rasch) model, both of which are intended to be addressed in this proposed study. The full research design for this study is presented in Chapter Three.
Chapter Three: Research Design

This chapter presents the complete research design for the current study, including an introduction, hypotheses, methodology and analyses. First, the instrument and data source are discussed, followed by hypotheses on the type of LID detected in the raw data and the impact of LID on certain test statistics and parameter estimates. Next, the selection of the score-based approach, used because of the type of LID, to accommodate the local dependence in the data is discussed. Then, six separate analyses are outlined to address each research question and the corresponding hypothesis. Last, in order to provide a reference point for comparison, each analysis conducted on the real data is also coupled with the same analysis conducted on simulated data where LID is absent.
Data Description

Instrument

The Pediatric Evaluation of Disability Inventory (PEDI) version 1.0, developed in the early 1990s (Haley, Coster, Ludlow, Haltiwanger, & Andrellos, 1992), is used for this study. The PEDI is a comprehensive clinical assessment instrument used to evaluate functional capability and performance in children ages 6 months to 7.5 years. In this instrument, functional capability is measured by identifying key daily activities the child demonstrates with mastery and competence. Functional performance of these daily activities is measured by the level of caregiver assistance needed to accomplish the activity. Both the capability and performance of functional activities are measured in three content domains: (i) self-care, (ii) mobility, and (iii) social function. These domains are represented by 197 items in the capability section and 20 items in the performance section of the PEDI. Within the section of 197 functional capability items, 73 are in the Self-Care domain, 59 in the Mobility domain, and 65 in the Social Function domain.

The present study will focus on the 59 capability items in the Mobility domain. These capability items are direct measures of the functional capability of the child, and these items provide sufficient detail to identify clinical patterns of deficiencies in functional skill attainment (Haley et al., 1992). The Mobility domain was selected based on a practical consideration of the small sample size involved in this study. Since only
347 participants will be included in the current study, the Mobility domain was chosen because it has the smallest number of functional capability items.

Each of the 59 Mobility items are dichotomously scored with “0” corresponding to \textit{unable} and “1” to \textit{capable} of performing the activity specified in each item. Items within this domain are further grouped by content topics, each content topic containing five or fewer items. In total, there are 13 content topics in the Mobility domain. Table 3.1 presents the complete set of content topics and the number of items per content. A full list of items in the Mobility domain is presented in Appendix A.

\begin{table}[h]
\centering
\caption{Content Coverage in the Mobility Domain}
\begin{tabular}{ll}
\hline
Content Topic & \# of items per content \\
\hline
1 & Toilet transfer & 5 \\
2 & Chair/wheelchair transfers & 5 \\
3 & Car transfers & 5 \\
4 & Bed mobility/transfer & 4 \\
5 & Tub transfers & 5 \\
6 & Indoor locomotion methods & 3 \\
7 & Indoor locomotion: Distance/speed & 5 \\
8 & Indoor locomotion: Pulls/carries objects & 5 \\
9 & Out door locomotion: Methods & 2 \\
10 & Indoor locomotion: Pulls/carries objects & 5 \\
11 & Outdoor locomotion: Surfaces & 5 \\
12 & Upstairs & 5 \\
13 & Downstairs & 5 \\
\hline
\textbf{Total} & \textbf{59} \\
\end{tabular}
\end{table}

\textit{Population and Sample}

The data used in this study are from the PEDI norms study, which collected
functional data on a representative sample of children ages 6 months to 7.5 years. The purposes of the norms study were to collect information from a “norm population” free from disabilities and to develop normative standard scores for each of the scales in the PEDI. The norm for the PEDI provides the reference point to which a child’s functional performance is compared and evaluated. The final norm group contains a regional New England sample of 412 children without disabilities, selected via a stratified quota sampling strategy according to demographic data derived from the 1980 United States Census (Haley, et al, 1992). In this final sample, ages are distributed almost equally across the span of 6 months through 7.5 years, with approximately 20-30 children in each age range of 0.5 years. Of the sample, 50.7% are female, 76.6% are white, and 69.6% are from urban areas (defined as populations greater than 50,000).

A strong ceiling effect was revealed after the examination of the raw score distribution of the entire sample. To reduce possible influence of ceiling effect on this study, 63 children aged 6 or above who showed the most severe ceiling effect (98% scored the highest or the next highest score) are excluded from this study. This reduced the final sample size from 412 to 347.

**Hypotheses**

After careful examining of the PEDI content, it is hypothesized that a certain degree of LID may exist among items within the same content topic. For example, five
items in the “Toilet Transfers” content topic read as

1. Can the child sit if supported by equipment or caregiver;
2. Can the child sit unsupported on toilet or potty chair;
3. Can the child get on and off low toilet or potty;
4. Can the child get on and off adult-sized toilet; and
5. Can the child get on and off toilet, not needing own arms.

Examination of the “Toilet Transfers” content items reveals that these five tasks are sorted in an increasing order by the difficulty level. It is almost certain that a child, who is capable of performing the activity stated in a specific item, is also capable of conducting the activities stated in all previous items. There is a deterministic pattern of the responses to these items. In other words, these items follow a Guttman scaling pattern, which is also known as cumulative scaling or scalogram analysis (Guttman, 1950). If responses to a set of items follow a Guttman scale strictly and these items are sorted by difficulty level, we would know exactly which items are answered correctly as long as we know the total score of these items.

For the five items in the “Toilet Transfers” content topic, response patterns 00000, 10000, 11000, 11100, 11110 and 11111 follow a Guttman scale strictly, and all other response patterns do not (e.g., 11010, 10110, 00111, etc.). According to Hoskens and De Boeck’s (1997) classification on local dependence (i.e., combination dependence vs.
order dependence), the Guttman scaling pattern could be regarded as a special case of order dependence. Therefore, being able to perform an activity stated in a certain item within a content topic, increases the probability of performing activities stated in subsequent items. The Guttman scaling pattern is also observed and hypothesized for most of the remaining content topics within the Mobility domain. The term \textit{deterministic order dependence} will be used to denote this type of LID.

Corresponding to the research questions posed in Chapter One and based on findings from previous research, five hypotheses are proposed:

1) the deterministic response pattern among items in the same content will cause items to be locally dependent;

2) the deterministic response pattern among items in the same content topic will influence the dimensionality assessment of test data;

3) the internal test reliability in Classical Test Theory will be overestimated with the presence of locally dependent items;

4) if a 2PL IRT model is fitted to the original data treating all items as if they were independent, the test precision level will be overestimated; whereas if a 1PL (Rasch) model is fitted, the test precision level will not be overestimated; and

5) score estimates will not be influenced when the deterministic pattern is
observed, regardless of which IRT model is employed.

Selection of an Approach to Manage LID

As introduced in Chapter Two, two basic methods are used to handle LID, the score-based method and the item-based method. Wainer et al. (2000) indicated that the score-based approach worked well in a broad array of situations and was a good practical approach except during two circumstances: 1) when an adaptive test is administered at the individual item level, and 2) when more information needs to be extracted from individual items. On the other hand, the item-based approach has an advantage over the score-based approach in that it retains the item-level information by explicitly modeling LID. However, the item-based approach involves building more complicated models and estimating more parameters.

Considering the nature of LID detected in the PEDI, as well as in other health surveys, the score-based method is assumed appropriate and will be adopted to manage potential local dependence for the following reasons.

The two circumstances under which the score-based approach is disadvantageous are not present in the current study. This study deals with the paper-and-pencil format of the survey, no adaptive testing is considered. Separate research should be conducted in the case of a computerized adaptive test. Further, when items combined into testlets follow a Guttman scaling, no item-level information will be lost because the total testlet
score becomes a sufficient statistic in describing the response pattern. Therefore, the benefit of using the item-based approach in retaining item-level information does not outbalance its disadvantages in estimating a more complicated model. Plenty of previous research has examined the impact of using the score-based approach on certain test statistics in the presence of combination dependence, but few have examined these impacts when items exhibit deterministic order dependence.

To use the score-based method, items within the same content topic will be formed into testlets. The same raw dataset is now presented in two forms – the dichotomously-scored data composed of 59 items, and the polytomously-scored data composed of 13 testlets. For notational purposes, the dichotomous real data will be referred to as Real_Dicho and the polytomous real data will be referred to as Real_Poly. The polytomous real data composed of testlets are free from LID. To investigate the impact of LID on test statistics, a conventional dichotomous IRT model will be fitted to the dichotomous data, treating all items as if they were independent of each other, and a polytomous IRT model will be fitted to the polytomous data. Then, internal reliability estimates, test information values, and score estimates generated from the two scoring schemes will be compared.

Since previous research found that LID had a negative impact on item discrimination estimates rather than on item difficulty parameter estimates, two separate
sets of IRT models will be employed for each analysis (outlined in the Analysis section):
1) IRT models not employing the item discrimination parameter (Rasch family models),
and 2) IRT models employing the item discrimination parameters (non-Rasch models).
The purpose of having two separate sets of IRT models is to address the last two research
questions, which examine whether the impact of LID on test statistics (i.e., test
information and score estimates) differs between models considering item discrimination
and models not considering item discrimination. No pseudo-guessing parameters will be
considered due to the fact that they are not appropriate in health surveys.

For the Rasch family of models, the standard dichotomous Rasch model (1PL
model) is used to analyze the dichotomous data, and the Partial Credit Model (Masters,
1982) is used to analyze the polytomous data. For the non-Rasch family models, the
standard two-parameter logistic (2PL) model is used to analyze the dichotomous data and
the Graded Response Model (Samejima, 1969) is used to analyze the polytomous real
data. Mathematical details for each model are illustrated in the next section.

\textbf{IRT Models}

\textit{Rasch Family Models}

The Item Response Function for the dichotomous Rasch model is defined as

\[ P_i(\theta) = \frac{e^{(\theta-b_i)}}{1 + e^{(\theta-b_i)}}, \quad (3.1) \]
Chapter Three: Research Design

\( P_i(\theta) = \) the probability that an individual with ability level \( \theta \) answers item \( i \) correctly. In the context of health surveys, it is the probability that an individual with ability level \( \theta \) is able to accomplish the task specified in item \( i \).
\( \theta = \) person ability parameter;
\( b_i = \) item difficulty parameter for item \( i \).

The Partial Credit Model (PCM) developed by Masters (1982) is the polytomous Rasch model. For item \( i \) taking scores of 0, 1, \( \ldots, m \), the PCM item response function is defined as:

\[
P_{ig}(\theta) = \frac{e^{\sum_{h}^{g} (\theta - b_h)}}{\sum_{h=0}^{m} e^{\sum_{l}^{g} (\theta - b_l)}}, \tag{3.2}
\]

\( P_{ig}(\theta) = \) the probability that an individual with ability level \( \theta \) responds in category \( g \) of item \( i \);
\( b_{il} = \) step parameter for the \( g^{th} \) category of item \( i \), the point at which the probabilities of endorsing adjacent categories are equal;
\( g = \) a specific category being modeled, \( g \in (0, 1, \ldots, m) \);
\( m = \) highest possible score of item \( i \);
\( l = 0, 1, \ldots, g \);
\( h = 0, 1, \ldots, m \).

**Non-Rasch Family Models**

The dichotomous 2PL model is defined as

\[
P_i(\theta) = \frac{e^{a_{i}(\theta - b_i)}}{1 + e^{a_{i}(\theta - b_i)}}, \tag{3.3}
\]

\( P_i(\theta) = \) the probability that an individual will answer an items correctly or in the context of the health survey, it is the probability that an individual is able to accomplish the task specified in item \( i \).
\( \theta = \) person ability parameter;
\( a_i = \) item discrimination (slope) parameter for item \( i \).
\( b_i = \) item difficulty parameter for item \( i \).
The Graded Response Model (GRM) is referred to as a two-step model, or a difference IRT model according to Thissen and Steinberg (1986). Computation of item category probability is achieved in two steps:

\[ P_{ig} = P_{ig}^* - P_{i(g+1)}^* \]  

(3.4)

\[ P_{ig}^* = \frac{e^{a_i(\theta - b_{ig})}}{1 + e^{a_i(\theta - b_{ig})}} \]  

(3.5)

- \( P_{ig} \) = the probability of responding in category \( g \) of item \( i \);
- \( P_{ig}^* \) = the probability of responding in \( g \) or higher categories of item \( i \);
- \( P_{i(g+1)}^* \) = the probability of responding in \((g+1)\) or higher categories of item \( i \);
- \( a_i \) = item discrimination parameter for item \( i \);
- \( b_{ig} \) = Thurstone threshold parameter for the \( g^{th} \) category of item \( i \).

Note that the \( b_{ig} \) in the GRM (Equation 3.5) is different from the \( b_{il} \) in the PCM (Equation 3.2). In the PCM, \( b_{il} \) is the step parameter; while in the GRM, \( b_{ig} \) is the Thurstone threshold. Step parameter is the intersection between two adjacent item category curves, or the point at which the probability of endorsing adjacent item categories are the same. Thurstone threshold is the point at which the probability of being observed below a given category is equal that of being observed in or above that category (Linacre, 1998). In other words, it is the point at which a subject has a 0.5 probability of endorsing the current or higher item categories. Step parameter estimates do not necessarily increase as the step sequence increases, while Thurstone thresholds do.
increase as the category sequence increases.

**Simulated Data**

The score-based approach accommodates LID by changing the scoring scheme from the dichotomous to polytomous models, therefore certain test statistics (i.e., test information) could be influenced by this change even when items are locally independent. To provide a reference for comparison when examining real data using the two scoring methods, simulated data consisting of independent items will be generated. Specifically, two dichotomous datasets will be simulated, each containing 59 items and 347 individuals, which is a reflection of the real dichotomous dataset structure. One dataset is generated from the Rasch dichotomous model and reflects a situation in which all items have similar slopes but different difficulty parameters. The other dataset is generated from the 2PL model and reflects a situation in which items differ not only in difficulties, but also in slopes. Each of the two dichotomous datasets will then be formed into two corresponding polytomous datasets by combining items into 13 testlets (as was done in the real dataset). Each simulation will be repeated 25 times to control for random error. Since the purpose of the simulation is to provide a reference point for comparison rather than examining standard errors, 25 simulations are considered enough. All simulations are conducted in SAS.

To generate the independent dichotomous Rasch data, the real data are first
analyzed using the dichotomous Rasch model. Then, the estimated item difficulty and person score parameters (derived from WINSTEPS) will be treated as the true item and person parameters. To simulate the binary response of individual $j$ on item $i$, the expected response (the probability of receiving a score of “1”) is first computed using the Item Response Function of the Rasch dichotomous model (Equation 3.1). Then, a random number is selected from the uniform distribution $[0,1]$. If the random number is smaller than or equal to the expected response, a value of 1 is assigned; otherwise, a value of 0 is assigned. For notation purposes, this simulated dataset is denoted as SIMU_dicho_1.

After this, the simulated Rasch dichotomous data will be converted to a polytomous dataset by combining items into testlets based on the same structure as the real dataset. This polytomous dataset is denoted as Simu_Poly_1.

To generate the independent dichotomous 2PL data, the real data is first analyzed using the 2PL model. Then, the estimated item difficulty, item discrimination and person parameters derived from MULTILOG will be treated as the true item and person parameters. A similar process is used to simulate the 2PL responses, except that the expected response is now computed from the Item Response Function for the 2PL model (Equation 3.3). This dataset is named SIMU_Dicho_2. Next, the polytomous dataset is generated by forming items into 13 testlets, and the polytomous dataset is named SIMU_Poly_2. To summarize, Table 3.2 lists dataset names.
Chapter Three: Research Design

Table 3.2: Dataset Names

<table>
<thead>
<tr>
<th>Real Data</th>
<th>Simulated Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch dataset</td>
<td>Simu_Dicho_1</td>
</tr>
<tr>
<td>2PL dataset</td>
<td>Simu_Dicho_2</td>
</tr>
<tr>
<td>59 dichotomous items</td>
<td>Real_Dicho</td>
</tr>
<tr>
<td>13 testlets</td>
<td>Real_Poly</td>
</tr>
<tr>
<td></td>
<td>Simu_Poly_1</td>
</tr>
<tr>
<td></td>
<td>Simu_Poly_2</td>
</tr>
</tbody>
</table>

Analyses

The entire Analysis section is composed of six major sections. The first section involves a series of exploratory analyses on the real data to (i) verify whether responses to items within the same content topic follow a Guttman scaling pattern, (ii) conduct a traditional item analysis, and (iii) to examine item/person parameter estimates after fitting different IRT models. In the second section, Yen’s $Q_3$ statistic is used to assess whether items within the same content topic exhibit local dependence in the dichotomous real data. In the third section, dimensionality of the dichotomous and polytomous real data will be evaluated using factor analysis on the tetrachoric correlation matrix. In the fourth section, Cronbach’s alpha is used to estimate internal reliability. In the fifth section, test precision levels derived from the two scoring schemes will be compared. Lastly, the sixth section involves comparing score estimates between the two scoring schemes when LID is present in the dichotomous data. Analyses of the second through sixth sections directly address the five research questions formulated in Chapter One.

Analyses 1: Exploratory Analyses

Existence of the Guttman Scaling Pattern
The examination of the content of items results in the formation of a hypothesized Guttman pattern among items within the same content topic. In order to verify this hypothesis, a descriptive measure is first taken to check response patterns of items within the same testlet. Specifically, for each testlet, the percentage of children whose responses strictly satisfy a Guttman scale will be calculated. Since a zero or perfect score always follow the Guttman scale and inclusion of these children may inflate the percentage, the calculation will be adjusted to include only those who do not have a zero or perfect score on that content topic. A higher percentage would indicate strong agreement with the Guttman pattern. Disordered individual response patterns will be identified.

In addition to the descriptive measure, the coefficient of reproducibility, proposed by Guttman (1950), can also be used to evaluate whether responses in a testlet follow a Guttman model. The coefficient is defined as

$$1 - \frac{\text{number of errors}}{\text{number of items} \times \text{number of persons}}.$$  

To count the number of errors, each test taker’s response on each item is examined. For example, in a test of five items, the Guttman response pattern for a raw score of 3 should be 11100. If the response pattern is 01110, two errors have occurred – in the 1st item and in the 4th item; similarly, if the response pattern is 00111, four errors have occurred – in the 1st, 2nd, 4th and 5th item. The number of errors defined in the above formula is the total
number of errors across all test takers. In this analysis, the coefficient of reproducibility
will be generated for each testlet. The standard of 0.90 (Guttman, 1950) will be used as a
cut point.

CTT Analysis

The traditional CTT item analysis is conducted on the dichotomous and
polytomous real data for the purpose of providing an overview of item difficulties and
item discriminations. According to the traditional CTT terminology for a dichotomous
dataset, item difficulty is defined as the percentage of responses scored “able.” Item
discrimination is defined as the Pearson correlation between item and total scores. Since
the terms of “item difficulty” and “item discrimination” are also used in the IRT analyses
but with different meanings, to avoid terminology confusion, item easiness is used in
place of item difficulty and item-total correlation is used in place of item discrimination
in the CTT analysis. The higher the item easiness, the easier the item is.

For the polytomous data, average testlet scores across all subjects and the Pearson
correlation between testlet and total scores are computed. Since in the raw data, each
testlet consists of different number of dichotomous items ranging from two to five, the
average testlet scores are not comparable among the 13 testlets. For example, a testlet
composed of 5 dichotomous items has an average between zero and five, while a testlet
composed of 2 items has an average between zero and two. In order to solve this problem
and make the average testlet scores comparable among the 13 testlets, the average testlet score will be adjusted based on the number of items in each testlet. Specifically, the adjusted testlet average is computed as dividing the average testlet score by the number of dichotomous items in each testlet. For example if the average testlet score for a five-item testlet is 3.6, the adjusted testlet average is $3.6/5=0.72$; if the average testlet score for a two-item testlet is 1.8, the adjusted testlet average is $1.8/2=0.9$. Thus, the adjusted testlet averages are comparable among the 13 testlets. Since each testlet is divided by a constant, this adjustment does not change the testlet-total score correlations.

**IRT Analysis**

As mentioned above, the real data will be fitted with Rasch models and non-Rasch models (fixed item slopes versus varying item slopes) on the dichotomous data composed of individual items, as well as on the polytomous data composed of testlets. Specifically, three sets of models are employed to the real data, as shown in Table 3.3. Item parameters and item fit statistics for each model will be reported.

<table>
<thead>
<tr>
<th>Table 3.3. IRT Models Fitted to the Real Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dichotomous Data Model (Software)</strong></td>
</tr>
<tr>
<td>Rasch Model</td>
</tr>
<tr>
<td>Non-Rasch Model</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

$^a$Partial Credit Model.  $^b$Graded Response Model.
Rasch models are estimated by WINSTEPS (Linacre, 2002) and non-Rasch models are estimated by MULTILOG (Thissen, Wang, & Bock, 2002). For non-Rasch models, 1PL and 2PL models are fitted to the dichotomous data, and corresponding Graded Response Models are fitted to the polytomous data. Though a simple Rasch model and a 1PL model are identical in terms of theoretical framework, WINSTEPS and MULTILOG use different algorithm in parameter estimation. Essentially, WINSTEPS does not estimate item slope at all, while MULTILOG estimates one common slope to all items. In addition, WINSTEPS uses joint maximum likelihood to generate item parameters, while MULTILOG uses marginal maximum likelihood, which assumes a unit normal population distribution to generate item parameters.

It should be noted that direct comparison between the fit of the 2PL model and the Rasch model (1PL model) to the original dichotomous data is not proper due to the fact that the slope parameters may be influenced by the existing LID. However, for the polytomously-scored real data, it is possible to check whether the model with varying item slopes fits better than the model fixing all item slopes to be equal. To achieve this, the polytomous real dataset will be analyzed twice using MULTILOG, 1) using the standard GRM model with varying item slopes and 2) using the GRM model but constraining item slopes to be equal. Then, the negative two times log-likelihood (-2LL) is compared and tested between the two runs. Since the GRM with equal slopes is nested
within the GRM model with varying slopes, a comparison between the two -2LL’s can be used to judge which model has a better fit. The lower the -2LL, the better the model fits the data (du Toit, 2003). The -2LL follows a chi-square distribution, so does the difference between the -2LLs, with degrees of freedom equal to the difference of the number of free parameters being estimated in each run. For example, the GRM with equal slopes will estimate 347 person ability parameters and 60 item-related parameters (1 common slope plus 59 item threshold parameters), and the GRM with varying slopes will estimate 347 person ability parameters and 72 item-related parameters (13 slopes and 59 item threshold parameters). The difference in the number of free parameters between the two models is 12.

**Analysis 2: Assessing LID Using Yen’s $Q_3$ statistic**

The $Q_3$ statistic (Yen, 1984, 1993) will be used to assess potential LID in the dichotomous real data. To calculate $Q_3$, the residual score of each individual on each item is first calculated by taking the difference between the observed and expected scores. Then, Pearson’s product-moment correlation coefficient is obtained from the residual scores of any two items across all examinees. Note that the expected scores are decided by a specific IRT model fitted to the data. In order to avoid the influence of model selection (i.e., Rasch model vs. 2PL model) on $Q_3$ computation, two sets of $Q_3$ statistics will be computed, one based on the expected scores derived from the Rasch model and
one based on the expected scores derived from the 2PL model. To provide a reference point for comparison, $Q_3$ statistics will also be computed based on the two simulated dichotomous datasets free from LID, Simu_Dicho_1 and Simu_Dicho_2.

If the local item independence assumption holds, $Q_3$ values between any two items would be near zero after partialing out the latent trait. For any dataset, $Q_3$ represents a residual correlation matrix. If there are $n$ items in the test, there will be $(n*(n-1)/2)$ unique item pairs and thus the same number of unique $Q_3$ values. In this analysis, $Q_3$ is represented by the lower triangle of the 59-by-59 residual correlation matrix.

Yen (1993) showed that when local item independence holds, the expected $Q_3$ value is equal to $-1/(n-1)$, where $n$ is the total number of items in a test. A large deviation of the observed $Q_3$ from the expected $Q_3$ indicates possible existence of LID. However, no common agreement has been reached as to which specific cut value should be used to claim substantial local item dependence. Fitzgeralds proposed an absolute value of 0.2 as the uniform critical value (as cited in Chen & Thissen, 1997). However, Chen and Thissen note that using this criterion in a test with a realistic length of 40-80 items will underestimate the type I error rates substantially, and result in very low power for $Q_3$.

Alternatively, diagnosing item inter-dependence can be achieved by comparing the average $Q_3$ values among dependent items against the average $Q_3$ among independent
items. In this study, items within the same content topic are hypothesized to be positively related with each other. Average $Q_3$ values for the sub-matrix of items from the same testlet (within-content $Q_3$) should be positive and large in value; whereas average $Q_3$ for the sub-matrix of items from different content topics (between-content $Q_3$) should be close to the expected value.

Since there are 13 contents (testlets) in the mobility domain, there will be 13 within-content average $Q_3$ values, and 78 ($78=13*12/2$) between-content average $Q_3$ values. The observed averages within- and between-content average $Q_3$ will be compared with an expected value of $-1/(n-1) = -1/(59-1)=-0.017$. Figure 3.1 shows the $Q_3$ matrix structure partitioned by content sub-matrixes, with $W$ representing a within-content sub-matrix and $B$ representing a between-content sub-matrix. Calculation formulas for the within- and between-content average $Q_3$ are illustrated in Figure 3.2 and Figure 3.3.
Chapter Three: Research Design

Figure 3.1: \( Q_3 \) Matrix Structure Partitioned by Content Sub-matrixes

\( W = \text{Within-content sub-matrix}; B = \text{Between-content sub-matrix} \)

<table>
<thead>
<tr>
<th>Content</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
<th>C8</th>
<th>C9</th>
<th>C10</th>
<th>C11</th>
<th>C12</th>
<th>C13</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>W1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td></td>
<td>B2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>B3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>B4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td></td>
<td>B5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td></td>
<td>B6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C7</td>
<td></td>
<td></td>
<td>B7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C8</td>
<td></td>
<td>B8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C9</td>
<td></td>
<td>B9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C10</td>
<td></td>
<td>B10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C11</td>
<td></td>
<td>B11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C12</td>
<td></td>
<td>B12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C13</td>
<td></td>
<td>B13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.2: Calculation of the Average Within-content \( Q_3 \)

\[
\text{Average within-content } Q_3 = \frac{2}{n(n-1)} \sum_{i,j(i<j)} Q_{ij} \quad (i, j \leq n)
\]

\( n = \text{number of items in the content} \)

Figure 3.2 represents the sub-matrix of \( Q_3 \) for five items in content 1, as well as the formula for calculating the average within-content \( Q_3 \) (the arithmetic mean \( Q_3 \) values of all possible pairs of items in the same content). Figure 3.3 presents the sub-matrix of \( Q_3 \) among items from two different contents - content 1 consisting of 5 items and content 2 consisting of 5 items, as well as the formula for calculating the average between-content \( Q_3 \).
**Figure 3.3: Calculation of the Average Between-content $Q_3$ (B$_{1,2}$ sub-matrix)**

<table>
<thead>
<tr>
<th>Content 1 (item $i$)</th>
<th>Item1</th>
<th>Item2</th>
<th>Item3</th>
<th>Item4</th>
<th>Item5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item6</td>
<td>$Q_{6,1}$</td>
<td>$Q_{6,2}$</td>
<td>$Q_{6,3}$</td>
<td>$Q_{6,4}$</td>
<td>$Q_{6,5}$</td>
</tr>
<tr>
<td>Item7</td>
<td>$Q_{7,1}$</td>
<td>$Q_{7,2}$</td>
<td>$Q_{7,3}$</td>
<td>$Q_{7,4}$</td>
<td>$Q_{7,5}$</td>
</tr>
<tr>
<td>Item8</td>
<td>$Q_{8,1}$</td>
<td>$Q_{8,2}$</td>
<td>$Q_{8,3}$</td>
<td>$Q_{8,4}$</td>
<td>$Q_{8,5}$</td>
</tr>
<tr>
<td>Item9</td>
<td>$Q_{9,1}$</td>
<td>$Q_{9,2}$</td>
<td>$Q_{9,3}$</td>
<td>$Q_{9,4}$</td>
<td>$Q_{9,5}$</td>
</tr>
<tr>
<td>Item10</td>
<td>$Q_{10,1}$</td>
<td>$Q_{10,2}$</td>
<td>$Q_{10,3}$</td>
<td>$Q_{10,4}$</td>
<td>$Q_{10,5}$</td>
</tr>
</tbody>
</table>

Average between-content $Q_3 = \frac{1}{mn} \sum_{i,j} Q_{ij} \quad (i \leq m, \ j \leq n)$

$m =$ number of items in content 1;  
$n =$ number of item in content 2.

The average within-content $Q_3$ will be plotted against the average between-content $Q_3$ for a visual presentation. Specifically, the relative frequency distribution of the 78 between-$Q_3$ values will be overlaid with the relative frequency distribution of the 13 within-$Q_3$ values. The within-$Q_3$ distribution is expected to lay on the right side of the between-$Q_3$ distribution.

In addition to the visual presentation, the Kruskal-Wallis test (Baglivo, 2005) will be used to test whether the 13 within-content $Q_3$ values are statistically larger than the 78 between-content $Q_3$ values. The Kruskal-Wallis (KW) test is a non-parametric approach and detects group differences by examining the rank order of observations. To conduct the KW-test in this analysis, the 13 averages of within-content and 78 between-content $Q_3$ values will be pooled and ranked. The KW statistic, $K$, is calculated from Equation 3.6:

$$K = \frac{12}{N(N + 1)} \sum_{i=1}^{2} n_i \left( \bar{R}_i - \frac{N + 1}{2} \right)^2 ,$$

where $N$ is the total number of observations in both groups, $i$ represents group
identity, \( n_i \) is the number of data points in the \( i^{th} \) group, and \( \bar{R}_i \) is the average rank of observations in the \( i^{th} \) group. The average of all \( N \) ranks is \((N+1)/2\). In this analysis, \( i \) takes two values – one for the within group and one for the between group. The distribution of \( K \) is approximately chi-square with one degrees of freedom for a two-group comparison, under the null hypothesis that the within-content \( Q_3 \) and the between-content \( Q_3 \) are equal (if there are no ties in the observations and when \( N \) is large).

When LID is not present in the dataset, the average rank of the within-\( Q_3 \) should be about the same as the average rank of the between-\( Q_3 \). However, when LID is present among items in the same content, the average rank of the within-\( Q_3 \) should be larger than that of the between-\( Q_3 \).

One potential problem of using the Kruskal-Wallis test to compare the within- and between-\( Q_3 \)’s is that the assumption of group independence required by the test might be violated, because the average within- and between-\( Q_3 \) originally come from pair-wise residual correlations of the same set of items. However, the results of the test are still considered valid since the purpose of this test is to conduct qualitative comparisons between the observed within- and between-\( Q_3 \)’s, as well as between the simulated within- and between- \( Q_3 \)'s, rather than conduct formal significance test.
Analysis 3: Dimensionality Assessment

Dimensionality assessment on both the dichotomous and polytomous real data will be accomplished through factor analysis (FA) using the computer program Mplus, which uses the tetrachoric correlation matrix as an input. Despite the previously discussed disadvantages of using the tetrachoric correlation matrix as an input matrix for factor analysis, this method remains a standard approach and has been used widely in the current literature to conduct factor analysis on categorical data.

The dimensionality assessments of the dichotomous and polytomous real data will be compared. The purpose of this analysis is to examine whether a primary factor can be observed in both the dichotomous dataset and the polytomous dataset.

To provide a reference for comparison, the same dimensionality assessment will be conducted on the simulated data (Simu_Dicho_1 and Simu_Poly_1), where LID is absent.

Analysis 4: Internal Reliability

Previous studies have shown that presence of LID may inflate test reliability. This is because interdependent items have a higher correlation with each other and thus show more consistent responses for the entire test, which results in a higher reliability. To examine this, test reliability for both the dichotomous and polytomous real data will be estimated by Cronbach’s alpha using Equation 3.7,
\[
\alpha = \frac{k}{k-1} \left( 1 - \frac{\sum \sigma_i^2}{\sigma_X^2} \right),
\]

where \(k\) is the number of items in the test, \(\sigma_i^2\) is the variance of item \(i\), and \(\sigma_X^2\) is the total score variance.

Also, to provide a reference for comparison, the alpha coefficient will be obtained from two sets of simulated datasets where LID is absent. Previous research (Sireci, Thissen, & Wainer, 1991; Zenisky, Hambleton, & Sireci, 2002; Wainer, 1995) shows that when LID is absent in a test, combining items into testlets will lead to an expected reliability drop of 2%-3%. However, when LID is present, the decrease ranges from 5%-10%.

**Analysis 5: Impact of the deterministic response pattern on test information and whether the impact differs between Rasch and non-Rasch models**

Test precision, as indicated by the test information function, derived from the dichotomous data will be compared to the test information derived from the polytomous data. The same analysis will be conducted both on the real data and the simulated data. To investigate whether the impact is the same when different IRT models are applied, three sets of contrasts are made between a dichotomous model and an equivalent polytomous model. For each contrast, total test information, as well as testlet information, will be compared between the two scoring methods.
Three sets of contrasts

Table 3.4 lists three contrasts corresponding to three different sets of IRT models. The term “contrast” used in this study refers to the comparison of analyses between two scoring schemes on the same test data: the dichotomously-scored test analyzed by a dichotomous IRT model versus the poltymously-scored test analyzed by a poltymous IRT model.

For Contrast 1, Rasch models are investigated. The simple dichotomous Rasch model is fitted to the dichotomous real data and the Partial Credit Model is fitted to the poltymous real data, both of which are estimated by WINSTEPS. To provide a reference point for comparison, the same set of analyses will be repeated on the simulated dichotomous Rasch data (Simu_Dicho_1) and the equivalent poltymous data (Simu_Poly_1).

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Software</th>
<th>Dichotomous</th>
<th>Poltymous</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>WINSTEPS</td>
<td>Rasch</td>
<td>PCM</td>
</tr>
<tr>
<td>2</td>
<td>MULTILOG</td>
<td>1PL</td>
<td>GRM – equal slopes</td>
</tr>
<tr>
<td>3</td>
<td>MULTILOG</td>
<td>2PL</td>
<td>GRM – varying slopes</td>
</tr>
</tbody>
</table>

For Contrast 2, the same 1-parameter logistic (1PL) model is fitted to the dichotomous data, but analyzed using MULTILOG. The corresponding poltymous
model will be Samejima’s (1969) Graded Response Model with equal slopes. Although the theoretic framework is the same for the simple Rasch model analyzed in WINSTEP and the 1PL model in MULTILOG, there are some technical differences in the estimation process conducted by the two computer programs. In WINSTEPS, item slopes are set to one and item/person parameters are estimated using the Joint Maximum Likelihood method. Whereas in MULTILOG, slope is estimated but restricted to be equal for all the items, and item parameters are estimated using Marginal Maximum Estimation, assuming the population distribution of person ability is normal with a mean of zero and a standard deviation of one, $\theta \sim N(0,1)$.

The purpose of including this contrast is to provide a transition when comparing the Rasch models estimated in WINSTEPS and the 2PL models estimated in MULTILOG. Direct comparison between these two sets of models is less informative because differences in computer programs are not considered. Contrast 2 serves as an intermediate between contrast 1 and contrast 3. For example, comparison between Contrasts 1 and 2 will suggest possible differences when two computer programs are used to estimate the same IRT model; and Contrast 2 and 3 will suggest a difference between two models estimated by the same program.

Contrast 2 will be conducted on the dichotomous and polytomous real data – Real_Dicho and Real_Poly, as well as on the simulated dichotomous and polytomous
Rasch (1PL) dataset – Simu_Dicho_1 and Simu_Poly_1.

For Contrast 3, the 2PL model is fitted to the dichotomous data and the Graded Response Model (allowing slopes to vary) is fitted to the polytomous test. Data are analyzed using MULTILOG. Contrast 3 will be conducted on both the real data - Real_Dicho and Real_Poly, and the two parameter simulated data - Simu_Dicho_2 and Simu_Poly_2.

*Item Information and Test Information*

Test information is an indicator of score precision level. Test information is the arithmetic sum of item information:

\[ T(\theta) = \sum_i I_i(\theta) \] (3.8)

The item information function is defined as

\[ I_i(\theta) = \frac{[P_i'(\theta)]^2}{P_i(\theta)Q_i(\theta)}, \] (3.9)

where \( P_i(\theta) \) is obtained from the Item Response Function of a specific IRT model, \( P_i'(\theta) \) is the first derivative of \( P_i(\theta) \) with respect to \( \theta \), and \( Q_i(\theta) = 1 - P_i(\theta) \).

For the dichotomous models, the item information function can be reduced to:

\[ I_i(\theta) = a_i^2 P_i(\theta)Q_i(\theta); \] (3.10)

and for the polytomous IRT models, the item information function, for an item with \( m \) categories, can be transformed to:
Chapter Three: Research Design

\( I_i(\theta) = \sum_{g=1}^{m} \left( \frac{[P^\prime_{ig}(\theta)]^2}{P_{ig}(\theta)} - P^\prime_{ig}(\theta) \right) \)  \( (3.11) \)

where \( P_{ig}(\theta) \) is the probability of giving a response in category \( g \) of item \( i \). \( P^\prime_{ig}(\theta) \) is the first derivative of \( P_{ig}(\theta) \) with respect to \( \theta \), and \( P^\prime\prime_{ig}(\theta) \) is the second derivative of \( P_{ig}(\theta) \) with respect to \( \theta \).

**Linking item parameters from two separate calibrations**

Since the estimation process for the dichotomous data and the polytomous data in each contrast involves separate calibrations, item parameter estimates from the two analyses have to be on the same metric in order to compare test information values. For Contrasts 2 and 3, since MULTILOG always sets the mean of the ability estimates at zero and the standard deviation at one and since the same group of subjects is involved in each calibration, item parameters from the two runs are automatically put on the same scale and no linking is needed (Thompson & Pommerich, 1996).

However, Rasch models in Contrast 1 are identified by standardizing item difficulties rather than person abilities (Baker & Kim, 2004). In other words, the metric of parameter estimates are determined by setting the mean and standard deviation of item difficulty estimates at zero and one respectively. Since two different sets of items are calibrated (i.e., 59 dichotomous items versus 13 polytomous items), the metric of item parameters derived from two separate calibrations would be different. To solve this
problem, person scores are used to equate the two sets of item parameters. Specifically, the PCM model is fitted to the polytomous data. Then, score estimates derived from the polytomous run will be fixed when estimating item parameters in the dichotomous data. In this case, item parameters from two runs are put on the same scale.

**Analysis 6: Impact of the deterministic response pattern on score estimates and whether the impact differs between 1PL and 2PL models**

Score estimates and standard errors of estimates for the 347 subjects derived from the dichotomous data will be compared to those derived from the polytomous data. The same analysis will be conducted both on the real data and the simulated data. To investigate whether the impact is the same when different IRT models are applied, the same sets of contrasts, as used in the previous section, will be repeated here.

**Score Estimates**

Scores generated from each estimation process will be rescaled to have a mean of zero and a standard deviation of one. The dichotomously-estimated scores will be correlated with the polytomously-estimated scores in each contrast. Scatter plots and Pearson correlations will be reported for the two sets of scores.

**Standard Error of Estimates**

The standard error of estimate is a statistic closely related to the test information function and is defined as the square root of the inverse test information:
Standard errors of estimates provide very similar results as the test information. In addition, it provides a specific error value for each individual estimate and thus, can be used to establish confidence intervals around the point estimate for each person.

**Summary**

This Chapter outlines the complete analysis plan to address the research questions posed in Chapter One. To investigate the local dependence problem detected in health surveys, the dataset collected from the norms study of the Pediatric Evaluation of the Disability Inventory is employed. It is observed that items attached to the same content topic may have a deterministic response pattern, or the Guttman scaling pattern. Exploratory analysis is first conducted to verify the existence of the Guttman scaling pattern and to examine item parameter estimates when different models are fitted to the original data. The score-based method is proposed to manage LID, then five major analyses are conducted to test the hypotheses of whether the Guttman scaling pattern causes items to be locally dependent, affects assessment of data dimensionality, inflates test reliability, and influences score estimates. The analyses will also explore whether differences occur when different IRT models are fitted to the original data. Results of the above outlined analyses are presented in Chapter Four.
Chapter Four: Results

This Chapter presents the data analyses outlined in Chapter Three. The results are organized into six sections, including a series of exploratory analyses and five subsequent analyses corresponding to the five research questions and hypotheses. To recapitulate, the score-based method is adopted to accommodate LID in the original dichotomous data. Items belonging to the same testlet – content topic – are combined and treated as one polytomous item. Dimensionality assessment, internal reliability estimates, test information, and person score estimates are compared between the two scoring schemes.

A series of exploratory analysis is first conducted on the raw data, including an examination of the Guttman scaling pattern of item responses in the same testlet, a traditional CTT item analysis, and several IRT analyses on the dichotomous and the polytomous data. In the second section, local item dependence is assessed using $Q_3$ statistics proposed by Yen (1984, 1993). In the third section, dimensionality structure is evaluated using Mplus. In the fourth section, internal reliability is estimated. In the last two sections, score estimates and score precision levels are compared between the two scoring schemes when different IRT models are fitted.
Analysis 1- Exploratory Analyses

The following exploratory analyses include a Guttman scaling analysis examining item response patterns in each testlet; a traditional CTT analysis on item difficulties and item discriminations; and several IRT analyses on the dichotomously-scored and polytomously-scored raw data.

Guttman Scaling Analysis

This analysis aims to verify whether responses to the dichotomous items in each testlet follow a Guttman scale. A matching response (e.g., 10000, 11100) indicates agreement with the cumulative Guttman scaling, while a disordered response (e.g., 10100, 01100) suggests a non-Guttman scale. Table 4.1 presents the results for the Guttman analysis. The Table lists the number of items in each testlet, the percentage of matching responses (not counting perfect/zero testlet scores), the coefficients of reproducibility proposed by Guttman (1950), the total number of disordered responses, and the frequencies of disordered patterns.

As shown, both the percentages of matching responses and coefficients of reproducibility are very high for each of the 13 testlets, indicating high agreement with the Guttman scale. Even the lowest coefficient of reproducibility for Testlet 5 is 0.97, which is much higher than the cut point of 0.90 suggested by Guttman (1950) to determine a Guttman scale. This high agreement provides evidence that combining dichotomous items into testlets leads to a minimal loss of item-level information, since the total testlet score is sufficient to describe a response pattern. Therefore, the score-based method is deemed reasonable to accommodate potential LID in the survey.
Table 4.1. Guttman Scaling Analysis, by Testlet

<table>
<thead>
<tr>
<th>Testlet</th>
<th>Number of Items</th>
<th>Percent of Matching Patterns</th>
<th>Coefficient of Reproducibility</th>
<th>Number of Disordered Responses</th>
<th>Frequency of Disordered Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>98.8%</td>
<td>0.998</td>
<td>2</td>
<td>10101 10110 1</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>99.6%</td>
<td>0.999</td>
<td>1</td>
<td>01111 1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>100.0%</td>
<td>1.000</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>100.0%</td>
<td>1.000</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>82.6%</td>
<td>0.972</td>
<td>24</td>
<td>01010 11010 1 23</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>100.0%</td>
<td>1.000</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>100.0%</td>
<td>1.000</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>97.5%</td>
<td>0.997</td>
<td>3</td>
<td>10100 10110 11011 1</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>100.0%</td>
<td>1.000</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>98.1%</td>
<td>0.998</td>
<td>1</td>
<td>00011 1</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>96.3%</td>
<td>0.998</td>
<td>2</td>
<td>00100 11011 1</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>97.7%</td>
<td>0.998</td>
<td>2</td>
<td>10100 01111 1</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>98.8%</td>
<td>0.999</td>
<td>1</td>
<td>10100 1</td>
</tr>
</tbody>
</table>

*aNumber of dichotomous items in each testlet.
*bPercent of responses (not including perfect and zero testlet scores) following a Guttman scaling pattern.
*cCoefficient of Reproducibility = 1-[number of errors / (number of examinees*number of items)].

An examination of the number of disordered responses and frequency of disordered patterns, listed in the last three columns of the table, helps us identify possible reasons for the occurrence of disorder. Occurrence of a single disordered pattern may indicate an occasional error such as an input error, while occurrence of repeated disordered pattern may indicate that item difficulty levels do not have a hierarchical order as hypothesized. For example, Testlet 5 has 24 disordered responses. However, only two patterns emerged – pattern 01010 occurred once and pattern 11010 occurred 23 times. To further investigate possible causes, it is necessary to look at the actual items in this testlet:
1. Sits if supported by equipment or caregiver in a tub or sink;
2. Sits unsupported and moves in tub;
3. Climbs or scoots in and out of tub;
4. Sits down and stands up from inside tub; and
5. Steps/transfers into and out of an adult-sized tub.

An input error could have occurred for the first item in pattern 01010 (Item 20) because it is unlikely that a child is unable to do Item 1, but able to do Item 2. Correcting this error leads to a response pattern of 11010, the same as the remaining 23 disordered responses, in which a disorder happens between Item 3 and Item 4 – unable to do Item 3 (Climbs or scoots in and out of tub) but able to do Item 4 (Sits down and stands up from inside tub) - which is likely to happen for some children. An alternative explanation for this disorder is that parents or caregivers who filled out the survey never saw their children doing this before and thus assumed they were unable to do it.

**CTT Item Analysis**

Figure 4.1 presents the distributions of item easiness and item-total score correlations for the dichotomous dataset. Figure 4.2 presents the same information for the polytomous raw data. Actual values used to generate these two figures are presented in Table 4.2 and Table 4.3.

As shown in Figure 4.1 and Table 4.2, for the dichotomous dataset, with an average item easiness of around 0.80, this survey is relatively easy for the sample. High item easiness values indicate a possible ceiling effect, as many health outcome measures may encounter. The easiest item is Item 20 - only one out of the 347 sampled children
was unable to do this item. As discussed in the previous section, this “unable” response for Item 20 (first item in Testlet 5) is very likely to be an input error. The hardest item is Item 10 – less than one third (31.4%) of the children were able to do this item.

The average item-total correlation is around 0.69. Item 20 has a negative correlation, which is probably caused by an input error for the only one “unable” response to this item. Should this error be corrected, there would be no variation for this item and thus the item-total score covariance would be zero (item-total score correlation would not be determined because the item variance is zero). Other items having low item-total correlations (lower than 0.3) are Item 6, Item 25 and Item 33, all of which have very high item easiness values (larger than 0.98).

The results of the CTT analysis on the polytomous data convey similar information, as illustrated in Figure 4.2 and Table 4.3. Since test length has reduced from 59 items in the dichotomous data to 13 testlets in the polytomous dataset, distributions of the adjusted testlet averages (i.e., testlet score divided by the number of items in the testlet) and testlet-total score correlations are less various than those in the dichotomous data. The ceiling effect in the polytomous data is not as severe as that in the dichotomous data. As shown, the highest adjusted testlet average is 0.926, as compared to the highest item easiness of 0.997 in the dichotomous data. This is because when dichotomous items are combined into testlets, the chance of having a perfect testlet score is smaller than the chance of having a perfect dichotomous item score. All testlet-total score correlations are very high. Two testlets have correlations larger than 0.90, 12 larger than 0.80, and even the lowest correlation is 0.79.
To investigate the relationship between item easiness and item-total score correlations in the dichotomous data, and between the adjusted testlet score and testlet-total score correlations, scatter plots are constructed between the two statistics in Figure 4.3. As shown in Figure 4.3A for the dichotomous data, there is a curvilinear relationship for the dichotomous data. For item easiness smaller than 0.8, item-total score correlation increases as items get easier; for item easiness larger than 0.8, item-total score correlation tends to decrease as items get easier. Whereas for the polytomous data, as shown in Figure 4.3B, there is no relationship between adjusted testlet scores and testlet-total score correlations.
Figure 4.1. Dichotomous Raw Data: Distributions of Item Easiness and Item-Total Score Correlations

A. Item Easiness

B. Item-Total Score Correlations

Figure 4.2. Polytomous Raw Data: Distributions of Adjusted Testlet Averages and Testlet-Total Score Correlations

A. Adjusted Testlet Averages

B. Testlet-Total Score Correlations

Figure 4.3. Scatter Plot between Item Easiness and Item-Total Score Correlations

A. Dichotomous Data

B. Polytomous Data
Table 4.2. CTT Analysis on the Dichotomous Raw Data

<table>
<thead>
<tr>
<th>Item</th>
<th>Testlet</th>
<th>Easiness&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Item-Total Correlation&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Item</th>
<th>Testlet</th>
<th>Easiness&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Item-Total Correlation&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.818</td>
<td>0.739</td>
<td>31</td>
<td>7</td>
<td>0.925</td>
<td>0.644</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.769</td>
<td>0.836</td>
<td>32</td>
<td>7</td>
<td>0.700</td>
<td>0.729</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.729</td>
<td>0.843</td>
<td>33</td>
<td>8</td>
<td>0.988</td>
<td>0.297</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.617</td>
<td>0.772</td>
<td>34</td>
<td>8</td>
<td>0.954</td>
<td>0.543</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.331</td>
<td>0.509</td>
<td>35</td>
<td>8</td>
<td>0.939</td>
<td>0.584</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0.988</td>
<td>0.178</td>
<td>36</td>
<td>8</td>
<td>0.876</td>
<td>0.735</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.942</td>
<td>0.589</td>
<td>37</td>
<td>8</td>
<td>0.643</td>
<td>0.735</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0.879</td>
<td>0.770</td>
<td>38</td>
<td>9</td>
<td>0.893</td>
<td>0.769</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0.790</td>
<td>0.762</td>
<td>39</td>
<td>9</td>
<td>0.844</td>
<td>0.866</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.314</td>
<td>0.481</td>
<td>40</td>
<td>10</td>
<td>0.911</td>
<td>0.693</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>0.833</td>
<td>0.813</td>
<td>41</td>
<td>10</td>
<td>0.841</td>
<td>0.856</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>0.729</td>
<td>0.851</td>
<td>42</td>
<td>10</td>
<td>0.804</td>
<td>0.853</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>0.637</td>
<td>0.790</td>
<td>43</td>
<td>10</td>
<td>0.784</td>
<td>0.837</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>0.458</td>
<td>0.630</td>
<td>44</td>
<td>10</td>
<td>0.758</td>
<td>0.799</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>0.334</td>
<td>0.506</td>
<td>45</td>
<td>11</td>
<td>0.914</td>
<td>0.702</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>0.977</td>
<td>0.393</td>
<td>46</td>
<td>11</td>
<td>0.876</td>
<td>0.784</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>0.850</td>
<td>0.777</td>
<td>47</td>
<td>11</td>
<td>0.850</td>
<td>0.788</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>0.807</td>
<td>0.831</td>
<td>48</td>
<td>11</td>
<td>0.816</td>
<td>0.864</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>0.429</td>
<td>0.549</td>
<td>49</td>
<td>11</td>
<td>0.764</td>
<td>0.841</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>0.997</td>
<td>-0.019</td>
<td>50</td>
<td>12</td>
<td>0.908</td>
<td>0.703</td>
</tr>
<tr>
<td>21</td>
<td>5</td>
<td>0.942</td>
<td>0.598</td>
<td>51</td>
<td>12</td>
<td>0.876</td>
<td>0.796</td>
</tr>
<tr>
<td>22</td>
<td>5</td>
<td>0.738</td>
<td>0.793</td>
<td>52</td>
<td>12</td>
<td>0.784</td>
<td>0.869</td>
</tr>
<tr>
<td>23</td>
<td>5</td>
<td>0.784</td>
<td>0.768</td>
<td>53</td>
<td>12</td>
<td>0.735</td>
<td>0.860</td>
</tr>
<tr>
<td>24</td>
<td>5</td>
<td>0.545</td>
<td>0.684</td>
<td>54</td>
<td>12</td>
<td>0.660</td>
<td>0.787</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>0.991</td>
<td>0.267</td>
<td>55</td>
<td>13</td>
<td>0.862</td>
<td>0.811</td>
</tr>
<tr>
<td>26</td>
<td>6</td>
<td>0.922</td>
<td>0.666</td>
<td>56</td>
<td>13</td>
<td>0.821</td>
<td>0.874</td>
</tr>
<tr>
<td>27</td>
<td>6</td>
<td>0.865</td>
<td>0.842</td>
<td>57</td>
<td>13</td>
<td>0.764</td>
<td>0.870</td>
</tr>
<tr>
<td>28</td>
<td>7</td>
<td>0.980</td>
<td>0.390</td>
<td>58</td>
<td>13</td>
<td>0.723</td>
<td>0.853</td>
</tr>
<tr>
<td>29</td>
<td>7</td>
<td>0.960</td>
<td>0.500</td>
<td>59</td>
<td>13</td>
<td>0.631</td>
<td>0.774</td>
</tr>
<tr>
<td>30</td>
<td>7</td>
<td>0.931</td>
<td>0.635</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Mean 0.792  0.693  (SD) 0.167  0.190  Median 0.821  0.769  Min 0.314  -0.019  Max 0.997  0.874

<sup>a</sup>Proportion of participants who are able to perform the task.

<sup>b</sup>Pearson correlation between the item score and the total test score.
Table 4.3. CTT Analysis on the Polytomous Raw Data

<table>
<thead>
<tr>
<th>Testlet</th>
<th>Number of items</th>
<th>Average Testlet Score&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Adjusted Average Testlet Score&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Testlet-Total Score Correlation&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>3.265</td>
<td>0.653</td>
<td>0.883</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>3.914</td>
<td>0.783</td>
<td>0.851</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>2.991</td>
<td>0.598</td>
<td>0.866</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3.063</td>
<td>0.766</td>
<td>0.861</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>4.006</td>
<td>0.801</td>
<td>0.876</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>2.778</td>
<td>0.926</td>
<td>0.815</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>4.496</td>
<td>0.899</td>
<td>0.787</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>4.401</td>
<td>0.880</td>
<td>0.850</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>1.738</td>
<td>0.869</td>
<td>0.864</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>4.098</td>
<td>0.820</td>
<td>0.896</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>4.219</td>
<td>0.844</td>
<td>0.895</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>3.963</td>
<td>0.793</td>
<td>0.940</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>3.801</td>
<td>0.760</td>
<td>0.939</td>
</tr>
</tbody>
</table>

Mean (SD) 3.595 (0.786) 0.799 (0.093) 0.871 (0.043)

Median 3.914 0.801 0.866
Min 1.738 0.598 0.787
Max 4.496 0.926 0.940

<sup>a</sup>Average testlet score across all subjects.
<sup>b</sup>Average testlet score divided by the number of items in each testlet.
<sup>c</sup>Pearson correlation between the testlet score and the total test score.

**IRT Analysis**

The structure of the raw data is presented in two formats, the dichotomous data composed of 59 individual items and the polytomous data composed of 13 testlets. Three dichotomous IRT models are fitted to the dichotomous raw data and three comparable polytomous IRT models are fitted to the polytomous raw data. The detailed model specifications are shown in Table 3.3, which is repeated as follows.

Table 3.3. IRT Models Fitted to the Real Data

<table>
<thead>
<tr>
<th></th>
<th>Dichotomous Data Model (Software)</th>
<th>Polytomous Data Model (Software)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rasch Model</td>
<td>Simple Rasch (WINSTEPS)</td>
<td>PCM&lt;sup&gt;c&lt;/sup&gt; (WINSTEPS)</td>
</tr>
<tr>
<td>Non-Rasch Model</td>
<td>1PL Model (MULTILOG)</td>
<td>GRM&lt;sup&gt;b&lt;/sup&gt; – fixed slopes (MULTILOG)</td>
</tr>
<tr>
<td></td>
<td>2PL Model (MULTILOG)</td>
<td>GRM&lt;sup&gt;b&lt;/sup&gt; – varying slopes (MULTILOG)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Partial Credit Model.  <sup>b</sup>Graded Response Model.
Item parameter estimates and relevant fit statistics are presented for each set of models. Since the polytomous data is reconstructed from the dichotomous data, comparisons are made between item parameters estimated from the dichotomous data and from the polytomous data. In the dichotomous data, 59 item difficulties are estimated; and in the polytomous data, 59 item step/threshold parameters are estimated. The 59 item difficulties are comparable to the 59 step/threshold parameters. A scatter plot is constructed between both the 59 dichotomous item difficulty estimates and the 59 steps/thresholds estimated from the corresponding polytomous model. In each scatter plot, a reference line of \( y = x \) is added to indicate possible scaling differences between item parameter estimates generated from a dichotomous model and item parameter estimates generated from a corresponding polytomous model. Results from each of the three sets of models, as specified in Table 3.3, are illustrated in the following section.

**Rasch Models**

The simple Rasch model is fitted to the dichotomous data. Table 4.4 presents item difficulty estimates, standard error (SE) of estimates, and two fit statistics. In addition, item number, testlet number each item belongs to, and the order of the item appearing in each testlet are also listed in the table.

Item difficulty estimates range from -8.40 (Item 20) to 7.53 (Item 10). Most of the extremely easy items (item difficulty estimate smaller than -4) are the first items in the testlets, and most of the extremely hard items (item difficulty estimate larger than 4) are the last items in the testlets. The Pearson correlation between item orders and item difficulty estimates is 0.75. In general, the item difficulty estimate increases as the item
order increases in each testlet, which also supports the original Guttman scaling hypothesis. The only exception occurs on Items 22 and 23. From the Guttman scaling analysis at the beginning of this chapter, it is shown that the difficulty level of these two items (the 3rd and 4th item in Testlet 5) might be reversed.

Two fit statistics are reported: the information-weighted mean-square (INFIT MSQ) and the outlier-sensitive mean-square (OUTFIT MSQ). INFIT MSQ is computed as

\[ V_i = \frac{\sum_{j=1}^{N} w_{ij} z_{ij}^2}{\sum_{j=1}^{N} w_{ij}}, \]

and OUTFIT MSQ is computed as,

\[ V_i = \frac{\sum_{j=1}^{N} z_{ij}^2}{N-1}, \]

where

\[ z_{ij}^2 = \frac{X_{ij} - P_{ij}(\theta)}{\sqrt{[P_{ij}(\theta)][1-P_{ij}(\theta)]}} \quad \text{and} \quad w_{ij} = [P_{ij}(\theta)][1-P_{ij}(\theta)], \]

\[ X_{ij} = \text{observed score of person } v \text{ on item } i; \]
\[ P_{ij}(\theta) = \text{predicted score of person } v \text{ on item } i; \]
\[ N = \text{total number of examinees.} \]

Since the OUTFIT mean-square is highly influenced by response outliers, the INFIT mean-square is used to identify misfit items. It is commonly accepted that items with INFIT values between 0.7 and 1.3 are considered a good fit. INFIT larger than 1.3 is considered misfit and INFIT smaller than 0.7 is considered overfit or redundancy. Based
on this criterion, 5 out of the 59 items are classified as misfitting and 11 are classified as overfitting.

The Partial Credit Model is fitted to the polytomously-scored data, and the results are presented in Table 4.5. The Partial Credit Model is used in place of the Rating Scale model because the number of categories in each testlet is different - ranging from two to five. Item category steps, INFIT and OUTFIT mean-squares are all listed in the table. Most step estimates are within -5 to 5. The average step estimate, across the 13 testlets, increases as the number of step increases. No misfitting items are identified and two items are identified as overfitting.

To investigate possible relationships between the item parameter estimates generated from the two models, a scatter plot of the 59 item difficulty estimates (based on the dichotomous Rasch model) against the 59 item step estimates (based on the Partial Credit Model) is presented in Figure 4.4. An obvious linear relationship is detected, with one outlier – Item 20. The difficulty estimate for Item 20, generated from the dichotomous model, seems much lower than what can be predicted from its equivalent step estimate. The Pearson correlation between the two sets of item parameter estimates is .89 when including all items, and .95 after excluding the outlier.

In the PCM, the item step parameters are estimated, while in GRM, as will be discussed later, the Thurstone thresholds are modeled and estimated. In order to make the scatter plot generated in Rasch models comparable to the other two plots generated from non-Rasch models, step parameters estimated in PCM are converted to Thurstone threshold estimates – the point at which a subject has a 0.5 probability of endorsing the
current or higher item categories, as shown in Table 4.6. The scatter plot between the two sets of item parameters is presented in Figure 4.5. As shown, when dichotomous items are formed into testlets, item difficulty estimates derived from the dichotomous model correlated higher with Thurstone threshold estimates ($r=0.96$) than with step estimates ($r=0.89$). Figure 4.5 also reveals that item parameter estimates generated from the dichotomous model and those generated from the PCM do not lie around the reference line of $y=x$, which, to some degree, implies that these two sets of estimates are not on the same metric. As explained previously in Chapter Three, this is because the metric of parameters in Rasch models is determined by standardizing item parameter estimates separately in each calibration. Since two different sets of items are estimated in the dichotomous and polytomous models, parameter estimates generated from these two separate runs are not on the same metric.

As a general summary, both Rasch models generated comparable item parameters for non-extreme items. Five misfitting items are identified in the dichotomous model and no misfitting items in the polytomous model. Since fit is not the focus of this study and item misfit may due to the existence of LID, these misfitting items will stay in the data for other analyses.
### Table 4.4. Item Parameter Estimates and Fit Statistics - Dichotomous Rasch Model

<table>
<thead>
<tr>
<th>Item</th>
<th>Testlet</th>
<th>Item Order&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Diff</th>
<th>SE</th>
<th>IN.MSQ</th>
<th>OUT.MSQ</th>
<th>Item</th>
<th>Testlet</th>
<th>Item Order&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Diff</th>
<th>SE</th>
<th>IN.MSQ</th>
<th>OUT.MSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.35</td>
<td>0.27</td>
<td>1.83</td>
<td>3.64</td>
<td>31</td>
<td>7</td>
<td>4</td>
<td>-2.86</td>
<td>0.32</td>
<td>1.05</td>
<td>0.26</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1.47</td>
<td>0.24</td>
<td>0.99</td>
<td>1.02</td>
<td>32</td>
<td>7</td>
<td>5</td>
<td>2.74</td>
<td>0.22</td>
<td>1.60</td>
<td>1.73</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2.25</td>
<td>0.23</td>
<td>0.78</td>
<td>0.31</td>
<td>33</td>
<td>8</td>
<td>1</td>
<td>-6.55</td>
<td>0.62</td>
<td>0.72</td>
<td>0.07</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>3.97</td>
<td>0.20</td>
<td>0.67</td>
<td>0.33</td>
<td>34</td>
<td>8</td>
<td>2</td>
<td>-4.02</td>
<td>0.37</td>
<td>0.86</td>
<td>0.14</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>7.32</td>
<td>0.19</td>
<td>0.86</td>
<td>0.87</td>
<td>35</td>
<td>8</td>
<td>3</td>
<td>-3.40</td>
<td>0.34</td>
<td>0.92</td>
<td>3.53</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>-6.55</td>
<td>0.62</td>
<td>1.95</td>
<td>9.90</td>
<td>36</td>
<td>8</td>
<td>4</td>
<td>-1.28</td>
<td>0.3</td>
<td>1.32</td>
<td>1.39</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>-3.51</td>
<td>0.34</td>
<td>0.97</td>
<td>0.24</td>
<td>37</td>
<td>8</td>
<td>5</td>
<td>3.61</td>
<td>0.20</td>
<td>1.30</td>
<td>1.99</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>3</td>
<td>-1.37</td>
<td>0.30</td>
<td>0.86</td>
<td>6.97</td>
<td>38</td>
<td>9</td>
<td>1</td>
<td>-1.82</td>
<td>0.3</td>
<td>0.72</td>
<td>0.14</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>4</td>
<td>1.04</td>
<td>0.25</td>
<td>1.51</td>
<td>8.53</td>
<td>39</td>
<td>9</td>
<td>2</td>
<td>-0.35</td>
<td>0.29</td>
<td>0.49</td>
<td>0.26</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>5</td>
<td>7.53</td>
<td>0.19</td>
<td>0.99</td>
<td>1.88</td>
<td>40</td>
<td>10</td>
<td>1</td>
<td>-2.37</td>
<td>0.31</td>
<td>0.88</td>
<td>2.32</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>1</td>
<td>0.03</td>
<td>0.28</td>
<td>1.12</td>
<td>0.85</td>
<td>41</td>
<td>10</td>
<td>2</td>
<td>-0.27</td>
<td>0.28</td>
<td>0.64</td>
<td>0.41</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>2</td>
<td>2.25</td>
<td>0.23</td>
<td>0.73</td>
<td>0.27</td>
<td>42</td>
<td>10</td>
<td>3</td>
<td>0.70</td>
<td>0.26</td>
<td>0.85</td>
<td>0.58</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>3</td>
<td>3.69</td>
<td>0.20</td>
<td>0.66</td>
<td>0.30</td>
<td>43</td>
<td>10</td>
<td>4</td>
<td>1.17</td>
<td>0.25</td>
<td>1.02</td>
<td>1.04</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>4</td>
<td>5.86</td>
<td>0.18</td>
<td>0.77</td>
<td>0.43</td>
<td>44</td>
<td>10</td>
<td>5</td>
<td>1.71</td>
<td>0.24</td>
<td>1.32</td>
<td>4.16</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>5</td>
<td>7.28</td>
<td>0.19</td>
<td>1.16</td>
<td>0.68</td>
<td>45</td>
<td>11</td>
<td>1</td>
<td>-2.47</td>
<td>0.31</td>
<td>0.84</td>
<td>0.23</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>1</td>
<td>-5.38</td>
<td>0.48</td>
<td>0.98</td>
<td>0.16</td>
<td>46</td>
<td>11</td>
<td>2</td>
<td>-1.28</td>
<td>0.3</td>
<td>1.00</td>
<td>0.25</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>2</td>
<td>-0.51</td>
<td>0.29</td>
<td>1.39</td>
<td>0.57</td>
<td>47</td>
<td>11</td>
<td>3</td>
<td>-0.51</td>
<td>0.29</td>
<td>1.27</td>
<td>1.44</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>3</td>
<td>0.64</td>
<td>0.26</td>
<td>1.12</td>
<td>0.45</td>
<td>48</td>
<td>11</td>
<td>4</td>
<td>0.42</td>
<td>0.27</td>
<td>0.76</td>
<td>0.23</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>4</td>
<td>6.18</td>
<td>0.18</td>
<td>1.30</td>
<td>3.89</td>
<td>49</td>
<td>11</td>
<td>5</td>
<td>1.59</td>
<td>0.24</td>
<td>0.93</td>
<td>1.19</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>1</td>
<td>-8.40</td>
<td>1.08</td>
<td>1.33</td>
<td>9.90</td>
<td>50</td>
<td>12</td>
<td>1</td>
<td>-2.28</td>
<td>0.31</td>
<td>0.86</td>
<td>0.89</td>
</tr>
<tr>
<td>21</td>
<td>5</td>
<td>2</td>
<td>-3.51</td>
<td>0.34</td>
<td>0.86</td>
<td>0.13</td>
<td>51</td>
<td>12</td>
<td>2</td>
<td>-1.28</td>
<td>0.3</td>
<td>0.81</td>
<td>0.21</td>
</tr>
<tr>
<td>22</td>
<td>5</td>
<td>3</td>
<td>2.09</td>
<td>0.23</td>
<td>1.24</td>
<td>0.98</td>
<td>52</td>
<td>12</td>
<td>3</td>
<td>1.17</td>
<td>0.25</td>
<td>0.76</td>
<td>0.23</td>
</tr>
<tr>
<td>23</td>
<td>5</td>
<td>4</td>
<td>1.17</td>
<td>0.25</td>
<td>1.59</td>
<td>1.12</td>
<td>53</td>
<td>12</td>
<td>4</td>
<td>2.14</td>
<td>0.23</td>
<td>0.66</td>
<td>0.26</td>
</tr>
<tr>
<td>24</td>
<td>5</td>
<td>5</td>
<td>4.87</td>
<td>0.19</td>
<td>0.98</td>
<td>0.95</td>
<td>54</td>
<td>12</td>
<td>5</td>
<td>3.36</td>
<td>0.21</td>
<td>0.92</td>
<td>0.69</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>1</td>
<td>-6.98</td>
<td>0.69</td>
<td>0.52</td>
<td>0.02</td>
<td>55</td>
<td>13</td>
<td>1</td>
<td>-0.85</td>
<td>0.29</td>
<td>0.94</td>
<td>0.24</td>
</tr>
<tr>
<td>26</td>
<td>6</td>
<td>2</td>
<td>-2.76</td>
<td>0.32</td>
<td>0.94</td>
<td>0.18</td>
<td>56</td>
<td>13</td>
<td>2</td>
<td>0.27</td>
<td>0.27</td>
<td>0.65</td>
<td>0.16</td>
</tr>
<tr>
<td>27</td>
<td>6</td>
<td>3</td>
<td>-0.94</td>
<td>0.29</td>
<td>0.51</td>
<td>0.13</td>
<td>57</td>
<td>13</td>
<td>3</td>
<td>1.59</td>
<td>0.24</td>
<td>0.69</td>
<td>0.25</td>
</tr>
<tr>
<td>28</td>
<td>7</td>
<td>1</td>
<td>-5.62</td>
<td>0.50</td>
<td>0.58</td>
<td>0.05</td>
<td>58</td>
<td>13</td>
<td>4</td>
<td>2.35</td>
<td>0.22</td>
<td>0.67</td>
<td>0.28</td>
</tr>
<tr>
<td>29</td>
<td>7</td>
<td>2</td>
<td>-4.30</td>
<td>0.38</td>
<td>0.99</td>
<td>0.25</td>
<td>59</td>
<td>13</td>
<td>5</td>
<td>3.77</td>
<td>0.20</td>
<td>0.83</td>
<td>0.53</td>
</tr>
<tr>
<td>30</td>
<td>7</td>
<td>3</td>
<td>-3.07</td>
<td>0.32</td>
<td>0.93</td>
<td>0.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Order of the dichotomous item appeared in the testlet – the lower the order, the easier the item is estimated to be.
Table 4.5. Step Parameter Estimates and Fit Statistics – PCM

<table>
<thead>
<tr>
<th>Testlet</th>
<th>Slope</th>
<th>Step Estimates</th>
<th>IN.MSQ</th>
<th>OUT.MSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1   2   3   4   5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1   1.00</td>
<td>1.12  1.13 0.76 2.17 4.92</td>
<td>1.04</td>
<td>2.01</td>
<td></td>
</tr>
<tr>
<td>2   1.00</td>
<td>-4.84 -1.92 -0.68 0.09 5.30</td>
<td>1.07</td>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td>3   1.00</td>
<td>0.36  1.56 1.91 4.06 4.40</td>
<td>0.85</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>4   1.00</td>
<td>-3.78 0.72 -0.60 4.06 1.17</td>
<td>1.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5   1.00</td>
<td>0.65 -2.25 0.86 1.27 2.72</td>
<td>1.11</td>
<td>1.08</td>
<td></td>
</tr>
<tr>
<td>6   1.00</td>
<td>-5.00 -1.09 -1.23 0.54 0.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7   1.00</td>
<td>-2.97 -2.44 0.15 -3.92 1.44</td>
<td>1.21</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>8   1.00</td>
<td>-4.19 -1.78 -2.47 -0.99 2.02</td>
<td>0.98</td>
<td>1.23</td>
<td></td>
</tr>
<tr>
<td>9   1.00</td>
<td>-0.22 -1.08 0.44 0.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10  1.00</td>
<td>-1.18 0.25 1.13 1.19 -0.49</td>
<td>1.36</td>
<td>4.82</td>
<td></td>
</tr>
<tr>
<td>11  1.00</td>
<td>-0.93 -0.38 -0.28 0.46 0.08</td>
<td>1.34</td>
<td>1.07</td>
<td></td>
</tr>
<tr>
<td>12  1.00</td>
<td>-0.46 -1.46 1.52 1.18 1.26</td>
<td>0.74</td>
<td>0.68</td>
<td></td>
</tr>
<tr>
<td>13  1.00</td>
<td>0.45 -0.17 1.68 1.02 1.62</td>
<td>0.72</td>
<td>0.61</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>1.00</td>
<td>-1.61 -0.69 0.23 0.96 2.33</td>
<td>0.97</td>
<td>1.22</td>
</tr>
</tbody>
</table>

Figure 4.4. Scatter Plot of Item Difficulty Estimates from the Dichotomous Rasch Model versus Item Step Estimates from the PCM

Item Difficulty Estimates from Dichotomous Rasch Model
Table 4.6. Thurstone Threshold Estimates - PCM

<table>
<thead>
<tr>
<th>Testlet</th>
<th>Slope</th>
<th>Thurstone Threshold Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1.00</td>
<td>0.45</td>
</tr>
<tr>
<td>2</td>
<td>1.00</td>
<td>-4.89</td>
</tr>
<tr>
<td>3</td>
<td>1.00</td>
<td>0.12</td>
</tr>
<tr>
<td>4</td>
<td>1.00</td>
<td>-3.79</td>
</tr>
<tr>
<td>5</td>
<td>1.00</td>
<td>-0.99</td>
</tr>
<tr>
<td>6</td>
<td>1.00</td>
<td>-5.00</td>
</tr>
<tr>
<td>7</td>
<td>1.00</td>
<td>-3.34</td>
</tr>
<tr>
<td>8</td>
<td>1.00</td>
<td>-4.28</td>
</tr>
<tr>
<td>9</td>
<td>1.00</td>
<td>-0.97</td>
</tr>
<tr>
<td>10</td>
<td>1.00</td>
<td>-1.37</td>
</tr>
<tr>
<td>11</td>
<td>1.00</td>
<td>-1.36</td>
</tr>
<tr>
<td>12</td>
<td>1.00</td>
<td>-1.28</td>
</tr>
<tr>
<td>13</td>
<td>1.00</td>
<td>-0.28</td>
</tr>
<tr>
<td>Average</td>
<td>1.00</td>
<td>-2.08</td>
</tr>
</tbody>
</table>

Figure 4.5. Scatter Plot of Item Difficulty Estimates from the Dichotomous Rasch Model versus Thurstone Threshold Estimates from the PCM

Item Difficulty Estimates from Dichotomous Rasch Model
Chapter Four: Results

1PL Model and GRM with Equal Slopes

The 1PL model is fitted to the dichotomous data, and the Graded Response Model with equal slopes is fitted to the polytomous model. Parameter estimation is conducted in MULTILOG and analysis results are illustrated in Table 4.7 and Table 4.8 respectively.

Unlike the simple Rasch model which fixed item slopes to be one, the 1PL model in MULTILOG estimates one common slope for all the items. The common slope is 4.17 after estimating the slope and then multiplying by the 1.7 scaling factor. Item difficulty estimates generated from the 1PL model range from -2.49 for Item 20 to 1.16 for Item 10. The $\chi^2$ statistic listed in the last column of Table 4.7 and Table 4.8 is the item fit statistic generated from the program PARSACLE after fitting the same model. Some $\chi^2$ statistics cannot be calculated due to too few degrees of freedom (i.e., too few responses in one of the response categories). All the un-estimated $\chi^2$ statistics happened on the first items in the testlets. Other than that, no misfit items are identified. In the GRM, the common slope estimate is 4.33 and the Thurstone threshold estimates range from -5.71 (Item 20) to 0.89 (Item 10). Note that the slope estimate of 4.33 in the GRM is different from the slope estimate of 4.17 in the 1PL model. This is because the two models were calibrated separately by MULTILOG. A common slope was estimated separately for each model.

Two polytomous items - Testlet 1 and Testlet 2 - are identified as misfitting according to the $\chi^2$ statistic. Since the $\chi^2$ statistic is highly influenced by sample size, identifying misfitting items cannot be solely determined by this single index. Other measures, such as visual examinations of the observed and fitted item characteristic curves, should be used to facilitate judgement of misfitting items.
Figure 4.6 plots the item difficulty estimates derived from the 1PL model against the Thurstone threshold estimates derived from the GRM. A strong, positive, linear pattern is presented with Item 20 as an outlier, which is similar to the scatter plot in the Rasch model comparison for Thurstone thresholds. The correlation between the dichotomous item difficulty estimates and the polytomous Thurstone threshold estimates is .91 when including every item, and .99 after excluding Item 20.

Figure 4.6 also reveals that most data points in the scatter plot, except for Item 20, lie on the reference line of $y=x$. To some degree, this implies that the two sets of item parameters are on the same metric. As illustrated in Chapter Three, MULTILOG identifies a model in each calibration by setting the mean and the standard deviation of person ability estimates at zero and one respectively. In this analysis, the two separate calibrations involved the same group of subjects, therefore, item parameter estimates generated from these two separate calibrations are actually linked and put on the same scale through the common person ability scale. This also holds true for the 2PL contrast in the next section.

As shown, when fixing item slopes to be equal, both dichotomous and polytomous models seem appropriate. The two models generate highly correlated item parameter estimates. Two misfitting items are identified in the polytomous data according to the $\chi^2$ statistic. However, $\chi^2$ statistics are used in this analysis only as an informative index to roughly evaluate possible misfit items. Using only a single $\chi^2$ statistic to judge a misfit item is inadequate. Since this is not a major concern in this study, no further analysis is conducted and the misfitting items will be kept in the data for all other
analyses.

The PCM and the GRM with equal slopes are both polytomous models with common item slopes. In the PCM, item slope parameters are fixed to one in WINSTEPS; while the GRM in Contrast 2 has a common item slope in MULTILOG. The Thurstone thresholds estimated from both models are compared in a scatter plot constructed between the two sets of estimates, presented in Figure 4.7. As shown, the two sets of Thurstone threshold estimates have some scaling differences due to the fact that they are estimated separately by two different software programs. Other than that, they are highly correlated, with Item 20 being identified as an outlier.
## Table 4.7. Item Parameter Estimates and Fit Statistics - 1PL Model

<table>
<thead>
<tr>
<th>Item</th>
<th>Testlet</th>
<th>Item order</th>
<th>Slope</th>
<th>Slope_se</th>
<th>Diff</th>
<th>Diff_se</th>
<th>( \chi^2 )</th>
<th>Item</th>
<th>Testlet</th>
<th>Item order</th>
<th>Slope</th>
<th>Slope_se</th>
<th>Diff</th>
<th>Diff_se</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4.17</td>
<td>0.10</td>
<td>-0.43</td>
<td>0.07</td>
<td>19.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4.17</td>
<td>0.10</td>
<td>-0.17</td>
<td>0.08</td>
<td>2.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4.17</td>
<td>0.10</td>
<td>0.00</td>
<td>0.08</td>
<td>2.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>4.17</td>
<td>0.10</td>
<td>0.39</td>
<td>0.07</td>
<td>9.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>4.17</td>
<td>0.10</td>
<td>1.12</td>
<td>0.06</td>
<td>14.65</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>4.17</td>
<td>0.10</td>
<td>-2.04</td>
<td>0.14</td>
<td>n/a†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>4.17</td>
<td>0.10</td>
<td>-1.32</td>
<td>0.10</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>3</td>
<td>4.17</td>
<td>0.10</td>
<td>-0.82</td>
<td>0.09</td>
<td>1.78</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>4</td>
<td>4.17</td>
<td>0.10</td>
<td>-0.27</td>
<td>0.07</td>
<td>3.69</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>5</td>
<td>4.17</td>
<td>0.10</td>
<td>1.16</td>
<td>0.06</td>
<td>21.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>1</td>
<td>4.17</td>
<td>0.10</td>
<td>-0.52</td>
<td>0.10</td>
<td>8.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>2</td>
<td>4.17</td>
<td>0.10</td>
<td>0.00</td>
<td>0.08</td>
<td>4.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>3</td>
<td>4.17</td>
<td>0.10</td>
<td>0.33</td>
<td>0.07</td>
<td>12.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>4</td>
<td>4.17</td>
<td>0.10</td>
<td>0.81</td>
<td>0.06</td>
<td>13.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>5</td>
<td>4.17</td>
<td>0.10</td>
<td>1.11</td>
<td>0.05</td>
<td>22.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>1</td>
<td>4.17</td>
<td>0.10</td>
<td>-1.76</td>
<td>0.15</td>
<td>n/a†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>2</td>
<td>4.17</td>
<td>0.10</td>
<td>-0.63</td>
<td>0.14</td>
<td>1.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>3</td>
<td>4.17</td>
<td>0.10</td>
<td>-0.36</td>
<td>0.09</td>
<td>2.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>4</td>
<td>4.17</td>
<td>0.10</td>
<td>0.88</td>
<td>0.05</td>
<td>39.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>1</td>
<td>4.17</td>
<td>0.10</td>
<td>-2.49</td>
<td>0.29</td>
<td>n/a†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>5</td>
<td>2</td>
<td>4.17</td>
<td>0.10</td>
<td>-1.32</td>
<td>0.11</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>5</td>
<td>3</td>
<td>4.17</td>
<td>0.10</td>
<td>-0.03</td>
<td>0.07</td>
<td>4.12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>5</td>
<td>4</td>
<td>4.17</td>
<td>0.10</td>
<td>-0.24</td>
<td>0.07</td>
<td>10.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>5</td>
<td>5</td>
<td>4.17</td>
<td>0.10</td>
<td>0.59</td>
<td>0.06</td>
<td>8.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>1</td>
<td>4.17</td>
<td>0.10</td>
<td>-2.14</td>
<td>0.27</td>
<td>n/a†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>6</td>
<td>2</td>
<td>4.17</td>
<td>0.10</td>
<td>-1.14</td>
<td>0.09</td>
<td>0.17</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>6</td>
<td>3</td>
<td>4.17</td>
<td>0.10</td>
<td>-0.72</td>
<td>0.11</td>
<td>6.15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>7</td>
<td>1</td>
<td>4.17</td>
<td>0.10</td>
<td>-1.82</td>
<td>0.20</td>
<td>n/a†</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>7</td>
<td>2</td>
<td>4.17</td>
<td>0.10</td>
<td>-1.51</td>
<td>0.12</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>7</td>
<td>3</td>
<td>4.17</td>
<td>0.10</td>
<td>-1.21</td>
<td>0.10</td>
<td>3.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p<.05 (after Bonferroni adjustment .05/59=.0008).
†Fit statistics cannot be calculated due to too few responses in the “unable” category.
Table 4.8. Thurstone Threshold Estimates and Fit Statistics –GRM with Equal Slopes

<table>
<thead>
<tr>
<th>Testlet</th>
<th>Slope</th>
<th>Thurstone Threshold Estimates</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4.33</td>
<td>-0.62 -0.43 -0.25 0.18 0.87</td>
<td>54.78 *</td>
</tr>
<tr>
<td>2</td>
<td>4.33</td>
<td>-2.15 -1.41 -0.88 -0.42 0.89</td>
<td>31.19</td>
</tr>
<tr>
<td>3</td>
<td>4.33</td>
<td>-0.67 -0.20 0.12 0.59 0.87</td>
<td>68.01 *</td>
</tr>
<tr>
<td>4</td>
<td>4.33</td>
<td>-1.80 -0.74 -0.51 0.62</td>
<td>33.03</td>
</tr>
<tr>
<td>5</td>
<td>4.33</td>
<td>-5.71 -1.36 -0.50 -0.16 0.34</td>
<td>22.90</td>
</tr>
<tr>
<td>6</td>
<td>4.33</td>
<td>-2.19 -1.24 -0.82</td>
<td>7.08</td>
</tr>
<tr>
<td>7</td>
<td>4.33</td>
<td>-1.83 -1.52 -1.25 -1.20 -0.08</td>
<td>17.89</td>
</tr>
<tr>
<td>8</td>
<td>4.33</td>
<td>-2.07 -1.54 -1.31 -0.85 0.10</td>
<td>15.44</td>
</tr>
<tr>
<td>9</td>
<td>4.33</td>
<td>-1.04 -0.70</td>
<td>7.02</td>
</tr>
<tr>
<td>10</td>
<td>4.33</td>
<td>-1.16 -0.68 -0.44 -0.35 -0.25</td>
<td>21.73</td>
</tr>
<tr>
<td>11</td>
<td>4.33</td>
<td>-1.14 -0.86 -0.73 -0.55 -0.29</td>
<td>9.80</td>
</tr>
<tr>
<td>12</td>
<td>4.33</td>
<td>-1.15 -0.93 -0.39 -0.19 0.08</td>
<td>21.56</td>
</tr>
<tr>
<td>13</td>
<td>4.33</td>
<td>-0.85 -0.62 -0.31 -0.15 0.15</td>
<td>18.71</td>
</tr>
<tr>
<td>Average</td>
<td>4.33</td>
<td>-1.72 -0.94 -0.61 -0.23 0.27</td>
<td></td>
</tr>
</tbody>
</table>

*p<.05 (after Bonferroni adjustment .05/13=.004)

Figure 4.6. Scatter Plot of Item Difficulty Estimates from the 1PL Model versus Thurston Threshold Estimates from the GRM with Equal Slopes

\( r = .91 \)

\( r = .99 \) (excluding Item 20)
Figure 4.7. Comparison of Thurstone Threshold Estimates between the PCM and the GRM with Equal Slopes

![Graph showing comparison of Thurstone Threshold Estimates between the PCM and the GRM with Equal Slopes]

**Thurstone Threshold Estimates from the PCM**

**2PL Model and GRM with Varying Slopes**

Table 4.9 and Table 4.10 present the analysis results for the 2PL model and for the GRM with varying slopes. Slope estimates of the 2PL model range from 0.27 for Item 20 to 10.87 for Item 25. Since previous literature has found that LID may inflate slope values, interpretations of the slope estimates from the 2PL model should be considered with caution. The $\chi^2$ statistic identified three misfit items, all of which are the highest-order items in their testlets. In the GRM, the average slope is 4.71 (ranging from 3.23 to 7.76). Testlet 1 and Testlet 9 are identified as misfitting according to the $\chi^2$ statistic. As shown earlier, when item slopes are fixed to be equal in the GRM with equal slopes, Testlets 1 and 3 were identified as misfitting. To further study the extent to which the expected models deviate from the observed data, other fit statistics, or a visual
examination of the expected item category characteristic curves and the observed category characteristic curves, can be conducted. However, a comprehensive fit analysis is not within the scope of this study.

Figure 4.8 plots item difficulty estimates generated from the 2PL model against the Thurstone threshold estimates from the GRM. A similar pattern is presented as in Figure 4.5 and Figure 4.6, except that two outliers are detected – Item 6 and Item 20, both of which have the lowest item slope estimates among all items. Correlation between the two sets of item parameter estimates is .85 when including every item, .97 after excluding Item 20, and .99 after excluding Items 20 and 6.

To compare whether allowing slopes to vary in the GRM provides a better overall fit, the difference of the -2 log likelihood (-2LL) between the GRM with equal slopes (-2LL = 2163.1, df = 60) and the GRM with varying slopes (-2LL = 2091.9, df = 72) is computed. The difference of 52.2, with degrees of freedom equal to 12, is significant ($p < 0.001$), which indicates that allowing slopes to vary has a better fit.

The IRT analyses conducted on the dichotomous and on the polytomous raw data revealed that for non-extreme item difficulty estimates, the location parameters between the two sets of models (item difficulty estimates derived from dichotomous models and Thurstone threshold estimates derived from corresponding polytomous models) were highly correlated. An obvious difference appeared on two extremely easy items and this difference became larger when item slopes were taken into account (2PL vs. GRM with varying slopes). Such difference was probably caused from the handling of extreme scores in the two models, during the estimation process in each software program.
| Item | Testlet | Item order | Slope | Slope_se | Diff | Diff_se | $\chi^2$ | Item | Testlet | Item order | Slope | Slope_se | Diff | Diff_se | $\chi^2$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2.44</td>
<td>0.40</td>
<td>-0.40</td>
<td>0.12</td>
<td>3.13</td>
<td>31</td>
<td>7</td>
<td>4</td>
<td>3.81</td>
<td>1.17</td>
<td>-1.12</td>
<td>0.13</td>
<td>3.61</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3.76</td>
<td>0.62</td>
<td>-0.01</td>
<td>0.09</td>
<td>4.17</td>
<td>32</td>
<td>7</td>
<td>5</td>
<td>2.49</td>
<td>0.51</td>
<td>0.20</td>
<td>0.1</td>
<td>4.77</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4.87</td>
<td>0.99</td>
<td>0.22</td>
<td>0.07</td>
<td>3.43</td>
<td>33</td>
<td>8</td>
<td>1</td>
<td>5.87</td>
<td>3.63</td>
<td>-1.95</td>
<td>0.24</td>
<td>n/a</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>7.13</td>
<td>1.43</td>
<td>0.63</td>
<td>0.05</td>
<td>10.03</td>
<td>34</td>
<td>8</td>
<td>2</td>
<td>4.61</td>
<td>2.11</td>
<td>-1.38</td>
<td>0.13</td>
<td>3.10</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>4.51</td>
<td>0.74</td>
<td>1.32</td>
<td>0.06</td>
<td>32.25</td>
<td>35</td>
<td>8</td>
<td>3</td>
<td>3.72</td>
<td>0.87</td>
<td>-1.26</td>
<td>0.15</td>
<td>1.88</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>1</td>
<td>1.12</td>
<td>0.68</td>
<td>-3.91</td>
<td>0.13</td>
<td>1.1</td>
<td>36</td>
<td>8</td>
<td>4</td>
<td>2.93</td>
<td>0.60</td>
<td>-0.75</td>
<td>0.13</td>
<td>0.78</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>2</td>
<td>3.68</td>
<td>1.17</td>
<td>-1.29</td>
<td>0.13</td>
<td>3.01</td>
<td>37</td>
<td>8</td>
<td>5</td>
<td>2.95</td>
<td>0.50</td>
<td>0.45</td>
<td>0.09</td>
<td>6.10</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>3</td>
<td>3.55</td>
<td>0.64</td>
<td>-0.74</td>
<td>0.13</td>
<td>3.20</td>
<td>38</td>
<td>9</td>
<td>1</td>
<td>5.74</td>
<td>2.55</td>
<td>-0.84</td>
<td>0.09</td>
<td>1.93</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>4</td>
<td>2.41</td>
<td>0.37</td>
<td>-0.24</td>
<td>0.14</td>
<td>3.61</td>
<td>39</td>
<td>9</td>
<td>2</td>
<td>6.17</td>
<td>1.44</td>
<td>-0.44</td>
<td>0.1</td>
<td>6.23</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>5</td>
<td>3.54</td>
<td>0.67</td>
<td>1.38</td>
<td>0.07</td>
<td>60.01*</td>
<td>40</td>
<td>10</td>
<td>1</td>
<td>3.82</td>
<td>0.67</td>
<td>-0.99</td>
<td>0.12</td>
<td>0.00</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>1</td>
<td>3.39</td>
<td>0.61</td>
<td>-0.41</td>
<td>0.11</td>
<td>5.12</td>
<td>41</td>
<td>10</td>
<td>2</td>
<td>5.29</td>
<td>0.97</td>
<td>-0.42</td>
<td>0.1</td>
<td>3.87</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>2</td>
<td>5.75</td>
<td>1.42</td>
<td>0.24</td>
<td>0.06</td>
<td>2.77</td>
<td>42</td>
<td>10</td>
<td>3</td>
<td>4.14</td>
<td>0.84</td>
<td>-0.19</td>
<td>0.12</td>
<td>4.48</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>3</td>
<td>9.48</td>
<td>2.5</td>
<td>0.57</td>
<td>0.04</td>
<td>6.08</td>
<td>43</td>
<td>10</td>
<td>4</td>
<td>3.77</td>
<td>0.74</td>
<td>-0.09</td>
<td>0.12</td>
<td>5.02</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>4</td>
<td>7.22</td>
<td>1.62</td>
<td>1.02</td>
<td>0.04</td>
<td>29.76</td>
<td>44</td>
<td>10</td>
<td>5</td>
<td>3.10</td>
<td>0.54</td>
<td>0.00</td>
<td>0.12</td>
<td>5.87</td>
</tr>
<tr>
<td>15</td>
<td>3</td>
<td>5</td>
<td>4.23</td>
<td>0.68</td>
<td>1.32</td>
<td>0.06</td>
<td>61.87*</td>
<td>45</td>
<td>11</td>
<td>1</td>
<td>4.56</td>
<td>1.36</td>
<td>-1.00</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>1</td>
<td>3.73</td>
<td>1.76</td>
<td>-1.78</td>
<td>0.28</td>
<td>n/a†</td>
<td>46</td>
<td>11</td>
<td>2</td>
<td>4.15</td>
<td>1.38</td>
<td>-0.71</td>
<td>0.1</td>
<td>1.79</td>
</tr>
<tr>
<td>17</td>
<td>4</td>
<td>2</td>
<td>3.12</td>
<td>0.66</td>
<td>-0.55</td>
<td>0.11</td>
<td>2.07</td>
<td>47</td>
<td>11</td>
<td>3</td>
<td>3.33</td>
<td>0.63</td>
<td>-0.53</td>
<td>0.11</td>
<td>0.88</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>3</td>
<td>3.94</td>
<td>0.73</td>
<td>-0.22</td>
<td>0.09</td>
<td>1.36</td>
<td>48</td>
<td>11</td>
<td>4</td>
<td>4.96</td>
<td>1.36</td>
<td>-0.24</td>
<td>0.09</td>
<td>5.02</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
<td>4</td>
<td>2.71</td>
<td>0.45</td>
<td>1.08</td>
<td>0.08</td>
<td>39.55*</td>
<td>49</td>
<td>11</td>
<td>5</td>
<td>3.89</td>
<td>0.54</td>
<td>0.02</td>
<td>0.1</td>
<td>7.14</td>
</tr>
<tr>
<td>20</td>
<td>5</td>
<td>1</td>
<td>0.27</td>
<td>N/A</td>
<td>-21.17</td>
<td>N/A</td>
<td>n/a†</td>
<td>50</td>
<td>12</td>
<td>1</td>
<td>3.65</td>
<td>0.82</td>
<td>-0.97</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td>21</td>
<td>5</td>
<td>2</td>
<td>4.87</td>
<td>2.01</td>
<td>-1.24</td>
<td>0.12</td>
<td>0.75</td>
<td>51</td>
<td>12</td>
<td>2</td>
<td>4.46</td>
<td>1.30</td>
<td>-0.70</td>
<td>0.1</td>
<td>1.18</td>
</tr>
<tr>
<td>22</td>
<td>5</td>
<td>3</td>
<td>3.19</td>
<td>0.47</td>
<td>0.10</td>
<td>0.09</td>
<td>4.39</td>
<td>52</td>
<td>12</td>
<td>3</td>
<td>5.56</td>
<td>1.25</td>
<td>-0.03</td>
<td>0.08</td>
<td>2.31</td>
</tr>
<tr>
<td>23</td>
<td>5</td>
<td>4</td>
<td>2.77</td>
<td>0.51</td>
<td>-0.16</td>
<td>0.18</td>
<td>0.69</td>
<td>53</td>
<td>12</td>
<td>4</td>
<td>6.18</td>
<td>1.46</td>
<td>0.22</td>
<td>0.07</td>
<td>1.49</td>
</tr>
<tr>
<td>24</td>
<td>5</td>
<td>5</td>
<td>3.85</td>
<td>1.64</td>
<td>0.79</td>
<td>0.06</td>
<td>10.54</td>
<td>54</td>
<td>12</td>
<td>5</td>
<td>4.14</td>
<td>0.68</td>
<td>0.45</td>
<td>0.08</td>
<td>10.70</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>1</td>
<td>10.87</td>
<td>N/A</td>
<td>-1.99</td>
<td>0.2</td>
<td>n/a†</td>
<td>55</td>
<td>13</td>
<td>1</td>
<td>4.06</td>
<td>1.04</td>
<td>-0.59</td>
<td>0.1</td>
<td>3.83</td>
</tr>
<tr>
<td>26</td>
<td>6</td>
<td>2</td>
<td>4.13</td>
<td>1.83</td>
<td>-1.08</td>
<td>0.12</td>
<td>2.58</td>
<td>56</td>
<td>13</td>
<td>2</td>
<td>5.76</td>
<td>1.87</td>
<td>-0.27</td>
<td>0.08</td>
<td>2.66</td>
</tr>
<tr>
<td>27</td>
<td>6</td>
<td>3</td>
<td>6.81</td>
<td>2.02</td>
<td>-0.60</td>
<td>0.09</td>
<td>5.18</td>
<td>57</td>
<td>13</td>
<td>3</td>
<td>6.53</td>
<td>1.59</td>
<td>0.09</td>
<td>0.07</td>
<td>1.76</td>
</tr>
<tr>
<td>28</td>
<td>7</td>
<td>1</td>
<td>8.90</td>
<td>N/A</td>
<td>-1.65</td>
<td>0.36</td>
<td>n/a†</td>
<td>58</td>
<td>13</td>
<td>4</td>
<td>6.96</td>
<td>1.36</td>
<td>0.28</td>
<td>0.06</td>
<td>1.96</td>
</tr>
<tr>
<td>29</td>
<td>7</td>
<td>2</td>
<td>3.92</td>
<td>1.50</td>
<td>-1.48</td>
<td>0.19</td>
<td>0.87</td>
<td>59</td>
<td>13</td>
<td>5</td>
<td>4.80</td>
<td>1.08</td>
<td>0.56</td>
<td>0.07</td>
<td>11.71</td>
</tr>
</tbody>
</table>

*p<.05 (after Bonferroni adjustment .05/59=.0008)
†Fit statistics cannot be calculated due to too few responses in the category
Table 4.10. Thurstone Threshold Estimates and Fit Statistics – GRM with Varying Slopes

<table>
<thead>
<tr>
<th>Testlet</th>
<th>Slope</th>
<th>Thurstone Threshold Estimates</th>
<th>$\chi^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>4.00</td>
<td>-0.57</td>
<td>-0.36</td>
</tr>
<tr>
<td>2</td>
<td>3.27</td>
<td>-2.32</td>
<td>-1.46</td>
</tr>
<tr>
<td>3</td>
<td>4.18</td>
<td>-0.61</td>
<td>-0.12</td>
</tr>
<tr>
<td>4</td>
<td>3.36</td>
<td>-1.88</td>
<td>-0.73</td>
</tr>
<tr>
<td>5</td>
<td>3.54</td>
<td>-5.59</td>
<td>-1.37</td>
</tr>
<tr>
<td>6</td>
<td>7.76</td>
<td>-2.03</td>
<td>-1.11</td>
</tr>
<tr>
<td>7</td>
<td>3.23</td>
<td>-1.96</td>
<td>-1.59</td>
</tr>
<tr>
<td>8</td>
<td>3.79</td>
<td>-2.14</td>
<td>-1.55</td>
</tr>
<tr>
<td>9</td>
<td>7.66</td>
<td>-0.96</td>
<td>-0.63</td>
</tr>
<tr>
<td>10</td>
<td>3.91</td>
<td>-1.14</td>
<td>-0.63</td>
</tr>
<tr>
<td>11</td>
<td>4.30</td>
<td>-1.10</td>
<td>-0.82</td>
</tr>
<tr>
<td>12</td>
<td>6.07</td>
<td>-1.06</td>
<td>-0.86</td>
</tr>
<tr>
<td>13</td>
<td>6.16</td>
<td>-0.76</td>
<td>-0.53</td>
</tr>
<tr>
<td>Average</td>
<td>4.71</td>
<td>-1.70</td>
<td>-0.90</td>
</tr>
</tbody>
</table>

* $p<.05$ (after Bonferroni adjustment $0.05/13=0.004$)

Figure 4.8. Scatter Plot of Item Difficulty Estimates from the 2PL Model versus Thurstone Threshold Estimates from the GRM with Varying Slopes

Item Difficulty Estimates from 2PL Model

$r = .97$ (excluding Item 20)

$r = .99$ (excluding Items 20 and 6)
Analysis 2: Assessment of Local Item Dependence

Yen’s $Q_3$ statistic is implemented to evaluate LID among items within the same testlet in the dichotomous data. Since $Q_3$ is derived after fitting a specific IRT model, in order to control the influence of fitting different models on the calculation, the $Q_3$ statistic in this study is computed twice – based on the Rasch Model and on the 2PL model.

Table 4.11 and Figure 4.9 present the analysis results for $Q_3$’s derived from the Rasch model; Table 4.12 and Figure 4.10 present the same analysis for $Q_3$’s derived from the 2PL model. Since the two models generated highly similar patterns, only the analysis based on the Rasch model is interpreted.

Table 4.11 lists the within-content $Q_3$’s for the 13 testlets, as well as the deviation of the within-$Q_3$ from the expected value of -0.017 (i.e., -1/58). When items are locally independent, the within-$Q_3$ should be around the expected value, and the deviation values should be near zero. Results from a Parallel analysis on the simulated 1PL data are presented in the last two columns of Table 4.11. As shown, in the real data, the average within-$Q_3$ is 0.167 and the average deviation from the expected value is 0.184, while for the simulated data the average within-$Q_3$ is -0.019 and the average deviation is -0.002. The within-$Q_3$ values in the real data are all positive and much higher than the expected value. Specifically, all of the observed within-$Q_3$ values are at least two standard deviations above the average simulated within-$Q_3$. In general, among the 13 testlets, three
testlets have $Q_3$ values larger than 0.2, eight testlets between 0.1 and 0.2, and two testlets smaller than 0.1 but larger than zero. Testlet 10 has the highest within-$Q_3$ of 0.324, and Testlet 6 has the smallest within-$Q_3$ of 0.065.

Based on Reese’s (1995) classification on low-, medium-, and high-LID levels using $Q_3$ values (low-LID with $Q_3$ around 0.01, medium-LID with $Q_3$ around 0.02-0.05, and high-LID with $Q_3$ above 0.3), all 13 testlets have exhibited medium-high LID. The observed within-$Q_3$ values in this study are also much higher than findings from previous empirical studies. For example, Zenisky, Hambleton, & Sireci (2002) detected low to medium LID in the Medical College Admission Test (MCAT). The average within-passage (testlet) $Q_3$ values across three sections of the MCAT ranged from 0.01 to 0.03, with an average deviation from the expected value ranging from 0.027 to 0.051.

Figure 4.9 presents the overlay of between- and within-$Q_3$ distributions, for the real data and for simulated data. The y-axis in the histogram represents relative frequency (percentage of $Q_3$ values) among the 78 between-$Q_3$ values and among the 13 within-$Q_3$ values. These 78 ($78=13*12/2$) between-$Q_3$ values indicate the associations across the 13 testlets after taking into account the underlying dimension, and thus, suggest the degree of dependence or independence among the 13 testlets should the individual items be formed into testlets.

As shown from Figure 4.9A, in the real data, most of the between-$Q_3$’s are within
-0.1 and +0.1 and form a roughly normal distribution around the expected value of -0.017; whereas, most within-$Q_3$’s are above 0.1. In the simulated data, the distribution of the within-$Q_3$’s almost overlaps with the distribution of the between-$Q_3$’s, and the variance of both distributions are smaller than that of the real data.

The Kruskal-Wallis test also reveals that the within-$Q_3$ distribution is significantly different from the between-$Q_3$ distribution in the real data. Specifically, the average rank for the 13 within-$Q_3$’s is 83, larger than the average rank of 40 for the 78 between-$Q_3$’s in the real data, whereas, no statistical difference is detected in the simulated data.

This analysis assesses the existence of LID using Yen’s $Q_3$ statistic. Examinations of the actual $Q_3$ values and of the within- and between-$Q_3$ distributions suggest that items within the same content exhibit a moderate to high degree of LID.
### Table 4.11. Within-Content $Q_3$ values, derived from the Rasch Model

<table>
<thead>
<tr>
<th>Testlet</th>
<th>1PL_Real</th>
<th>Deviation from Expected*</th>
<th>1PL_Simu Mean (SD)a</th>
<th>Deviation from Expected* Mean (SD)a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testlet 1</td>
<td>0.190</td>
<td>0.207</td>
<td>-0.027 (0.019)</td>
<td>-0.010 (0.006)</td>
</tr>
<tr>
<td>Testlet 2</td>
<td>0.098</td>
<td>0.115</td>
<td>-0.010 (0.023)</td>
<td>0.007 (0.005)</td>
</tr>
<tr>
<td>Testlet 3</td>
<td>0.111</td>
<td>0.128</td>
<td>-0.027 (0.017)</td>
<td>-0.010 (0.026)</td>
</tr>
<tr>
<td>Testlet 4</td>
<td>0.128</td>
<td>0.145</td>
<td>-0.015 (0.024)</td>
<td>0.002 (0.003)</td>
</tr>
<tr>
<td>Testlet 5</td>
<td>0.101</td>
<td>0.118</td>
<td>-0.014 (0.012)</td>
<td>0.003 (0.010)</td>
</tr>
<tr>
<td>Testlet 6</td>
<td>0.065</td>
<td>0.082</td>
<td>0.004 (0.056)</td>
<td>0.021 (0.056)</td>
</tr>
<tr>
<td>Testlet 7</td>
<td>0.249</td>
<td>0.266</td>
<td>-0.030 (0.022)</td>
<td>-0.013 (0.008)</td>
</tr>
<tr>
<td>Testlet 8</td>
<td>0.120</td>
<td>0.137</td>
<td>-0.015 (0.035)</td>
<td>0.002 (0.001)</td>
</tr>
<tr>
<td>Testlet 9</td>
<td>0.170</td>
<td>0.187</td>
<td>-0.022 (0.122)</td>
<td>-0.005 (0.154)</td>
</tr>
<tr>
<td>Testlet 10</td>
<td>0.324</td>
<td>0.341</td>
<td>-0.021 (0.025)</td>
<td>-0.004 (0.005)</td>
</tr>
<tr>
<td>Testlet 11</td>
<td>0.294</td>
<td>0.311</td>
<td>-0.031 (0.025)</td>
<td>-0.014 (0.039)</td>
</tr>
<tr>
<td>Testlet 12</td>
<td>0.140</td>
<td>0.157</td>
<td>-0.012 (0.027)</td>
<td>0.005 (0.009)</td>
</tr>
<tr>
<td>Testlet 13</td>
<td>0.179</td>
<td>0.196</td>
<td>-0.025 (0.023)</td>
<td>-0.008 (0.028)</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>0.167</strong></td>
<td><strong>0.184</strong></td>
<td><strong>-0.019 (0.033)</strong></td>
<td><strong>-0.002 (0.027)</strong></td>
</tr>
</tbody>
</table>

*Expected = $-1/59 = -0.017$; aMean and Standard Deviation (SD) across the 25 simulated data

### Figure 4.9. Distribution of Within- and Between-$Q_3$, derived from the Rasch Model

#### A. Real Data

#### B. Simulated Data

<table>
<thead>
<tr>
<th></th>
<th>Real Data Mean (SD)</th>
<th>Simulated Data Mean (SD)a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-Content $Q_3$</td>
<td>0.167 (0.079)</td>
<td>-0.019 (0.012)</td>
</tr>
<tr>
<td>Between-Content $Q_3$</td>
<td>-0.026 (0.056)</td>
<td>-0.016 (&lt;0.001)</td>
</tr>
<tr>
<td>Kruskal-Wallis Test</td>
<td>30.13 ($p=0.001$)</td>
<td>1.792 ($p=0.414$)</td>
</tr>
</tbody>
</table>

aAcross 25 simulated datasets
Table 4.12. Within-Content $Q_3$ values, derived from the 2PL Model

<table>
<thead>
<tr>
<th>Testlet</th>
<th>2PL_Real</th>
<th>Deviation from Expected*</th>
<th>2PL_Simu Mean (SD)a</th>
<th>Deviation from Expected* Mean (SD)a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testlet 1</td>
<td>0.193</td>
<td>0.199</td>
<td>-0.029 (0.023)</td>
<td>-0.012 (0.023)</td>
</tr>
<tr>
<td>Testlet 2</td>
<td>0.108</td>
<td>0.108</td>
<td>-0.003 (0.018)</td>
<td>0.014 (0.018)</td>
</tr>
<tr>
<td>Testlet 3</td>
<td>0.081</td>
<td>0.116</td>
<td>-0.038 (0.017)</td>
<td>-0.021 (0.017)</td>
</tr>
<tr>
<td>Testlet 4</td>
<td>0.134</td>
<td>0.143</td>
<td>-0.006 (0.020)</td>
<td>0.011 (0.020)</td>
</tr>
<tr>
<td>Testlet 5</td>
<td>0.110</td>
<td>0.114</td>
<td>-0.008 (0.011)</td>
<td>0.009 (0.011)</td>
</tr>
<tr>
<td>Testlet 6</td>
<td>0.078</td>
<td>0.096</td>
<td>-0.002 (0.041)</td>
<td>0.015 (0.041)</td>
</tr>
<tr>
<td>Testlet 7</td>
<td>0.225</td>
<td>0.263</td>
<td>-0.025 (0.034)</td>
<td>-0.008 (0.034)</td>
</tr>
<tr>
<td>Testlet 8</td>
<td>0.118</td>
<td>0.135</td>
<td>-0.010 (0.028)</td>
<td>0.007 (0.028)</td>
</tr>
<tr>
<td>Testlet 9</td>
<td>0.116</td>
<td>0.149</td>
<td>-0.049 (0.107)</td>
<td>-0.032 (0.107)</td>
</tr>
<tr>
<td>Testlet 10</td>
<td>0.319</td>
<td>0.363</td>
<td>-0.024 (0.029)</td>
<td>-0.007 (0.029)</td>
</tr>
<tr>
<td>Testlet 11</td>
<td>0.295</td>
<td>0.324</td>
<td>-0.018 (0.031)</td>
<td>-0.001 (0.031)</td>
</tr>
<tr>
<td>Testlet 12</td>
<td>0.127</td>
<td>0.166</td>
<td>-0.005 (0.019)</td>
<td>0.012 (0.019)</td>
</tr>
<tr>
<td>Testlet 13</td>
<td>0.138</td>
<td>0.208</td>
<td>-0.029 (0.029)</td>
<td>-0.012 (0.029)</td>
</tr>
<tr>
<td>average</td>
<td>0.157</td>
<td>0.183</td>
<td>-0.019 (0.031)</td>
<td>-0.002 (0.031)</td>
</tr>
</tbody>
</table>

*Expected = -1/59 = -0.017; aMean and Standard Deviation (SD) across the 25 simulated data

Figure 4.10. Distribution of Within- and Between-$Q_3$, derived from the 2PL Model

A. Real Data

B. Simulated Data

<table>
<thead>
<tr>
<th></th>
<th>Real Data</th>
<th>Simulated Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-Content $Q_3$</td>
<td>0.157 (0.078)</td>
<td>-0.019 (0.011)</td>
</tr>
<tr>
<td>Between-Content $Q_3$</td>
<td>-0.026 (0.053)</td>
<td>-0.016 (0.001)</td>
</tr>
<tr>
<td>Kruskal-Wallis Test ($p$)</td>
<td>29.76 ($p&lt;0.001$)</td>
<td>1.353 ($p=0.430$)</td>
</tr>
</tbody>
</table>

*aAcross 25 simulated datasets
Analysis 3: Assessment of Dimensionality

Exploratory factor analyses (EFA) were conducted on the dichotomous data and on the polytomous data using Mplus. Analysis results for the dichotomous data are presented in Table 4.13 and Figure 4.11, and analysis results for the polytomous data are presented in Table 4.14 and Figure 4.12.

EFA on the Dichotomous Data

Table 4.13 lists eigenvalues from EFA on the dichotomous data. Figure 4.11 presents the overlay of scree plots for the real data and the simulated data. A 95% confidence interval is constructed around the average eigenvalue across 25 simulations for each of the 59 factors. Since a tetrachoric correlation matrix is used as an input matrix, negative values occur for the last few factors.

In the real data, eigenvalues for the first few factors and last few factors fall outside of the 95% confidence intervals, but the deviation is not very big and barely noticeable in the scree plot as shown in Figure 4.11. Basically, the scree plot exhibits an obvious one dominant factor in the real data.

EFA on the Polytomous Data

Eigenvalues from EFA on the polytomous data are presented in Table 4.14. The overlay of scree plots for the real data and the simulated data are presented in Figure 4.12. As shown, the eigenvalue of the first factor in the real data is slightly lower than the
lower bound of the 95% confidence interval constructed from the simulated data, while
eigenvalues of all the remaining factors fall inside this interval. Although the raw data
show a stronger elbow than the simulated data, the eigenvalue for the second factor in the
real data falls inside the 95% confidence interval and the two scree plots cross between
factor 1 and factor 2, indicating that an essential unidimensional structure can be claimed
for the polytomous raw data.

Although LID caused by the deterministic response pattern has been detected
using $Q_3$, the unidimensionality assumption basically holds for both the dichotomous and
the polytomous data. Unlike previous literature which states the equivalence between the
dimensionality assumption and the local item independence assumption, results from this
study suggest that although LID is detected in the real data, the unidimensionality
assumption still holds. One explanation is that LID detected in this study is caused by an
intrinsic data structure rather than obvious extra dimensions. In other words, local
dependence is more likely caused by the Guttman scaling response pattern in each testlet
than by a shared secondary factor among items in the same testlet.

Identifying the fundamental causes of LID would assist in selecting an
appropriate method to accommodate LID. As discussed in Chapter Two, two major
approaches are available to handle LID – the score-based method, which has been
adopted in this study, and the item-based method. Recall that almost all item-based
models are grounded on the theoretical framework that LID is caused by the existence of extra dimension(s) and thus, these models tackle the LID problem by modeling extra dimensions either explicitly as in the bi-factor model, or implicitly as in the Testlet Response Model. However, if LID is more likely caused by an intrinsic data structure, or the extra dimensions are not as salient as they are hypothesized to be, the item-based approach might not be an optimal choice. Issues in this aspect deserve further investigating.
Table 4.13. Eigenvalues from Factor Analysis on the Dichotomous Data

<table>
<thead>
<tr>
<th>Factor</th>
<th>Eigenvalue</th>
<th>Eigenvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Real_data</td>
<td>Simu_data</td>
</tr>
<tr>
<td></td>
<td>Mean (SD)a</td>
<td>Mean (SD)a</td>
</tr>
<tr>
<td>1</td>
<td>52.380</td>
<td>53.530</td>
</tr>
<tr>
<td>2</td>
<td>4.799</td>
<td>1.971</td>
</tr>
<tr>
<td>3</td>
<td>2.586</td>
<td>1.107</td>
</tr>
<tr>
<td>4</td>
<td>1.482</td>
<td>0.852</td>
</tr>
<tr>
<td>5</td>
<td>1.111</td>
<td>0.731</td>
</tr>
<tr>
<td>6</td>
<td>0.975</td>
<td>0.661</td>
</tr>
<tr>
<td>7</td>
<td>0.928</td>
<td>0.603</td>
</tr>
<tr>
<td>8</td>
<td>0.743</td>
<td>0.566</td>
</tr>
<tr>
<td>9</td>
<td>0.670</td>
<td>0.520</td>
</tr>
<tr>
<td>10</td>
<td>0.539</td>
<td>0.481</td>
</tr>
<tr>
<td>11</td>
<td>0.519</td>
<td>0.451</td>
</tr>
<tr>
<td>12</td>
<td>0.460</td>
<td>0.420</td>
</tr>
<tr>
<td>13</td>
<td>0.403</td>
<td>0.387</td>
</tr>
<tr>
<td>14</td>
<td>0.382</td>
<td>0.359</td>
</tr>
<tr>
<td>15</td>
<td>0.323</td>
<td>0.334</td>
</tr>
<tr>
<td>16</td>
<td>0.277</td>
<td>0.312</td>
</tr>
<tr>
<td>17</td>
<td>0.267</td>
<td>0.288</td>
</tr>
<tr>
<td>18</td>
<td>0.210</td>
<td>0.265</td>
</tr>
<tr>
<td>19</td>
<td>0.184</td>
<td>0.244</td>
</tr>
<tr>
<td>20</td>
<td>0.157</td>
<td>0.225</td>
</tr>
<tr>
<td>21</td>
<td>0.144</td>
<td>0.209</td>
</tr>
<tr>
<td>22</td>
<td>0.122</td>
<td>0.194</td>
</tr>
<tr>
<td>23</td>
<td>0.099</td>
<td>0.175</td>
</tr>
<tr>
<td>24</td>
<td>0.089</td>
<td>0.157</td>
</tr>
<tr>
<td>25</td>
<td>0.077</td>
<td>0.140</td>
</tr>
<tr>
<td>26</td>
<td>0.058</td>
<td>0.125</td>
</tr>
<tr>
<td>27</td>
<td>0.052</td>
<td>0.110</td>
</tr>
<tr>
<td>28</td>
<td>0.042</td>
<td>0.094</td>
</tr>
<tr>
<td>29</td>
<td>0.030</td>
<td>0.081</td>
</tr>
<tr>
<td>30</td>
<td>0.018</td>
<td>0.067</td>
</tr>
<tr>
<td>31</td>
<td>0.011</td>
<td>0.056</td>
</tr>
<tr>
<td>32</td>
<td>0.009</td>
<td>0.044</td>
</tr>
<tr>
<td>33</td>
<td>0.002</td>
<td>0.035</td>
</tr>
<tr>
<td>34</td>
<td>-0.011</td>
<td>0.024</td>
</tr>
<tr>
<td>35</td>
<td>-0.015</td>
<td>0.010</td>
</tr>
<tr>
<td>36</td>
<td>-0.023</td>
<td>0.003</td>
</tr>
<tr>
<td>37</td>
<td>-0.027</td>
<td>-0.009</td>
</tr>
<tr>
<td>38</td>
<td>-0.035</td>
<td>-0.020</td>
</tr>
<tr>
<td>39</td>
<td>-0.052</td>
<td>-0.032</td>
</tr>
<tr>
<td>40</td>
<td>-0.059</td>
<td>-0.044</td>
</tr>
<tr>
<td>41</td>
<td>-0.073</td>
<td>-0.055</td>
</tr>
<tr>
<td>42</td>
<td>-0.079</td>
<td>-0.068</td>
</tr>
<tr>
<td>43</td>
<td>-0.087</td>
<td>-0.083</td>
</tr>
<tr>
<td>44</td>
<td>-0.112</td>
<td>-0.101</td>
</tr>
<tr>
<td>45</td>
<td>-0.118</td>
<td>-0.117</td>
</tr>
<tr>
<td>46</td>
<td>-0.129</td>
<td>-0.132</td>
</tr>
<tr>
<td>47</td>
<td>-0.156</td>
<td>-0.150</td>
</tr>
<tr>
<td>48</td>
<td>-0.170</td>
<td>-0.172</td>
</tr>
<tr>
<td>49</td>
<td>-0.191</td>
<td>-0.195</td>
</tr>
<tr>
<td>50</td>
<td>-0.229</td>
<td>-0.216</td>
</tr>
<tr>
<td>51</td>
<td>-0.258</td>
<td>-0.239</td>
</tr>
<tr>
<td>52</td>
<td>-0.292</td>
<td>-0.265</td>
</tr>
<tr>
<td>53</td>
<td>-0.399</td>
<td>-0.301</td>
</tr>
<tr>
<td>54</td>
<td>-0.419</td>
<td>-0.338</td>
</tr>
<tr>
<td>55</td>
<td>-0.476</td>
<td>-0.390</td>
</tr>
<tr>
<td>56</td>
<td>-0.648</td>
<td>-0.481</td>
</tr>
<tr>
<td>57</td>
<td>-1.487</td>
<td>-0.649</td>
</tr>
<tr>
<td>58</td>
<td>-2.633</td>
<td>-1.135</td>
</tr>
<tr>
<td>59</td>
<td>-2.968</td>
<td>-1.978</td>
</tr>
</tbody>
</table>

*aAcross 25 simulated datasets*
Figure 4.11. Scree Plot of the Dichotomous Data

- - - - Simulated Data  - - Real Data
Table 4.14. Eigenvalues from Factor Analysis on the Polytomous Data

<table>
<thead>
<tr>
<th>Factor</th>
<th>Real_data</th>
<th>Simu_data Mean (SD)*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Eigenvalue</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>11.285</td>
<td>11.864 (0.186)</td>
</tr>
<tr>
<td>2</td>
<td>1.990</td>
<td>1.023 (0.976)</td>
</tr>
<tr>
<td>3</td>
<td>0.307</td>
<td>0.206 (0.239)</td>
</tr>
<tr>
<td>4</td>
<td>0.245</td>
<td>0.173 (0.013)</td>
</tr>
<tr>
<td>5</td>
<td>0.229</td>
<td>0.141 (0.011)</td>
</tr>
<tr>
<td>6</td>
<td>0.153</td>
<td>0.114 (0.008)</td>
</tr>
<tr>
<td>7</td>
<td>0.135</td>
<td>0.084 (0.007)</td>
</tr>
<tr>
<td>8</td>
<td>0.125</td>
<td>0.059 (0.007)</td>
</tr>
<tr>
<td>9</td>
<td>0.086</td>
<td>0.044 (0.007)</td>
</tr>
<tr>
<td>10</td>
<td>0.051</td>
<td>0.033 (0.006)</td>
</tr>
<tr>
<td>11</td>
<td>-0.010</td>
<td>0.021 (0.009)</td>
</tr>
<tr>
<td>12</td>
<td>-0.035</td>
<td>-0.013 (0.191)</td>
</tr>
<tr>
<td>13</td>
<td>-1.561</td>
<td>-0.749 (0.830)</td>
</tr>
</tbody>
</table>

*Across 25 simulated datasets

Figure 4.12. Scree Plot of the Polytomous Data
Analysis 4: Internal Reliability

Figure 4.13 presents the alpha coefficients obtained from the real and simulated dichotomous and polytomous data. For the simulated data, a 95% confidence interval is constructed around the average alpha values of the 25 simulations. As shown for the real data, the alpha coefficient drops from .980 in the dichotomous scoring scheme to .965 in the polytomous scoring scheme; and the average alpha value across the 25 simulations drops from .982 to .975.

The observed alpha value in the real dichotomous data falls in the 95% interval around the average simulated alpha, while the observed alpha in the polytomous data is below the lower bound of the 95% interval. Specifically, there is a 1.7\% [(0.982-0.965)/0.982] drop in the alpha coefficient when changing the scoring scheme in the real data, while there is only a 0.7\% [(0.982 - 0.975)/0.982] drop in the simulated data, which to some degree indicates a potential local dependency. However, although the drop of alpha value in the real data seems dramatic when compared with the simulated data, the practical significance is not dramatic. The alpha coefficient in the polytomous data is still very high at 0.965. Besides, the drop of 1.8\% in reliability in the real data is not large at all when compared with similar empirical studies in the literature. Sireci, Thissen, & Wainer (1991) detected a reliability drop by about 10\% when individual dichotomous items were combined into testlets; Zenisky et al. (2002) detected reliability drops that
ranged from 4.6% to 7.1% for the three sections of the MCAT; and Keller, Swaminathan, & Sireci (2003) found a reliability drop of about 5.4%.

**Figure 4.13. Alpha Coefficients for the Dichotomous and the Polytomous Data**
Analysis 5: Test Information

The test information function provides a means for evaluating test precision levels and can provide an overview of the most precisely and the least precisely estimated regions on the ability spectrum measured by an instrument. The analysis in this section examines the difference in test information generated from the two scoring methods in the presence of local dependence.

Since test information is directly computed from item parameter estimates (which are derived separately from each calibration), item parameter estimates generated separately from the dichotomous and polytomous models have to be on the same metric in order to compare the two sets of test information values. This is not a problem for Contrast 2 and Contrast 3. Theoretically, since MULTILOG sets the mean and the standard deviation of the ability estimates at zero and one in each calibration, and since the same group of subjects is involved in each calibration, the two sets of item parameter estimates are linked through the common ability scale and thus, test information can be directly compared. Empirically, the equivalence of item parameter estimates between the two sets of models is not only confirmed by Thompson and Pommerich (1996), but also implied by comparing item parameter estimates in the exploratory IRT analyses of this dissertation. As shown from the scatter plots in Figure 4.6 and Figure 4.8, comparing item difficulty estimates from the dichotomous model with Thurstone threshold estimates from
the GRM, the two sets of item parameters lie along the line of $y=x$, implying that they are basically on the same metric.

However, for Contrast 1 involving Rasch models, the two sets of item parameter/test information estimates are not placed on the same scale. In each calibration process, the metric of parameter estimates is determined by standardizing item parameters rather than by standardizing person abilities (Baker & Kim, 2004). Since two different sets of items are calibrated, the metric of the two sets of item parameter estimates would be different. This metric difference is implied by the scatter plot of Figure 4.5 comparing item difficulty estimates from the dichotomous Rasch model against Thurstone threshold estimates from the PCM. Data points in the scatter plot do not lie along the line of $y=x$. To solve this problem, a linking between the two sets of parameter estimates is realized by forcing individual person-ability estimates to be equal in the two calibrations. Specifically, the polytomously-scored person ability estimates are carried into and fixed in the dichotomous calibration process. Consequently, two sets of test information estimates are placed on the same scale through person score anchoring. To further illustrate item parameter estimates are comparable after conducting person score anchoring, the scatter plots of item difficulty estimates against Thurstone threshold estimates before and after person score anchoring are presented in Figure 4.14A and Figure 4.14B, respectively. Figure 4.14A is reproduced from Figure 4.5 in the exploratory
IRT analysis. When person scores are not anchored, as shown in Figure 4.14A, the two sets of item parameters do not have the same measurement unit; while when person scores are anchored, as shown in Figure 4.14B, the two sets of parameters are comparable.

Figure 4.14. Scatter Plots of Item Difficulty Estimates against Thurstone Threshold Estimates, before and after Person Score Anchoring

A. Before Person Score Anchoring

Now that item parameter estimates derived from the two models are put on the same metric, comparisons between the two sets of test information are possible. Total test information curves derived from the two scoring methods are presented in Figure 4.15, 4.16, and 4.17 for the three contrasts respectively. In each figure, the A-panel presents test information derived from the real data, and the B-panel presents test information derived from the simulated data. In each chart, the thin line represents the sum of the dichotomously-generated item information and the thick line represents the sum of testlet
information derived from the polytomous model.

To locate the sources of difference between the two test information curves in each contrast, each total test information curve is further broken down by testlets; the testlet information curves are shown in Figure 4.18, 4.19 and 4.20 for the three contrasts, respectively. The observed within-$Q_3$ value of each testlet in the real data, as computed from Analysis 2, is also listed. Item and test information for the simulated data are computed using the average item parameter estimates from the 25 simulations.

**Rasch Model Comparison**

The Rasch model comparison is illustrated in Contrast 1. When LID is absent, as shown from the simulated data in Figure 4.15B, the two information curves are almost identical, which indicates that when items forming a testlet do not exhibit LID, changing scoring schemes does not change test information estimates for Rasch models.

Whereas, for the real data where deterministic local dependence is present, the polytomously-generated test information is higher in the middle range (-2 to +2) but lower at both ends of the ability spectrum than the dichotomously-generated information. An examination of each testlet information curve helps locate where the discrepancy occurs and whether it is related to local dependence. As shown in Figure 4.18 for the real data, a general pattern emerges that as the within-testlet $Q_3$ value increases, or as the degree of LID increases, the discrepancy between the two information curves increases.
Among the 13 testlets in the real data, Testlet 10 exhibits the highest degree of LID (within-$Q_3 = 0.324$) and Testlet 6 exhibits the lowest degree of LID (within-$Q_3 = 0.065$). The difference in the two information curves is much larger in Testlet 10 than in Testlet 6.

In summary, the pattern in Contrast 1 suggests that test information tends to be underestimated in the middle range but overestimated at both ends of the ability continuum when LID is present.

**Non-Rasch model comparison: Contrast 2 and Contrast 3**

The two information curve patterns in Contrast 2 and Contrast 3 are very similar to each other, but different from findings in Contrast 1. First, for both Contrast 2 and Contrast 3, even when LID is absent, as shown from the simulated data in Figure 4.16B and Figure 4.17B, there is an expected loss of information after forming testlets—test information generated from the polytomous scoring method is less than the information generated from the dichotomous scoring method; while in Contrast 1, the change of the scoring scheme does not have any influence on test information curves when LID is absent. Second, when LID is present, as shown in the real data, the drop in test information after forming testlets is larger than the information drop in the simulated data. Figure 4.19 and 4.20 suggests that as the degree of LID increases, the drop in test information after forming testlets get larger. The extra drop of test information in the real polytomous data indicates that an overestimation of test information occurs when locally
dependent items are treated as independent in the dichotomous data.

Findings from Contrast 2 and Contrast 3 agree with findings from previous studies. When the score-based method is used to accommodate LID, test information derived from the polytomous scoring method is less than test information derived from the dichotomous scoring method. The amount of information drop is actually composed of two parts – an expected drop due to the change in scoring schemes and a further drop due to the existence of LID. The expected drop in test information is caused by the loss of item-level responses in each testlet. When items are formed into testlets, the item-level response pattern is subsumed in a single total testlet score, which ultimately leads to the loss of total test information (the Rasch model is an exception and will be discussed later). When LID items are present, there is a further loss of test information in addition to the expected information loss. Previous studies have shown that the additional loss of information is a result of an overestimation of item slopes when LID is present. Therefore, treating dependent items as independent can lead to an overestimation on test information when non-Rasch models are employed.

The analysis results from this section suggest that selection of IRT models influence the pattern of test information change when the score-based method is used to handle LID. Changes in test information are actually composed of two parts – the change
due to the scoring scheme change when LID is absent and the change due to the existence of LID items. The change pattern of test information differs between Rasch models and non-Rasch models, in the situation where LID is absent, as well as in the situation where LID is present.

Under the null condition – when LID is absent, test information comparison between the dichotomous and the polytomous models informs the test information change solely caused by the scoring scheme change. Results from the simulated data in the three contrasts revealed that when LID is absent, changing the scoring scheme in Rasch models did not change test information; while for non-Rasch models, changing from a dichotomous to a polytomous scoring scheme lead to an expected loss of information. For non-Rasch models, the loss of information due to the scoring scheme change is the result of losing item-level response patterns after combing items into testlets. However, when only Rasch models are involved in such an analysis, the total score is a sufficient statistic and the pattern of responses is irrelevant. Loss of the item-level response pattern does not lead to test information change.

When LID is present in the data, the additional change of test information implies whether test information derived from LID-laden data is overestimated or underestimated. Depending on which models (i.e., Rasch models versus non-Rasch models) are used, the conclusion with regards to test information change is different.
This difference is related to whether item slopes are estimated in an IRT model. Test information is the sum of item information which is closely related to the $a$-parameter (item slope), as discussed in Chapter Two. Whether the $a$-parameter is estimated or fixed influences the calculation of item information. For Rasch models, $a$-parameter is fixed to one for both the simple dichotomous and the PCM; for non-Rasch models, item slopes are estimated during each calibration process (i.e., in Contrast 1, one common slope is estimated for the 1PL model and a different common slope is estimated for the GRM; in Contrast 2, each item has its own estimated slope value for the 2PL and the GRM). Since LID tends to inflate the $a$-parameter, test information is likely to be overestimated for models estimating the $a$-parameter.

However, for Rasch models, the change of test information when LID is present is in the opposite direction. For the ability range of -2 to 2 logits, where most ability estimates fall inside, an underestimation of test information has occurred in the LID-laden dichotomous data. In other words, the PCM catches more information than does the dichotomous model. A proposed explanation for this observation is illustrated as follows.

Take a 5-item testlet as an example. In the dichotomous Rasch model, each of the five item difficulty estimates is derived separately from each of the five dichotomous item scores. In the PCM, item parameters (i.e., five step parameter estimates) are based
on a total testlet score. Since in this study, items within the same testlet follow a Guttman scale, total testlet scores determine the response patterns. In this sense, each step parameter is actually estimated after considering all five dichotomous items simultaneously and thus extracts more information than if the five items were estimated separately. Again, this is a proposed explanation. A strict derivation or further studies are needed in the future.
Figure 4.15. Total Test Information Comparison: Contrast 1

A. Real Data

B. Simulated Data

Figure 4.16. Total Test Information Comparison: Contrast 2

A. Real Data

B. Simulated Data

Figure 4.17. Total Test Information Comparison: Contrast 3

A. Real Data

B. Simulated Data
Figure 4.18. Test Information Derived from the Dichotomous and the Polytomous Scoring Schemes, by Testlets - Contrast 1: Rasch Models

Real Data

Simulated Data

Testlet 1 (Item 1 – Item 5)

Testlet 2 (Item 6 – Item 10)

Testlet 3 (Item 11 – Item 15)

Within-$Q^3=0.190$

Within-$Q^3=0.098$

Within-$Q^3=0.111$
Chapter Four: Results

Testlet 4 (Item 16 – Item 19)

Within-QL = 0.128

Testlet 5 (Item 20 – Item 24)

Within-QL = 0.101

Testlet 6 (Item 25 – Item 27)

Within-QL = 0.065
Chapter Four: Results

Testlet 7 (Item 28 – Item 32)

Testlet 8 (Item 33 – Item 37)

Testlet 9 (Item 38 – Item 39)

Within-\text{Q3}=0.249
Within-\text{Q3}=0.120
Within-\text{Q3}=0.170
Chapter Four: Results

Testlet 10 (Item 40 – Item 44)

Within-\(Q3=0.324\)

Testlet 11 (Item 45 – Item 49)

Within-\(Q3=0.294\)

Testlet 12 (Item 50 – Item 54)

Within-\(Q3=0.140\)
Figure 4.19. Test Information Derived from the Dichotomous and the Polytomous Scoring Schemes, by Testlets - Contrast 2: 1PL vs. GRM with Equal Slopes

**Real Data**

Testlet 1 (Item 1 – Item 5)

Within-$Q^2$=0.190

Testlet 2 (Item 6 – Item 10)

Within-$Q^2$=0.098

**Simulated Data**

Testlet 1 (Item 1 – Item 5)

Within-$Q^2$=0.190

Testlet 2 (Item 6 – Item 10)

Within-$Q^2$=0.098
Chapter Four: Results

Testlet 3 (Item 11 – Item 15)

Testlet 4 (Item 16 – Item 19)

Testlet 5 (Item 20 – Item 24)

Within-Q3 = 0.111

Within-Q3 = 0.128

Within-Q3 = 0.101
Chapter Four: Results

Testlet 6 (Item 25 – Item 27)

Testlet 7 (Item 28 – Item 32)

Testlet 8 (Item 33 – Item 37)
Chapter Four: Results

Testlet 9 (Item 38 – Item 39)

- Within-Q3 = 0.170

Testlet 10 (Item 40 – Item 44)

- Within-Q3 = 0.324

Testlet 11 (Item 45 – Item 49)

- Within-Q3 = 0.294
Figure 4.20. Test Information Derived from the Dichotomous and the Polytomous Scoring Schemes, by Testlets - Contrast 3: 2PL vs. GRM with Varying Slopes
Chapter Four: Results

Testlet 2 (Item 6 – Item 10)

Within-Q3=0.108

Testlet 3 (Item 11 – Item 15)

Within-Q3=0.081

Testlet 4 (Item 16 – Item 19)

Within-Q3=0.134
Chapter Four: Results

Testlet 5 (Item 20 – Item 24)

Theta Information

Within-\(Q^2\)=0.110

Testlet 6 (Item 25 – Item 27)

Theta Information

Within-\(Q^2\)=0.078

Testlet 7 (Item 28 – Item 32)

Theta Information

Within-\(Q^2\)=0.225
Chapter Four: Results

Testlet 8 (Item 33 – Item 37)

Testlet 9 (Item 38 – Item 39)

Testlet 10 (Item 40 – Item 44)
Chapter Four: Results

Testlet 11 (Item 45 – Item 49)

\[ \text{Within-Q3} = 0.295 \]

Testlet 12 (Item 50 – Item 54)

\[ \text{Within-Q3} = 0.127 \]

Testlet 13 (Item 55 – Item 59)

\[ \text{Within-Q3} = 0.138 \]
Analysis 6: Score Estimates and Standard Error of Estimates

This analysis examines the relationship between the dichotomously-scored and the polytomously-scored person estimates and the standard errors of estimates, for the 347 subjects in the sample. To make the two sets of scores comparable, person scores generated from each estimation process are re-scaled to have a mean of zero and standard deviation of one. Standard error (SE) of estimates associated with each person is also transformed by dividing the SE by the observed standard deviation of the ability estimates across all subjects, so that the measurement unit of SE’s is expressed in the standard deviation unit of the ability estimates.

Score Estimates

Figure 4.21 presents the scatter plots and correlation coefficients between the two sets of scores, for real data and simulated data. Since the scatter plots for the 25 simulations are very similar, only the scatter plot from one simulation is presented for each contrast. In each scatter plot, the x-axis represents scores estimated from the dichotomous model and the y-axis represents scores estimated from the polytomous model, and a reference line of $y=x$ is provided. When the dichotomously-scored and polytomously-scored estimates are identical, all the scatter points would fall on this line.

Pearson’s correlation coefficient between the two sets of scores is very high (larger than .99) for all contrasts and for both real and simulated data. The scatter plot
patterns exhibit some discrepancies among the three contrasts. For Contrast 1, the simple Rasch model is fitted to the dichotomous data and the Partial Credit Model is fitted to the polytomous model using WINSTEPS. As shown, data points in both the observed and simulated scatter plots form a single line rather than a “scattered” pattern. There is a one-to-one relationship between the two sets of scores: subjects having the same dichotomously-scored estimate would have the same polytomously-scored estimate, or vice versa. This pattern results from the “specific objectivity” property of Rasch models, in which only total raw score decides the estimated person measure. Changing the scoring scheme to accommodate LID does not alter the total raw score for each subject. Therefore, subjects with the same total raw score would have the same ability estimates within each scoring method.

A comparison between the observed and simulated scatter plots suggests a possible impact of LID on score estimates. The simulated scatter plot shows that almost all data points are on the line of $y=x$, indicating that when LID is absent, the two scoring schemes from Rasch models can generate not only highly correlated results but almost identical results. One obvious difference that emerged in the observed scatter plot is a curvilinear pattern at the lower end of the ability spectrum, while in the simulated data, the relationship between the two sets of scores is linear for the entirety of the ability range. This difference suggests that when LID is present, scoring bias might exist for
subjects with low scores. However, this study cannot answer the question of whether or to what extent the scores are biased when LID exists, because “true” ability parameters are not known. However, despite the visual difference, all the dichotomously-generated scores are within the 95% confidence interval around their polytomously-generated scores. The largest score difference is 0.15 standard deviations. About 96% of subjects have a score difference within 0.10 standard deviations.

For Contrast 2, the 1PL model is fitted to the dichotomous data and the GRM with equal slopes is fitted to the polytomous data. The 1PL model has the same theoretic framework as the simple Rasch model, while the GRM does not belong to the Rasch family and does not possess the “specific objectivity” property. Therefore, person measures estimated from the GRM are not only decided by total raw scores, but also by specific response patterns. Subjects having the same dichotomously-scored estimates might have different polytomously-generated estimates, or vice versa. Comparing the scatter plots between the real and simulated data reveals a very similar pattern: both present a clear linear relationship between the two scoring schemes across the entire ability range. More variations are observed in the real data, but again, all the dichotomously-generated scores are within the 95% confidence interval around the polytomously-generated scores. The largest score difference is 0.20 standard deviations and about 93% of the subjects have a score difference within 0.10 standard deviations.
For Contrast 3, the 2PL model is fitted to the dichotomous data and the GRM with varying slopes is fitted to the polytomous model. Neither model has the “specific objectivity” property, and thus, the one-to-one relationship is not observed between the two sets of scores as in Contrast 1. The scatter plots in Contrast 3 reveal a very similar pattern as that in Contrast 2, except the scatter plots in Contrast 3 bear more variations. 99.4% (345 out of 347) of all the subjects do not differ in the two scores after considering the standard error of estimate. The largest score difference is 0.39 standard deviations and about 80% of subjects have a score difference within 0.1 standard deviation.
Figure 4.21. Score Estimates from the Dichotomous and the Polytomous Data

A. Contrast 1: Rasch Models

- **Real Data**
  - Dichotomously-scored Estimates
  - Polytomously-scored Estimates

- **Simulated Data**
  - Dichotomously-scored Estimates
  - Polytomously-scored Estimates

B. Contrast 2: 1PL Model vs. GRM with Equal Slopes

- **Real Data**
  - Dichotomously-scored Estimates
  - Polytomously-scored Estimates

- **Simulated Data**
  - Dichotomously-scored Estimates
  - Polytomously-scored Estimates

C. Contrast 3 - 2PL Model vs. GRM with Varying Slopes

- **Real Data**
  - Dichotomously-scored Estimates
  - Polytomously-scored Estimates

- **Simulated Data**
  - Dichotomously-scored Estimates
  - Polytomously-scored Estimates
Standard Error of Estimates

Figure 4.22 presents the scatter plots of standard error (SE) of estimates derived from the two scoring methods in Contrast 1. Figure 4.22A plots SE against the score estimates derived from the corresponding scoring method. Figure 4.22B plots the difference between the two SE’s for each subject (SE derived from the dichotomous scoring method minus SE from the polytomous scoring method) against the subject’s polytomously-scored estimate. If the difference is negative, the dichotomously-generated SE is smaller than the polytomously-generated SE, or the dichotomous scoring method provides more precise estimates. Figure 4.23 and 4.24 present the same information for Contrast 2 and Contrast 3, respectively.

The SE analysis in this section duplicates the findings from the test information analysis in that it indicates the most precisely and the least precisely estimated regions on the ability continuum. However, the test information analysis is only good at comparing test information generated from the dichotomous and polytomous models within the same contrast, provided that they are on the same metric. Test information cannot be compared among models from different contrasts and the magnitude of test information is not directly interpretable. On the other hand, the SE values in this analysis are rescaled so that they are expressed in the unit of standard deviation of the score estimates (i.e., SE=0.1 represents 0.1 standard deviation on the estimated ability score distribution).
For Contrast 1, we already knew from previous analysis, as shown in Figure 4.15, that the test information curve generated from the polytomous scoring scheme is higher in the middle range of the ability spectrum, but lower at both ends. In other words, the polytomous scoring method generates more precise estimates in the middle range but less precise estimates at the extreme ends. Figure 4.22 reveals that the difference in SE’s between the two scoring methods is hardly noticeable in the middle range (for score estimates within -2 and -0.5), but at both ends, the dichotomously-generated SE’s are lower. In the real data, the average difference in SE’s between the two scoring methods is 0.045, ranging from 0.003 to 0.113; in the simulated data, the average difference is 0.004, ranging from a value less than 0.001 to 0.010.

For both Contrast 2 and Contrast 3 in the real data, all dichotomously-generated SE’s are smaller than the polytomously-generated SE’s. For Contrast 2, the average difference in the real data is 0.038, ranging from 0.015 to 0.078, and the average difference in the simulated data is 0.017 ranging from a value less than 0.002 to 0.065. For Contrast 3, the average difference in the real data is 0.047, ranging from less than 0.001 to 0.109, and the average difference in the simulated data is 0.016, ranging from a value less than 0.001 to 0.091. In general, although the SE difference between the two scoring schemes in the real data is larger than in the simulated data, almost all differences are within 0.1 standard deviations.
Figure 4.22. Standard Error of Estimates from the Dichotomous and the Polytomous Scoring Methods - Contrast 1: Rasch Models

A. Standard Errors

Real Data

Simulated Data

B. Difference in Standard Errors between the Two Scoring Methods

Real Data

Simulated Data
Figure 4.23. Standard Error of Estimates from the Dichotomous and the Polytomous Scoring Methods - Contrast 2: 1PL vs. GRM with Equal Slopes

A. Standard Errors

Real Data

Simulated Data

B. Difference in Standard Errors between the Two Scoring Methods

Real Data

Simulated Data
Figure 4.24. Standard Error of Estimates from the Dichotomous and the Polytomous Scoring Methods - Contrast 3: 2PL vs. GRM-Varying Slopes

A. Standard Errors

Real Data

Simulated Data

B. Difference in Standard Errors between the Two Scoring Methods

Real Data

Simulated Data

This analysis compared the score estimates and standard errors of estimates between the dichotomous and the polytomous models for the three Contrasts. The results showed that highly correlated person score estimates were generated. In general, the dichotomously-scored estimates were more precise than the polynamously-scored estimates for non-Rasch models, but the differences in SE’s for all the three contrasts were within 0.1 standard deviations.
Summary

This Chapter presents the results of the six analyses outlined in Chapter Three. The exploratory analysis conducted on the real data has confirmed the existence of a Guttman scaling pattern, or a deterministic response pattern, among items in the same testlet. Further analysis using Yen’s $Q_3$ has shown that items within the same testlet exhibit LID with various degrees. However, in contrast to published literature, LID caused by the deterministic pattern seems to have an insignificant impact on a dimensionality assessment. The internal reliability, estimated by the alpha coefficient, is slightly overestimated when LID is present. Nevertheless, the impact of LID on the internal reliability is not dramatic when compared with similar previous studies.

When the score-based method is used to accommodate LID, person score estimates from the two scoring schemes are highly correlated regardless of which model is fitted. Although the difference in score estimates between the two scoring methods gets larger when item slopes are considered, most of the differences are within measurement errors. Test precision is overestimated when standard dichotomous IRT models are fitted to a LID-laden dataset. However, the actual differences in standard errors between the two scoring schemes for most subjects are within 0.1 standard deviations. Chapter Five discusses interpretations and implications of the results in the health care field, acknowledges limitations of this study, and provides suggestions for future research.
Chapter Five: Conclusions

This chapter recapitulates the purpose statement and research questions; presents a summary of findings based on the five research questions and corresponding hypotheses; discusses results that went against the initial hypothesis; talks about implications for measurement practitioners in the field of public health; acknowledges limitations of this dissertation; and finally, makes recommendations for further research.
Local item independence is a key assumption to be satisfied when applying IRT models. A wealth of research has been conducted to address the problem of violating the local independence assumption on large-scale achievement tests in the field of education. However, in the field of health care, where IRT has been experiencing a greater popularity, the problem of local item dependence (LID) was studied inadequately. Some unique features about LID detected in health surveys deserve a separate and systematic examination. In patient-reported outcome measures, it is more common to see two or more items exhibit a similar structure in wording or content, which leads to a deterministic response pattern, or a Guttman scaling pattern. This special response pattern might exhibit LID.

The purposes of this study were to investigate whether such deterministic response structure, or a Guttman scaling pattern, that appeared among a subset of items in patient-reported outcome measures exhibited local dependence; to examine the impact of the deterministic order dependence on test statistics and on parameter estimates; and to determine whether these effects differed when different IRT models were employed.

Based on theoretic analyses and previous research findings, it was hypothesized that the deterministic response pattern among a subset of items would (1) cause items to be locally dependent; (2) influence the dimensionality assessment of test data; (3) inflate the internal test reliability in Classical Test Theory; (4) have minimal impact on person
ability estimates regardless of which IRT model was applied; and (5) overestimate test information when item slopes (e.g., 2PL model) were considered, but have little impact on test information when item slopes were not considered (e.g., Rasch models).

Data used in this dissertation come from a norms study on the Pediatric Evaluation of Disability Inventory (PEDI) – Mobility Domain, consisting of 59 dichotomously-scored items grouped into 13 content topics. Items within the same content are structured in such a way that early-appearing items are easier than late-appearing items. Thus, a deterministic response pattern, or a Guttman scaling pattern, is observed among items in the same content area.

To accommodate potential LID caused by the deterministic response pattern, the score-based method was employed. Dichotomous items within the same content were combined to form one polytomous item. Polytomous IRT models were then fitted to the “new” data composed of testlets. In essence, this method eliminates LID by making interdependent items “disappear” and replacing them with a new polytomous testlet. It should be noted that the score-based method accommodates within-content LID, but still assumes local independence among items from different content topics. For each analysis, results based on the dichotomous scoring method (LID-laden data) were compared to the results based on the corresponding polytomous scoring method (LID-free data).

Since research questions #4 and #5 investigated whether the impact of LID
differed when item slopes were estimated, three contrasts were designed to compare the results derived from the two models. The three contrasts were: 1) the simple Rasch model versus the PCM, estimated by WINSTEPS; 2) the 1PL model versus the GRM fixing item slopes to be equal, estimated by MULTILOG; and 3) the 2PL model versus the GRM allowing slopes to vary, estimated by MULTILOG. Contrast 1 involved only Rasch models, and Contrasts 2 and 3 involved non-Rasch models.

The complete analysis plan contained a series of exploratory analyses and five subsequent analyses to address five research questions and their corresponding hypotheses. The exploratory analyses included verification of the existence of a deterministic response pattern among items in the same content topic, a Classical Test Theory (CTT) analysis providing an overview of item easiness and item discriminations, and several IRT analyses evaluating item parameter estimates and general fits of different models.

**Summary of Findings**

**Exploratory Analyses**

It was shown in Chapter Four that a Guttman Scaling pattern was confirmed for all the 13 testlets, each with a very high value of *coefficient of reproducibility*, an index for determining a Guttman Scale. At least 96% of the responses matched a Guttman scale pattern for 12 out of the 13 testlets. It was obvious that there was a deterministic response
pattern among items belonging to the same content topic.

The CTT analysis revealed that this inventory was relatively easy for this sample, with an average item easiness of 0.792 for the dichotomous items. This was within expectation, because the PEDI was designed to evaluate developmental disabilities and data used in this study were collected from children without disability. It was reasonable to see that a large proportion of children were able to perform most of the items in the inventory.

In the IRT analysis, three sets of models (i.e., three contrasts), were fitted to the dichotomous and the polytomous raw data: 1) the simple Rasch model versus the PCM, 2) the 1PL model versus the GRM with equal item slopes, and 3) the 2PL model versus the GRM with varying item slopes. The first contrast reflected a Rasch model comparison; the second contrast reflected a non-Rasch model comparison but fixing item slopes to be equal in each model; and the third contrast reflected a non-Rasch model comparison allowing slopes to vary.

To compare location parameter estimates generated from the two models in each contrast, item difficulty estimates derived from a dichotomous model were correlated and plotted against the Thurstone threshold estimates derived from a corresponding polytomous model. Results showed that the two sets of location parameter estimates in each contrast were highly correlated for most items, with Item 20 being identified as an
outlier for all three contrasts. This was probably because Item 20 was extremely easy – only one out of the 347 subjects was not able to perform the task stated in this item. Parameter estimation procedures on items with perfect or almost perfect scores might be different between the two models in each software program.

Unlike item location parameter estimates, which were highly similar between the two models in all the three contrasts, item slope parameters might not be estimated reliably in the presence of LID. In Contrast 3, item slope estimates fluctuated greatly and some items had unusually high slope estimates in the 2PL model. Specifically, item slope estimates in the 2PL model ranged from as low as 0.27 for Item 20 to as high as 10.87 for Item 25, with an average of 4.51 and standard deviation of 1.85. In the GRM, item slope estimates ranged from 3.27 for Testlet 7 to 7.76 for Testlet 6, with an average of 4.71 and standard deviation of 1.63. The range of item slope estimates in the 2PL model was much larger than that of the GRM. Also, in the 2PL model, three items had unusually high slope estimates which were larger than 8. The fluctuation of item slope estimates and some extremely high slope estimates in the 2PL model indicate a possible adverse impact of LID on item slope estimates.

Fit statistics under each model identified different sets of misfitting items. In this study, no misfitting items were excluded for the following reasons. First, decisions on keeping or deleting items could not be solely based on fit statistics, not to mention one
single fit statistic. Besides, the $\chi^2$ fit statistic was sensitive to sample size and its value might be inflated. Second, in this study, fit statistics were only used as an aid to better understand the data, rather than to make decisions as to whether to keep or delete items that did not fit. Therefore, no further fit analysis was conducted.

**Hypothesis 1: Deterministic response patterns cause items to be locally dependent.**

Findings from Chapter Four support hypothesis 1 that the deterministic response pattern did cause items to be locally dependent. The degree of local dependence exhibited among items in the same testlet was moderate to high. A comparison between the within-content $Q_3$ distribution with the between-content $Q_3$ distribution also revealed that within-content $Q_3$ values were all positive and large in magnitude compared to the expected value, while between-content $Q_3$ values were small in magnitude and roughly distributed symmetrically around the expected value. This finding suggests that LID exists among items within the same content topic but not among items from different content topics. Thus, the adoption of the score-based method is considered appropriate.

**Hypothesis 2: Deterministic response patterns influence the dimensionality assessment of test data.**

Hypothesis 2 is not supported by this study. Essentially, a unidimensional structure was detected for both the dichotomous and the polytomous data, as suggested from the scree plots. In other words, no obvious extraneous dimensions were identified in
the LID-laden dichotomous data. This finding seems to disagree with previous studies which state that the fundamental cause of LID is the introduction of extraneous dimensions common to a subset of items in the test. A detailed discussion in this regard is presented in the Discussion section.

**Hypothesis 3: Deterministic response patterns inflate the internal test reliability.**

This hypothesis is confirmed by findings from Chapter Four. However, the amount of inflation on the internal reliability was trivial in practice, especially after taking into account the big drop in the number of “items” from 59 in the dichotomous data to 13 in the polytomous data. Regardless of which scoring scheme was used, decent coefficient alpha values were achievable. In other words, LID caused by the deterministic response pattern has insignificant practical impact on the internal reliability estimate.

**Hypothesis 4: Deterministic response patterns overestimate test information when a 2PL IRT model is fitted to the original dichotomous data, but have little impact on test information function when a 1PL (Rasch) model is fitted to the original data.**

The first part of this hypothesis is supported - that is, when the 2PL model was fitted to the LID-laden dichotomous data, test information was overestimated: this agrees with previous literature. However, divergent results were reached for the second half of the hypothesis, even though the 1PL model and the dichotomous Rasch model are theoretically identical. When the 1PL model was estimated using MULTILOG, a similar
pattern as that in the 2PL contrast was observed - test information was overestimated along the entire ability continuum. However, for the Rasch model contrast, test information in the LID-laden dataset was underestimated for ability estimates within -2 and 2 logits, but overestimated for extreme ability estimates at both ends. This discrepancy implies a difference of LID’s impact on test information between Rasch and non-Rasch models.

**Hypothesis 5: Deterministic response patterns have minimal impact on person ability estimates regardless of which IRT model is applied.**

Hypothesis 5 is supported and agrees with previous findings. Highly correlated ability estimates were generated from the dichotomous and the polytomous scoring schemes regardless of the three contrasts. In the Rasch model contrast, the two sets of scores were not only highly correlated, but also highly similar due to the fact that total test score was a sufficient statistic in Rasch models and changing scoring scheme did not change total test score for each person. In the non-Rasch model contrasts, ability estimates were also highly correlated, but the difference between the two sets of scores bore more variations. Despite this, most of the differences between the two score estimates for each person were within the measurement errors.

**Discussions**

As stated in Chapter Two, this study aimed at filling two gaps in the current LID
literature involving the score-based method to handle LID. First, most previous LID studies examined combination dependence (Keller, Swaminathan, & Sireci, 2003; Lee, 2004; Sireci, Thissen, & Wainer, 1991; Wainer, 1995; Yen, 1993; Zenisky, Hambleton, & Sireci, 2002), while this study examined one special case of order dependence – the deterministic dependence, which was commonly observed in many patient-reported outcome measures. Second, most previous studies took into account item slopes in the model fitting process and employed the 2PL or 3PL models (Chen & Thissen, 1997; Keller, Swaminathan, & Sireci, 2003; Sireci, Thissen, & Wainer, 1991; Thompson & Pommerich, 1996; Wainer & Thissen, 1996; Yen, 1984, 1993), while this study investigated whether LID posed any adverse effect on Rasch models as they did on non-Rasch models.

This study has reached some similar findings as previous ones. For example, LID tends to inflate the internal reliability estimate and overestimate slope parameters, but has minimal impact on item location parameters and ability estimates. When compared with combination dependence, the deterministic LID influences test statistics at a lesser degree. Specifically, although LID inflates the internal reliability estimate, the amount of overestimation does not have practical significance; scores generated from the LID-laden data and LID-free data are not only highly correlated but also very similar to each other.

Other than these similarities, this study also identifies two findings that do not
agree with previous literature. First, the impact of LID on test information differs
between Rasch models and non-Rasch models; and second, a violation of the local item
independence assumption does not lead to an obvious violation of the unidimensionality
assumption. Technical discussions on each of these two points are presented as follows.

*LID’s Impact on Test Information*

Test information is another major concern when LID items are present. Previous
studies indicated that LID tended to overestimate test information (Keller et al., 2003;
Thompson & Pommerich, 1996; Zenisky, Hambleton, & Sireci, 2002). This study
identified a difference of LID’s impact on test information between Rasch models and
non-Rasch models.

For non-Rasch models, results from this study are consistent with findings from
previous studies: forming testlets leads to a natural loss of test information even when
items are locally independent. The existence of LID leads to a further loss of test
information, a result of an overestimation of item slopes. But for Rasch models, when
LID is absent, test information derived from the dichotomous Rasch model is virtually
identical to test information derived from the PCM; when LID is present, test information
derived from the LID-laden data is underestimated within the ability range of -2 to +2
logits, opposite to the test information change in non-Rasch models, where test
information is overestimated along the entire ability continuum.
Since most previous LID studies employed non-Rasch models when using the score-based method to accommodate LID, their results unanimously pointed out that the major disadvantage of the score-based method was loss of test information. Results of this dissertation supplement their findings and provide a more complete picture on LID’s impact on test information. Acknowledging the difference of LID’s impact on test information between Rasch and non-Rasch models better informs the consequences of applying the score-based method to handle LID. Since Rasch models were employed in the initial norms PEDI study, using the score-based method to accommodate LID would actually improve estimation precisions for most subjects whose ability levels are between -2 and +2 logits, but reduce estimation precisions for subjects with extreme scores. However, it should be noted that the change of test information after employing the score-based method would have a stronger impact on the standard error of estimates (SEE) for subjects with extreme scores than for subjects with mid-range scores. This is because SEE is the inverse of the square root of test information. The larger the test information, the smaller the SEE is. Due to the inverse relationship between test information and SEE, a large change of high-value test information may not lead to a large change in SEE; on the other hand, a small change of low-value test information may lead to a large change in SEE. Since SEE normally serves as stopping rules in a Computerized Adaptive Testing (CAT) session, the consequence of the SEE change in
Chapter Five: Conclusions

CAT sessions needs more research.

Local Item Dependence and Unidimensionality

Another finding in this dissertation that is not consistent with previous studies is the relationship between the local item independence assumption and the unidimensionality assumption. Compared with combination dependence, the deterministic order dependence detected in the PEDI tends to have a small effect on the assessment of dimensionality. Results from this study suggest that although the Guttman scaling pattern exhibits LID, it does not cause the data to deviate much from a unidimensional structure.

It is well documented from previous literature that the fundamental cause of LID is the introduction of certain extra dimensions that are common to a subset of items but different from the primary dimension (factor) being measured. The assumption of unidimensionality and the assumption of local independence agree with each other. This study identified several testlets with moderate to high degrees of local dependence, but no extra dimensions were obvious from the exploratory factor analysis.

One explanation for this observation is that items within the same content topic do, indeed, share a common extraneous dimension, but the extraneous dimension correlates highly with the primary dimension being measured. When two dimensions are highly correlated, it is difficult to separate one from the other; thus, it is difficult to detect
through an exploratory factor analysis. Geometrically, when the highly-correlated extra
dimension is decomposed into two vectors (dimensions) – one parallel to the primary
dimension and one orthogonal to the primary dimension, the orthogonal dimension would
be small in value. In this dataset, each testlet was highly correlated with the total score.
Specifically, 11 out of the 13 testlets had correlations higher than 0.85, and the other two
had correlations of 0.82 and 0.79, respectively. Such high correlations between each
testlet and the total score indicate that should the factor (dimension) measured by each
testlet be decomposed into the same two vectors, the one that is orthogonal to the primary
factor would be small in value, and thus would have a minor impact on the
unidimensional structure of the test data as a whole. In this sense, each testlet could be
regarded as a mini-scale measuring the same dimension as the main scale, and a
combination of these mini-scales makes a stronger and more reliable instrument.

An alternative explanation is that items within the same content topic correlate
more with each other not because they share common extra dimensions, but because of
the intrinsic deterministic response structure. If this is the case, it may be helpful to
distinguish two kinds of LID – underlying LID and surface LID. Underlying LID refers
to LID caused by extraneous dimensions, and surface LID caused by a superficial data
structure. Items exhibiting higher correlations could be due to either or both kinds of
LIDs. In this study, it is more likely that surface LID exists, because the unidimensional
assumption is essentially met.

Regardless of which of the above two explanations is true, findings from this study suggest that when LID is present without a violation of the unidimensionality assumption, its adverse impact is not as severe as the impact of LID caused by a violation of the unidimensionality assumption.

**Implications**

With regard to health survey designs, findings from this study indicate that it is fine to construct and include items showing deterministic order dependence as long as the unidimensionality assumption is basically satisfied. If no action is taken to handle the existing LID, employing the simple Rasch model to LID-laden data would minimize the adverse impact of LID on parameter estimates, since the slope parameter is fixed to a value of one. Rasch models were specified in developing the PEDI scales and in generating standardized scale scores in the PEDI norms study (Haley, Coster, Ludlow, Haltiwanger, & Andrellos, 1992), thus, the potential negative impact of the existing LID was minimal.

When high-degree of LID needs to be accommodated, the score-based method is sufficient and combining LID items into testlets prevents survey developers from unnecessarily deleting items. For example, when using residual correlations (i.e., $Q_3$) to evaluate and identify LID items, items with residual correlations higher than a critical
value become candidates for deletion. If LID items are combined to form a polytomous item, there is no need to delete any of these items.

With regard to clinical practices, a major concern rises from the precision level of the estimates. LID has a stronger impact on test information than on other test statistics. When a paper-and-pencil survey is administered, correct test information estimates could be obtained by using the score-based method after data are collected. However, when a Computerized Adaptive Test (CAT) is administered, stopping rules are often decided by test information. If test information is overestimated, a CAT session stops before it reaches the actually precision level; and if test information is underestimated, a CAT session continues when it should stop. If no action is taken to control LID, a practical adjustment on precision levels could be realized by setting the stopping rule at a higher level when test information is overestimated, or at a lower level when test information is underestimated. More research is needed in this regard.

LID’s influence on test information also plays an important role when cut scores are needed to categorize patients into different “stages”. The term stage is similar to the performance standards in the educational context. Currently, the major purpose of categorizing patients into stages is to better interpret continuous scores generated from IRT models (Jette, Tao, & Haley, in press; Tao, Haley, Coster, Ni, & Jette, in press). In this case, correct precision estimates, especially around the cut point areas, are important to
precisely categorize patients into stages.

As discussed in Chapter One, items sharing similar phrasing structures which lead to deterministic response patterns are necessary and commonly observed in outcome measures in the health care field. Given the findings from this study that deterministic response patterns do cause local dependence and do have an impact on test statistics and parameter estimates, two practical questions in managing LID in the health care field could be: 1) does LID caused by deterministic patterns needs to be taken care of? If yes, 2) which method is more appropriate for reducing the negative impact of LID? Based on findings from this study, as well as from previous studies, detailed discussions of these two questions are presented as follows.

*Does the deterministic LID need to be taken care of?*

It is always important to check the unidimensional structure of any dataset when IRT models are employed. If LID is caused by obvious extraneous dimensions, the impact of LID could be serious and approaches should be taken to control the multidimensional structure.

On the other hand, if the deterministic LID is not caused by any obvious extra dimensions, the adverse effect of LID on certain test statistics may be at a lesser degree. Under certain circumstance, LID items could stay in the scale as if they were independent (e.g., the ratio of the number of locally dependent items to the number of locally
independent items is relatively small, and LID items do not have a substantial impact on item parameter estimates).

In the PEDI data used in this study, all individual items are grouped into content units and items in each unit form a Guttman scale. This kind of data pattern may not be so common in other scales. Rather, it is usual to see pairs of interdependent items in a scale. When LID items appear in pairs, a simple examination of the impact of LID on item parameter estimates could be realized by calibrating the test twice – once including all items and once excluding one of the paired LID items (Hays et al., 2007; Hill et al., 2007; Reeve et al., 2007). If no substantial difference happens on item parameter estimates when including the LID items, it is probably fine to keep them in the scale. Otherwise, measures should be taken to control local dependence. Note that when the number of interdependent items forming a testlet is large, this method might not be practical.

Which method should be used to handle LID items?

As introduced in Chapter Two, four corrective options are available in the presence of LID items. First, delete LID items; second, ignore them and treat them as if they were independent; third, apply the score-based method by combining LID items into testlets and analyze the new test with polytomous IRT models; and four, apply the item-based approach.

Deleting items is usually not preferred, especially in certain fields where the
number of items is limited and generating items is difficult. Ignorance of LID items is an alternative only when the degree of local dependence is low or ignorance does not pose an adverse impact on item parameter estimates. As stated in the previous section, these effects could be examined by calibrating the data twice – including and excluding LID items to see their influence on item parameter estimates.

The item-based approach accommodates LID by building multidimensional models. It involves more complicated but more flexible models. When LID is caused by a salient multidimensional structure, an item-based approach would be appropriate. However, if LID is the result of an intrinsic data structure such as the deterministic response pattern, the item-based models might not be an optimal choice.

This study supports using the score-based method to accommodate deterministic LID that is not caused by any obvious extra dimensions. This approach is straightforward to understand in theory and easy to carry out in practice. However, when the score-based method is used, test information is an important consideration. Two complete different conclusions are reached in this study in terms of test information change. For non-Rasch models, which are commonly used in the educational context, it is the loss of test information that should be considered. However, in the health care field, Rasch models are as common as the 2PL model. If LID items are treated as if they were independent, the actual precision level, indicated by the test information, is actually lower than it
should be for non-extreme ability estimates.

**Limitations**

Three limitations of the study are acknowledged. The first limitation is the ceiling effect in the raw data. One consequence to the ceiling effect is low variance of the easiest dichotomous items. For very easy items, perfect or almost-perfect scores are very likely to be observed, which result in very low item variances and high standard error of estimates in the IRT analysis for the dichotomous data. Nevertheless, if a ceiling effect is a true situation for a population, using the score-based method could reduce the number of “items” having perfect item scores. When dichotomous items are combined into testlets, the chance of having a perfect testlet score would be smaller than the chance of having a perfect dichotomous item score, since having a perfect testlet score requires having perfect item scores for all the dichotomous items in the testlet.

The second limitation is using relatively small-sized sample. A size of 347 subjects is adequate for the Rasch analysis, but a little small for the 2PL model analysis. Fortunately, analysis results on the 2PL model agree with previous findings.

The third limitation is, strictly speaking, a delimitation. Narrowing of the scope is a delimitation, whereas a limitation is a potential weakness in the design (Creswell, 1994). This study is not a pure simulation study and therefore some questions cannot be answered, since certain variables are not controlled. The simulated data generated from
parameter estimates in the raw data, are only good at providing a LID-free null condition for comparison. For example, in testing hypothesis 4, it is only possible to examine the correlation between the two sets of scores. It is impossible to examine the bias of scores estimated from the LID-laden data because “true ability scores” are not known. To study the scoring bias resulted from LID, polytomous scores can be simulated on the 13 testlets. The polytomous data is the LID-free data and will be treated as the true model. Then the 13 testlets are unfolded back to the 59 dichotomous items. For example, a five-item testlet scored 3 will be unfolded to five dichotomous items with the first three items scored 1 and the last two items scored 0. Scores generated from the polytomous data will be treated as true scores, which are compared to scores generated from the dichotomous data.

Other simulation studies could include, but not limited to, examinations of LID’s impact when different proportions of LID items are included in a health survey, and when different numbers of LID items are formed into testlets. It is also interesting to see the extent to which LID influences categorizing patients into different stages.

**Future Research Recommendations**

This study was conducted on the paper and pencil format of the PEDI outcome measure. Computerized Adaptive Testing (CAT), in contrast, has become a trend in the health field. The next promising step is to investigate the impact of LID on CAT sessions.
Some specific topics of interest could be 1) the impact of LID on the stopping rules; 2) advantages and disadvantages of administering testlets as a basic unit in the CAT session over administering individual dichotomous items in the CAT session; and 3) employing the item-based method to manage LID in the CAT session.

The second recommendation is to apply a similar study on a different population. Diversified patient populations are a unique characteristic in the health care field. One health survey could be administered to a norm population or to a specific patient population. The purpose of administering a health survey to a norm population is often to establish a baseline/reference point for comparison, or to establish scale scores standardized on the norm population. Item parameter estimates generated from the norm population could also be used to link item parameter estimates generated from other populations. The target population in this study is the norm population. When the same scale is administered to a different population, e.g., a population with disabilities, the results might be different. For example, in this study, the Guttman scaling pattern is observed for items within the same content topic. This might not be true for another population. The ceiling effect, which was observed in the norm population, might be replaced by a floor effect in a different population.
Summary

This chapter provides an overview of the general research design and analysis results. There are detailed discussions about the relationship between the local item independence assumption and the unidimensionality assumption, and the difference in LID’s impact on test information for Rasch models and non-Rasch models. Then, practical implications of the study are discussed in terms of whether deterministic LID should be taken care of and which approach is more appropriate. Two limitations and one delimitation are acknowledged. Lastly, it is suggested that a future LID study in Computerized Adaptive Testing and a similar LID study on a different population might be useful.
References


References 210

S32-S38.


Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the
estimated discrimination parameters in item response theory. Psychological Methods, 6(2), 181-195.


References


Appendix A: Pediatric Evaluation of Disability Inventory

Mobility Domain, Item score 0 = unable; 1 = capable

A. Toilet Transfers
1. Sits if supported by equipment or caregiver
2. Sits unsupported on toilet or petty chair
3. Gents on and off low toilet or potty
4. Gets on and off adult-sized toilet
5. Gets on and off toilet, not needing own arms

B. Chair/Wheelchair Transfers
6. Sits if supported by equipment or caregiver
7. Sits unsupported on chair or bench
8. Gets on and off low chair or furniture
9. Gets in and out of adult-sized chair/wheelchair
10. Gets in and out of chair, not needing own arms

C. Car Transfers
11. Moves in car; scoots on seat or gets in and out of car seat
12. Gets in and out of car with little assistance or instruction
13. Gets in and out of car with no assistance or instruction
14. Manages seat belt or chair restraint
15. Gets in and out of car and opens and closes car door

D. Bed Mobility/Transfers
16. Raises to sitting position in bed or crib
17. Comes to sit at edge of bed; lies down from sitting at edge of bed
18. Gets in and out of own bed
19. Gets in and out of own bed, not needing own arms

E. Tub Transfers
20. Sits if supported by equipment or caregiver in a tub or sink
21. Sits unsupported and moves in tub
22. Climbs or scoots in and out of tub
23. Sits down and stands up from inside tub
24. Steps/transfers into and out of an adult-sized tub

F. Indoor Locomotion Methods
25. Rolls, scoots, crawls, or creeps on floor
26. Walks, but holds onto furniture, walls, caregivers or uses devices for support
27. Walks without support

G. Indoor Locomotion: Distance/Speed
28. Moves within a room but with difficulty (falls; slow for age)
29. Moves within a room with no difficulty
30. Moves between rooms but with difficulty (falls; slow for age)
31. Moves between rooms with no difficulty
32. Moves indoors 50 feet; opens and closes inside and outside doors
H. Indoor Locomotion: Pulls/Carries Objects
33. Changes physical location purposefully
34. Moves objects along floor
35. Carries objects small enough to be held in one hand
36. Carries object large enough to required two hands
37. Carries fragile or spillable objects

I. Outdoor Locomotion: Methods
38. Walks, but holds onto objects, caregiver, or devices for support
39. Walks without support

J. Outdoor Locomotion: Distance/Speed
40. Moves 10-50 feet (1-5 car lengths)
41. Moves 50-100 feet (5-10 car lengths)
42. Moves 100-150 feet (35-50 yards)
43. Moves 150 feet and longer, but with difficulty (stumbles; slow for age)
44. Moves 150 feet and longer with no difficulty

K. Outdoor Location: Surfaces
45. Level surfaces (smooth sidewalks, driveways)
46. Slightly uneven surfaces (cracked pavement)
47. Rough, uneven surfaces (lawns, gravel driveways)
48. Up and down incline or ramps
49. Up and down curbs

L. Upstairs
50. Scoots or crawls up partial flight (1-11 steps)
51. Scoots or crawls up full flight (12-15 steps)
52. Walks up partial flight
53. Walks up full flight, but with difficulty (slow for age)
54. Walks up entire flight with no difficulty

M. Downstairs
55. Scoots or crawls down partial flight (1-11 steps)
56. Scoots or crawls down full flight (12-15 steps)
57. Walks down partial flight
58. Walks down full flight, but with difficulty (slow for age)
59. Walks down full flight with no difficulty