Between-person and Within-person Subscore Reliability: Comparison of Unidimensional and Multidimensional IRT Models

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Okan Bulut

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Michael C. Rodriguez, Adviser

June, 2013

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Michael C. Rodriguez, for accepting me to the QME program and providing his excellent guidance and support during my graduate study at the University of Minnesota. He has been the main source of my inspiration for educational measurement and psychometrics. I feel very fortunate to be his advisee. I also would like to express my appreciation to Dr. Mark L. Davison for helping me to determine my dissertation topic and mentoring me patiently throughout my dissertation process. Without his valuable insights and feedbacks, this dissertation wouldn't have been completed.

I would like to thank Dr. Ernest C. Davenport, Jr. and Dr. David J. Weiss who served on my doctoral dissertation committee. Their constructive comments and valuable suggestions helped me a lot to understand the nuts and bolts of my dissertation.

Finally, I am grateful to my mom and brother for their constant support and love; and to my friends and colleagues at the University of Minnesota for their cooperation, support, and positive thoughts.

Dedication

This dissertation is dedicated to my parents, Kayhan and Ayfer Bulut, who have been a constant source of love and support during my life, and to my brother, Onur Bulut, who has always been there when I need.

Abstract

The importance of subscores in educational and psychological assessments is undeniable. Subscores yield diagnostic information that can be used for determining how each examinee's abilities/skills vary over different content domains. One of the most common criticisms about reporting and using subscores is insufficient reliability of subscores. This study employs a new reliability approach that allows the evaluation of between-person subscore reliability as well as within-person subscore reliability. Using this approach, the unidimensional IRT (UIRT) and multidimensional IRT (MIRT) models are compared in terms of subscore reliability in simulation and real data studies. Simulation conditions in the simulation study are subtest length, correlations among subscores, and number of subtests. Both unidimensional and multidimensional subscores are estimated with the maximum a posteriori probability (MAP) method. Subscore reliability of ability estimates are evaluated in light of between-person reliability, within-person reliability, and total profile reliability. The results of this study suggest that the MIRT model performs better than the UIRT model under all simulation conditions. Multidimensional subscore estimation benefits from correlations among subscores as ancillary information, and it yields more reliable subscore estimates than unidimensional subscore estimation. The subtest length is positively associated with both between-person and within-person reliability. Higher correlations among subscores improve between-person reliability, while they substantially decrease within-person reliability. The number of subtests seems to influence between-person reliability slightly but it has no effect on within-person reliability. The two estimation methods provide similar results with real data as well.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

Standardized tests are one of the most common measurement tools in educational and psychological assessment. Scores from standardized tests are often used for making important decisions such as, K-12 school accountability, high school graduation, college and graduate school admissions, professional certification, employment, etc. These tests are usually designed to measure several domains based on content areas, strands, attributes, or skills such as the Graduate Record Examination (GRE) and the SAT Reasoning Test. Similarly, test batteries (e.g., Woodcock-Johnson Test, MMPI-2) consist of several subtests, each of which measures a specific domain.

In testing programs such as SAT, two types of scores are typically reported. These are domain scores (i.e., subscores) based on examinees' performance on each domain and an overall composite score that is usually a weighted sum or a weighted average of the subscores. The provision of such test scores while meeting conventional requirements of quality for score reporting on high-stakes assessments has been a challenge in terms of test development and psychometrics (Thissen & Edwards, 2005). Although subscores from clusters of very small numbers of items may not be highly reliable, reporting subscores can be still useful because of their potential diagnostic value (Sinharay, 2010). Compared to an overall composite score, subscores may be more informative for determining how the examinee's abilities/skills vary over the different domains.

Why Subscores are important?

The usefulness of assessments that report subscores or domain scores is well accepted by policy makers, college admissions officers, school district administrators, and other educators. Under the No Child Left Behind Act of 2001 (NCLB) in the U.S., every state is required to administer statewide accountability assessments and report the students' scores in the major content domains – such as reading, writing, science, and mathematics – to measure school progress. In the classroom, teachers can benefit from subscores when they need to evaluate students' strengths and weaknesses. By using subscores, teachers can determine the most effective intervention or instructional program for the students based on their performance in each content domain. In addition to the educators and decision-makers, students and parents can also utilize the domain scores from assessments. Students can see their strengths and weaknesses in different content areas and use this information to plan their future studies (Haladyna & Kramer, 2004). Parents can use subscore information to monitor their child's achievement in the specific content areas, assess the effectiveness of the instruction that he/she receives, and identify potential learning difficulties that the student encounters.

In addition to the states, higher education institutions such as colleges and universities prefer to use scores from each domain for admissions because subscores can distinguish between candidates with the same or very similar total scores. In addition to admissions, colleges and universities use subscores as a summary of students' performance on different domains to better evaluate their training programs and determine content areas that need instructional improvement (Haladyna & Kramer, 2004).

Use of Subscores

Reliable and valid subscores could be very important for accountability assessments that could be used for diagnostic purposes. Subscores that provide diagnostic information about a skill or a cognitive behavior may lead to possible tailored instruction and remediation for students in classrooms and patients in clinical settings. Therefore, using subscores for important decisions such as diagnostic classifications potentially makes them more important and critical. Although it is useful to obtain and use subscores for making decisions or diagnoses, reasonable subscore performance should be empirically established before reporting subscores (Tate, 2004; Wainer et al., 2001).

Having a provision to report subscores requires a set of conditions that should be met to benefit from additional information derived from subscores. First, subscores should be distinct enough to be more useful and valuable than the total test score. As Haberman (2008) and Haberman, Sinharay, and Puhan (2009) suggested, whether subscores provide distinct information over the total score should be carefully examined before reporting the subscores. If the subscores function very similarly and do not differentiate enough, information obtained from the subscores may be negligible. Second, as total scores, subscores should also have high reliability and adequate psychometric qualities. Standard 5.12 of *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) suggests that test scores should not be reported for test takers unless the validity, comparability, and reliability of such scores have been established. If a subtest measures the intended construct poorly or produces scores that are not consistent over multiple administrations, then the information it yields may not be trustworthy (Monaghan, 2006). Lastly, depending on the scoring method, subscores may be

influenced by factors such as test length, number of subtests, and type of test takers. Considering the relatively small number of items used in subtests, the reliability and validity of subscores should be carefully examined before reporting the subscores to students, parents, schools, and the public.

Estimation of Subscores

The high demand for meaningful and reliable diagnostic information from tests that are not usually designed for this purpose requires robust and reliable scoring methods to enhance the quality of subscore estimates (Boughton, Yao, & Lewis, 2006). Considering the difficulty of obtaining reliable subscores from tests, it is important to use sophisticated and precise scoring methods for the content domains, which can provide a more reliable and valid scoring mechanism. In the literature, there are several classical test theory (CTT) and item response theory (IRT)-based methods for computing subscores from a test consisting of multiple subtests, such as number-correct scores, Kelly's univariate regression (Kelly, 1927, 1947), subscore augmentation (Wainer et al., 2001), the objective performance index (OPI; Yen, 1987), unidimensional IRT (UIRT), and multidimensional IRT (MIRT) scoring.

Summed scores and number-correct scores can be easily computed for each domain and used as estimates of subscores. However, these types of scores were judged unacceptable and inadequate for some stakeholders when the scores were subjected to intense public scrutiny in large-scale testing (Md Desa, 2012). In this approach, examinees with the same number of correct responses in a subtest receive the same subscore regardless of which items they respond to correctly because summed and number-correct scoring treat all items as equally difficult. Therefore, the scores are not

able to reflect examinees' real strengths and weaknesses on the subject areas. Kelly's univariate regression method (Kelly, 1927, 1947), Wainer's subscore augmentation procedure (Wainer et al., 2001), and the OPI method (Yen, 1987) weight the observed scores based on the test reliability and mean scores. Kelly's univariate regression uses CTT-based summed scores and weights them based upon test reliability (e.g., coefficient alpha) while Wainer's subscore augmentation and the OPI method use IRT scale score estimates.

Another approach that is commonly used to estimate subscores in large-scale assessments is to assume an independent unidimensional space for each subtest, and report domain scores as unidimensional IRT scale scores. IRT scale scores obtained from each subtest can be used as an indicator of the true proficiency in the subdomains measured on the test (Bock, Thissen, & Zimowksi, 1997; Hambleton & Jones, 1993; Hambleton & Swaminathan, 1985; Lord, 1980). Scoring in IRT models is neither sample- nor test-dependent, whereas the true-score approach in CTT is specific to the test and the sample of examinees. Also, item parameters and ability estimates obtained from unidimensional IRT models can be used for many psychometric procedures such as equating, linking, item banking, and computerized adaptive testing (Parshall et al., 2001; Wainer & Dorans, 2000).

Despite the advantages of unidimensional IRT models over conventional scoring methods, there are some limitations of these models in terms of subscore estimation. First, each test item is believed to measure a single trait, which is known as a simple structure. Also, subtests are calibrated independently by ignoring the relationship among them. To overcome these issues, other scoring approaches that allow for using complex

test structures have been introduced. MIRT is one of these approaches; it provides an alternative to the limitations in unidimensional scoring methods. The following section gives a brief introduction about the foundations of MIRT.

Application of MIRT

Most educational and psychological tests consist of either tests that measure different constructs or subtests that measure different content domains of a single construct. In these tests, items in a particular subtest are usually designed to measure a single (i.e., unidimensional) ability or latent trait. For instance, the SAT is a test with verbal and mathematics subtests that measure the verbal and mathematical reasoning skills of high school students. Similarly, the National Assessment of Educational Progress (NAEP) consists of mathematics, reading, science, history, and geography subtests, and each of these subtests measures a specific domain. The most common way to obtain information about examinees' abilities in multiple subtests is to apply UIRT for each subtest separately. UIRT models can be used to calibrate items and estimate person scores from the tests that measure a specific ability or proficiency (Hambleton, Swaminathan, & Rogers, 1991).

Assuming the simple structure of the subtests, each subtest can be easily scored with a UIRT model. However, in many instances, due to the lack of a satisfactory index for assessing the dimensionality assumption, the unidimensionality of the test structure may not be clear. The unidimensionality assumption is sometimes violated because items or item sets can measure multiple abilities no matter how carefully the items are constructed (Ackerman, 1992; DeMars, 2006; Reckase, 1985). Because of its design and content, a test item may require test takers to have two or more abilities to respond to the

item correctly. Also, an item can be related to a nuisance or irrelevant skill in addition to the target skill because of the item itself (e.g., contextual effects) or test takers (e.g., DIF). If the unidimensionality assumption does not hold, then the conclusions reached on the basis of a UIRT model may be misleading and summarizing the test performance of a test taker through a single score may not be sensible (Bartolucci, 2007).

Test items that simultaneously measure two or more abilities are called multidimensional in the literature, and they are usually evaluated within the framework of MIRT modeling. MIRT is an extension of UIRT, which relaxes the assumption of unidimensionality by estimating multiple abilities simultaneously and allowing for the inclusion of items that measure multiple abilities or traits. The estimation of item parameters and person abilities within the MIRT framework is not highly popular in testing due to the complexity of multidimensional models and the lack of software that can handle large numbers of items and examinees. However, there are several studies in the literature that have indicated the advantages of MIRT over other methods in estimating item parameters and abilities using simulations and real data (de la Torre, Song & Hong, 2011; Sheng & Wikle, 2007; Wang, Chen, & Cheng, 2004; Yao, 2010; Yao & Boughton, 2007).

MIRT models are flexible and efficient in various test situations. For example, MIRT models can be used for assessments with multiple subtests that have complex item structures. That is, the test includes items that measure two or more abilities (Finch, 2010; Zhang, 2012). Also, MIRT models can benefit from external information — such as the correlation between each ability dimension or other collateral information — when estimating subscores for each dimension. The intentional inclusion of this additional

information in ability estimation yields more precise scores, especially when tests are short and highly correlated (De la Torre, Song & Hong, 2011; Yao, 2010). Furthermore, if subtests are assumed to be related to each other in terms of scoring the items, MIRT models can be used to model compensation for low ability in one dimension by high ability on other dimensions. This type of MIRT model is called a "compensatory" MIRT model (Reckase, 2009). Noncompensatory MIRT models also exist in case one does not anticipate a compensatory relationship between the subscores.

Statement of the Problem

Goodman and Hambleton's (2004) review about the current subscore reporting practices has shown that most states report students' subscores based on raw score, percent correct metrics, or IRT scale scores obtained from the Rasch model. Raw scores (i.e., number of correct responses) or percentages of correct responses are very simple to compute, but they may be disadvantageous for reporting subscores. Since most testing programs consist of subtests based on a small number of items, the estimation of subscores using raw scores may lead to low reliability and precision of the scores (Haberman, 2008; Haberman, Sinharay, & Puhan, 2006; Monaghan, 2006).

To improve reliability and precision of subscores, researchers have proposed alternative methods for estimating persons' scores from subtests (e.g., de la Torre, Song, & Hong, 2011; Haberman & Sinharay, 2010; Kelly, 1947; Skorupski & Carvajal, 2009; Wainer et al., 2001; Yen, 1987). The main purpose of these studies is to propose a scoring method that provides more accurate and reliable test scores on the content domains. Multidimensional ability estimation is one of these methods; it allows for

estimating subscores from tests that have either a simple or complex structure. Multidimensional scoring of subscores is particularly useful when examinees' relative strengths and weaknesses in the different content domains need to be evaluated because of the diagnostic information that they provide (de la Torre, 2009).

Multidimensional ability scoring can incorporate the correlational structure of the latent abilities and ancillary or collateral information of other subtests into the estimation procedure (de la Torre, 2009; Edwards & Vevea, 2006; Wang, Chen, & Cheng, 2004). The correlational structure of the latent abilities refers to the correlation between the estimated latent abilities (i.e., subscores). Ancillary or collateral information can be obtained from any variable correlated with the target ability (de la Torre, 2009; Wang, Chen, & Cheng, 2004). For instance, the multidimensional scoring procedure can borrow ancillary information from other subtests or external variables such as previous test scores or grades. Compared to other estimation methods (e.g., UIRT, OPI, number-correct scoring), employing multidimensional scoring provides ability estimates that have smaller bias and standard error and higher reliability (de la Torre, 2009; de la Torre & Patz, 2005, Yao & Boughton, 2007).

Despite the promising findings from previous MIRT studies about obtaining more reliable subscore estimates, the requirements of MIRT estimation still seem to be in contradiction with the suggestions about when subscores provide valuable diagnostic information. For instance, Sinharay (2010) suggests that in order to have subscores that have added value, subscores should be based on an adequate number of items (at least 20 or more), and they should not be highly correlated (i.e., less than .85). However, multidimensional scoring is found to be advantageous when there are several short

subtests measuring highly correlated abilities that are also highly correlated with the ancillary information sources (de la Torre, 2009; Wang, Chen, & Cheng, 2004, Yao, 2010).

This contradiction causes a paradoxical situation between subscore reliability and the usefulness of subscores. This paradox leads to the question of whether the precision of subscores is more important than the diagnostic information that they provide. Does obtaining more precise estimates from highly correlated but less distinct subscores make the MIRT approach better than other alternative methods? These questions clearly reflect the nature of the relationship between the test and the methodology used to obtain subscores. In order to understand the benefits of MIRT in subscore estimation, both the reliability and added value of subscores should be evaluated together.

Purpose of the Study

As explained earlier, previous studies have revealed very limited amounts of information on the relationship between subscore reliability and how distinctly subscores function in the test. Previous MIRT studies have examined the reliability of subscores through different measures such as the correlation between true and estimated abilities, root mean squared error (RMSE) and bias using simulated data, root mean squared difference (RMSD), standard error, and relative efficiency in real data (e.g., de la Torre, 2009; DeMars, 2006; Wang, Chen & Cheng, 2004; Yao, 2010; Yao & Boughton, 2007). Although these measures are not direct indicators of reliability, researchers have used them as analogous terms to reliability and precision.

Given the need for a reliability index that indicates how reliable and distinct subscore estimates are, this study introduces a new reliability framework based on the variation among the subscores for each examinee. In contrast to other reliability coefficients that solely focus on the precision of individual subscores, the proposed reliability approach divides total subscore variation into within-person and between-person variations. This allows for evaluating not only the consistency of subscores among the examinees but also the distinctiveness of subscores within the examinees.

The main purpose of this study is to compare multidimensional and unidimensional subscore estimation procedures using the reliability framework described above. For multidimensional estimation, a compensatory MIRT model was used to estimate subscores from a test in which each subtest measures a unidimensional trait. This model was specifically chosen to make a direct comparison against the UIRT model, which also assumes that each subtest measures a unidimensional trait. However, the compensatory MIRT model estimates subscores simultaneously using the correlation between subtests as ancillary information, whereas a UIRT model estimates each subscore separately and ignores the relationships between the subtests. Simulated data based on various conditions and real data were used to compare the performances of the MIRT and UIRT scoring approaches in terms of within-person and between-person subscore reliabilities.

The specific research questions related to the performance of UIRT and MIRT for subscore precision are as follows:

1) Does the MIRT model perform better than the UIRT model in terms of within-person and between-person subscore reliability?
2) How are within-person and between-person subscore reliabilities from the UIRT and MIRT models affected by varying data conditions (test length, number of subtests, and correlation between subtests)?
3) How do the MIRT and the UIRT models perform in terms of within-person and between-person subscore reliability in real data?

Summary and Significance of the Study

The significance of this study lies in the fact that diagnostic information obtained from subscores can be very useful for test takers, teachers, and other stakeholders. Therefore, it is important to employ a scoring approach that provides not only reliable but also diagnostically informative subscores. The MIRT framework has received a great deal of attention lately as a method for estimating both item parameters and person abilities. Although simultaneous estimation of the subscores from multiple subtests seems to provide more accurate estimates of abilities, to what extent the estimated subscores provide distinct and useful information is not clearly known yet. Considering the computationally intensive requirements of the existing MIRT models in the estimation of subscores, the question of whether the MIRT framework is worth employing over the computationally simpler UIRT models needs to be addressed. The findings of this study will focus on this important question by illustrating an alternative method for subscore reliability.

In Chapter 2, the foundations of unidimensional and multidimensional IRT frameworks are discussed in detail. Also, this chapter presents a review of the previous studies on subscore estimation methods and subscore reliability. The merits of multidimensional scoring of subscores and challenges related to the estimation process are briefly presented. Furthermore, subscore reliability measures that have been used in previous research are described.

Chapter 3 provides the details of the studied MIRT and UIRT scoring methods employed in this study. Also, the concepts of within-person and between-person reliability are introduced. Then, the simulation studies and real data study are described, and evaluation criteria for the comparison of the MIRT and UIRT scoring approaches are addressed. Chapter 4 demonstrates the results from the simulation studies and real data, and Chapter 5 presents discussion and future directions based on the findings of this study.

CHAPTER 2

LITERATURE REVIEW

Unidimensional IRT (UIRT)

IRT is a psychometric framework that aims to measure the abilities, attitudes, interests, knowledge, or proficiencies of respondents independently from the items and the persons who respond to these items. IRT models focus on the interaction between persons and test items by placing persons' ability estimates and item difficulties on a continuous measurement scale so direct comparisons between respondents' abilities and items are possible (Hambleton, 2000). In IRT, test items are assumed to measure a latent trait (θ) that represents a person's location on the trait scale. When test items intend to measure a single latent trait, unidimensional IRT (UIRT) models are employed. The UIRT models express the probability of a correct response to a test item as a function of θ given one or more parameters of the item. When θ increases, the probability of a correct response to the items also increases monotonically. This monotonic relationship is defined by using item response functions. Figure 2.1 shows three hypothetical items and their relationships with the ability being measured. As the theta increases, the probability of correct response increases. Also, the slope of item characteristic curves represents item discrimination. As the steepness increases, the item discriminates better between examinees with low abilities and examinees with high abilities.

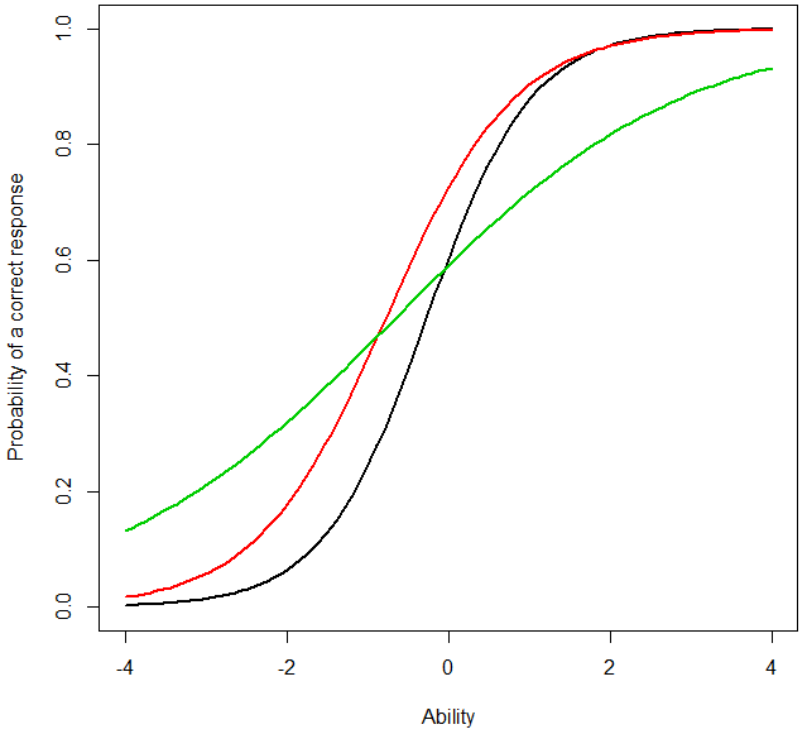

Figure 2.1. Item characteristic curves of three hypothetical test items

Assumptions of UIRT

There are two important assumptions of UIRT models. First, UIRT models require that items on a test should be independent of each other. This assumption is known as conditional local independence. Conditional on the latent trait level, the assumption of local independence requires the probability of responding to an item correctly to be independent of the responses given to the other items on the test. The assumption of conditional local independence has strong and weak versions (Embretson & Reise, 2000). When strong local independence is assumed, the responses to all items on a test should be independent of one another. Unlike strong local independence, weak local independence is met when the item covariances among all pairs of items decrease toward zero as test length approaches infinity. The assumption of conditional local independence is violated when the content or solution of a previous item on a test makes

a latter item easier for the respondents. Violation of this assumption may cause overestimation of test reliability or underestimation of the standard error of the ability estimates (Wainer, 1995; Wainer & Wang, 2000).

The second assumption of the UIRT models is that test items measure a unidimensional latent trait. Within the IRT framework, this assumption is known as the unidimensionality assumption (Embretson & Reise, 2000). In practice, test items may require several traits to obtain a correct response. However, if there is a single dominant trait that accounts for a majority of the variance in the correct responses to a set of items, then the test satisfies the assumption of unidimensionality. UIRT models are appropriate for items that involve a single underlying ability or combination of abilities that are constant across items (Embretson & Reise, 2000). When test items measure more than a single dominant ability, then the assumption of unidimensionality is violated. Violation of this assumption can be a risk to the reliability and validity of the test. For example, unintentional dimensions derived from a test may lead to item bias and differential item functioning (DIF) due to different distributions of construct-irrelevant abilities for different examinee subgroups (e.g., Ackerman & Evans, 1994; Douglas, Roussos, & Stout, 1996; Walker & Beretvas, 2001). Also, an unintended or irrelevant dimension is a potential threat to construct and test score validity because the existence of such a dimension indicates that the test does not merely measure the intended construct. In this case, estimation of ability is confounded because the UIRT model conditions response probabilities on a single ability that is in fact a composite of the target measure and a nuisance construct.

UIRT Models

There are several UIRT models based on the characteristics of item parameters. The simplest IRT model is the one-parameter logistic (1PL) model. The 1PL model assumes that all of the items have the same item discrimination, and the lower asymptote is assumed to approach zero. This model is as follows:

$$P\{X_{ij} = 1 | \theta_i, a, b_j\} = \frac{\exp[a(\theta_i - b_j)]}{1 + \exp[a(\theta_i - b_j)]} \quad (2.1)$$

where $P\{X_{ij} = 1 | \theta_i, a_j, b_j\}$ is the probability of an examinee *i* with ability θ_i answering item *j* correctly, b_j is the difficulty parameter of item *j*, *a* is a constant item discrimination parameter for all items, and θ_i is the ability level of examinee *i*. The Rasch model (Rasch, 1960) is a variant of 1PL model. However, item discrimination is not estimated in the Rasch model whereas the 1PL model estimates a constant discrimination parameter for all items. Depending upon the software used for estimating the Rasch model, theta scale can be fixed either on the average item location or on the average person location.

In contrast to the 1PL and the Rasch models that are restrictive in terms of item discrimination, the two-parameter logistic (2PL) model relaxes the restrictive discrimination assumption in the model. The 2PL model is as follows:

$$P\{X_{ij} = 1 | \theta_i, a_j, b_j\} = \frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} \quad (2.2)$$

where a_j is the discrimination parameter and varies across items, and the other model parameters can be interpreted as those presented for the 1PL model. Discrimination parameters typically range from 0 to 2 in the normal ogive metric and from 0 to 3.4 in the

logistic metric (Hambleton & Swaminathan, 1985), with high values being more effective for discriminating between respondents with low and high trait levels.

In multiple choice items, it's possible for examinees to guess the items correctly. The 1PL and 2PL models assume that item response functions have a zero lower asymptote (which is sometimes referred to as the guessing parameter). The three-parameter logistic (3PL) model includes a lower asymptote parameter, which is especially useful for multiple choice and other selected-response items. Among the UIRT models, the 3PL is the most popular (Kolen & Brennan, 2004). The 3PL model can be expressed as follows:

$$P\{X_{ij} = 1 | \theta_i, a_j, b_j, c_j\} = c_j + (1 - c_j)\frac{\exp[a_j(\theta_i - b_j)]}{1 + \exp[a_j(\theta_i - b_j)]} \quad (2.3)$$

where c_j is the lower asymptote for item j. Estimation of a unique lower asymptote for each item on a test can lead to some estimation problems. Because in most cases, there are only few test takers whose locations on the theta scale are in the region of the lower asymptote, there are few data points from which to estimate the location of the c parameter. Thus, a common guessing parameter is often assumed for all items or for groups of similar items (Embretson & Reise, 2000; Han, 2012).

Multidimensional IRT (MIRT)

Educational and psychological assessments have been consistently found to be more complex than intended. When designing an assessment or a test, items or tasks that are associated with a certain skill or ability are included in the assessment. For such assessments, a scoring method — such as UIRT — can be employed because the items

are thought to measure a single skill or ability. However, previous studies have shown that the unidimensionality assumption is often violated in real-world contexts (Ackerman, 1994; Nandakumar, 1994; Reckase, 1985), and the number of dimensions is underestimated (Reckase & Hirsh, 1991). When the items require multiple abilities to obtain a correct response, using a scoring approach based on the assumption of unidimensionality increases errors of measurement and the chances of making incorrect inferences about a student's proficiency in a given subject (Walker & Beretvas, 2003).

Previous studies that have examined the impact of using unidimensional IRT models for test items that are not strictly unidimensional suggest that if there is another strong dimension in the test beyond the major dimension being measured, unidimensional estimates of items and abilities may be drawn towards the secondary dimension. Thus, item parameter estimates would be biased, and the standard error estimates associated with ability estimates falsely become very small (Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Reckase, 1979; Way, Ansley, & Forsyth 1988). The use of a multidimensional IRT (MIRT) model can be helpful for addressing these issues due to the intended or unintended multidimensionality of test items.

A review of many forms of assessment and the associated scoring methods implies that MIRT is a promising framework for accounting for examinees' test performances adequately in simple and complex test structures (van der Linden & Hambleton, 1997, p. 221). Reckase (1997) defined MIRT as either an extension of item response theory applied to multidimensional data, or as a special case of confirmatory factor analysis. MIRT can deal with complex items and assessments by introducing ability and item discrimination parameters for each skill being measured by a test

question, and modeling the interaction between examinees and test items (Ackerman, 1992; Reckase, 1997). Modeling examinees' responses in a multidimensional manner allows for making separate inferences about each skill or ability being measured on the test (Walker & Beretvas, 2000). Furthermore, MIRT models can be used for both exploratory and confirmatory purposes (Embretson & Reise, 2000). Exploratory MIRT models can be used for determining the underlying dimensions of a test, and also checking the unidimensionality assumption. When the number of dimensions and the items defining each dimension are known, confirmatory MIRT models can be used for estimating item and person parameters for specific dimensions.

Common MIRT Models

Several multidimensional IRT models for dichotomous and polytomous responses have been proposed (Bock & Aitkin, 1981; Bock & Lieberman, 1970; McDonald, 1985; Mulaik, 1972; Sympson, 1978; Whitely, 1980). The most common MIRT models are summarized below.

Multidimensional Random Coefficients Multinomial Logit Model (MRCML).

The MRCML model is an extension of the Rasch family of item response models. It assumes that for an item j with ordered categories of response indexed by k, there corresponds a unique dimension among a larger set of possible dimensions denoted by m ($m = 1,\ldots, M$). For the presentation of the MRCML model, the notation developed in Adams, Wilson, and Wang (1997) will be used. Let items be indexed $j = 1, ..., N$ with each item having $K_j + 1$ possible response categories ($k = 0, 1, \ldots, K_i$). The random variable X_{jk} is introduced such that: $X_{jk} = 1$, if the response to item is in category or $X_{jk} = 0$, otherwise. The MRCML model can be written at the item category level as:

$$P(X_{ik} = 1; a, b, \boldsymbol{\xi}|\boldsymbol{\theta}) = \frac{\exp(b_{jk}\boldsymbol{\theta_i} + a'_{jk}\xi)}{\sum_{k=1}^{Kj} \exp(b_{jk}\boldsymbol{\theta_i} + a'_{jk}\xi)} \quad (2.4)$$

where $\boldsymbol{\theta_i}$ is an *M x 1* column vector with *M* corresponding to the number of hypothesized dimensions in a given instrument. Item and category parameters represented by δ_{ik} have been gathered into the vector ξ in this equation.

Multidimensional Two-Parameter Compensatory Logistic Model (MC2PL). In this model, the probability of a correct response to item *j* can be expressed using the M-dimensional compensatory two-parameter logistic model (Reckase, 1985) as:

$$P\left(X_{ij} = 1 \middle| \boldsymbol{\theta_{ip}}, \boldsymbol{a_j}, d_j\right) = \frac{\exp(\sum_{p=1}^{P} \boldsymbol{a_j}\boldsymbol{\theta_{ip}} + d_j)}{1 + \exp(\sum_{p=1}^{P} \boldsymbol{a_j}\boldsymbol{\theta_{ip}} + d_j)} \quad (2.5)$$

where X_{ij} represents the score (0,1) on item *j* person *i*, $\boldsymbol{a_j}$ represents a vector of multiple discrimination parameters associated with item *j*, d_j represents a scalar difficulty parameter of item *j*, and $\boldsymbol{\theta_p}$ ($\boldsymbol{\theta}_\mathbf{p}$ = {θ_1,...., θ_i} is the vector of ability parameters.

Multidimensional Three-Parameter Logistic Model (M3PL). The multidimensional 3-parameter logistic (M3PL) model (Reckase, 1985; Ackerman, 1996) is a compensatory MIRT model where decreasing an examinee's ability along one dimension can be offset by increasing ability along another dimension. Let $\boldsymbol{\theta}_\mathbf{p}$ = (θ_1,...., θ_i) denote the vector of abilities on a *J*-dimensional space. Let X_{ij}be an indicator variable such that $X_{ij} = 1$ if a given examinee responds correctly to item *j* and $X_{ij} = 0$ otherwise. Under the M3PL model, the probability of a correct response to item j given an examinee of ability (θ) is:

$$P\left(X_{ij}=1\middle|\boldsymbol{\theta_{ip}},\boldsymbol{a_j},d_j,c_j\right)=c_j+\left(1-c_j\right)\frac{\exp(\sum_{p=1}^{P}\boldsymbol{a_j\theta_{ip}}+d_j)}{1+\exp(\sum_{p=1}^{P}\boldsymbol{a_j\theta_{ip}}+d_j)} \tag{2.6}$$

where $\mathbf{a_i}$ is again a vector of multiple discrimination parameters, d_i is the difficulty parameter, and c_i is the lower asymptote for the item j. Although d_j is called the difficulty parameter, the higher value of d_j indicates the item being easier. The value of d_j can be computed as $d_j = a_j b_j$. In addition to the logistic form of the M3PL model, the normal ogive version of this model can be written using the same components in Equation 2.6 (Bock, Gibbons, and Muraki, 1988):

$$P\left(X_{ip}=1|\boldsymbol{\theta_p},\boldsymbol{a_i},d_i,c_i\right)=c_i+(1-c_i)\Phi(a_i\boldsymbol{\theta}_p-b_i), \tag{2.7}$$

where Φ is the cumulative distribution function of the Normal (Gaussian) distribution.

Multidimensional Two-Parameter Partial Credit Model (M-2PPC). For a polytomously scored item j, the probability of a response k − 1 to item j for an examinee with ability ($\bar{\theta}_i$) is given by the multidimensional version of the partial credit model (Yao & Schwarz, 2006):

$$P_{ijk}=P\left(x_{ij}=k-1\middle|\vec{\theta}_i,\vec{\beta}_j\right)=\beta_{3j}+\frac{e^{(k-1)\vec{\beta}_{2j}\odot\vec{\theta}_i^{\,T}-\sum_{t=1}^{k}\beta_{\delta tj}}}{\sum_{m=1}^{Kj}e^{((m-1)\vec{\beta}_{2j}\odot\vec{\theta}_i^{\,T}-\sum_{t=1}^{k}\beta_{\delta tj})}} \tag{2.8}$$

where x_{ij} = 0, ……., K_j-1 is the response of examinee i to item j, $\vec{\beta}_{2j}=(\beta_{2j1},\ldots\ldots,\beta_{2jD})$ is a vector of dimension D for item discrimination parameters, $\beta_{\delta kj}$ for k=1, 2, ….., K_j are the threshold parameters, $\beta_{\delta 1j}$= 0, and K_j is the number of response categories for the j^{th} item, and $\vec{\beta}_{2j}\odot\vec{\theta}_i^{\,T}=\sum_{l=1}^{D}\beta_{2jl}\theta_{il}$. The parameters for the j^{th} item become $\vec{\beta}_j=(\vec{\beta}_{2j},\beta_{\delta 2j},\ldots\ldots,\beta_{\delta K_j j})$.

Within-item and Between-item Multidimensionality

MIRT models are divided into two groups in terms of test structure. These are between-item models and within-item models (Adams, Wilson & Wang, 1997; Wang, Chen, & Cheng, 2004). Multidimensional between-item models focus on test structures where subtests are mutually exclusive and measure different latent variables. This type of test structure is also known as "simple structure" because each item is only associated with a single latent dimension. In contrast to between-item models, multidimensional within-item models are appropriate for test structures where items can be an indicator of multiple latent dimensions. This type of test structure is known as a "nonsimple structure" or "complex structure." Confirmatory models with complex structures or bi-factor models can be examples of within-multidimensional MIRT models. Figure 2.2 shows a graphical illustration of between-item and within-item test structures.

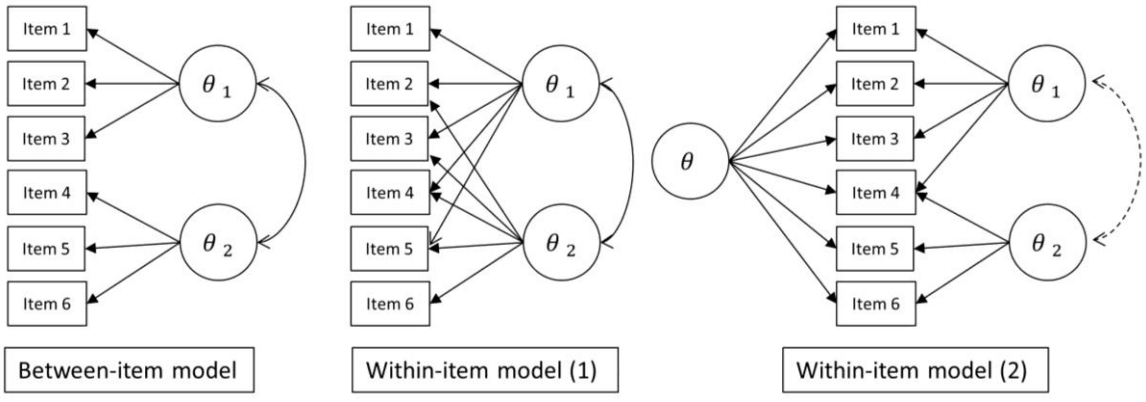

Figure 2.2 An example of between-item and within-item models (θ_1 & θ_2 = Domain abilities, θ = Overall ability)

In multidimensional between-item models, the test contains several subscales that measure related but distinct latent dimensions. These types of MIRT models are also known as multi-unidimensional models in the literature because each subtest is treated as unidimensional while the overall test structure becomes multidimensional (Sheng & Wikle, 2007, 2008). An example of a multi-unidimensional test is the Trends in International Mathematics and Science Study (TIMSS). TIMMS consists of mathematics and science subtests, and each of these subtests includes several content and cognitive domains. For example, the mathematics subtest includes several content domains (e.g., algebra, geometry, number, and data display) and three cognitive domains (knowing, applying, and reasoning). Each of these domains is treated as a unidimensional subtest to report raw scores and percent correct but the overall test is still assumed to measure a single construct, i.e. mathematics. Similarly, the Armed Services Vocational Aptitude Battery (ASVAB) is a multiple-aptitude test that measures developed abilities and helps predict future academic and occupational success in the military. The ASVAB tests are designed to measure four domains (verbal, math, science and technical, and spatial). Each domain also includes several subdomains that are used to obtain an overall measure of the domain.

The most common way of using IRT to analyze multi-unidimensional structures is to estimate item parameters and person abilities from each subtest separately with a UIRT model or to treat the whole test as unidimensional. Both of these methods have certain weaknesses that make them less desirable than undertaking a multidimensional calibration (see Adams, Wilson, & Wang, 1997; Zhang, 2012). When a test measures more than one latent dimension, and some of the test items require multiple abilities to be

responded to correctly, then the test displays within-item multidimensionality. Models that incorporate within-item multidimensionality are suitable for modeling interactions between different abilities and task demands. Here, the probability of solving an item can be modeled as a function of a combination of different dimensions of abilities. Hence, within-item multidimensional models imply explicit assumptions about the abilities required for the different items, which necessitate strong theoretical assumptions. Models with within-item multidimensionality are particularly interesting for modeling performance in complex tasks that cannot be explained by a single ability dimension for each task (Hartig & Hohler, 2009).

Compensatory and Noncompensatory MIRT Models

In addition to test structure, MIRT models can also be divided into two branches based on the presence of a compensatory relationship between the dimensions. MIRT models can be either compensatory or noncompensatory. Compensatory MIRT is additive in nature and therefore a respondent who happens to be weak in one dimension can make up for or compensate for this weakness with strength in another measured dimension (Reckase, 1997). For example, a child who is familiar with baseball but has poor reading skills may perform well on a test that requires him to read a passage on playing baseball and then write a brief essay about the reading passage. The compensatory model allows the dimensions to interact, with a high ability on one dimension compensating for a lower ability on a second dimension (Yao & Boughton, 2007). The noncompensatory model, which is also known as the partially compensatory model, is multiplicative in nature. Therefore, a respondent who is weak in one area cannot make up for this weakness by having strength in another area. Within the

noncompensatory version of MIRT, one must be proficient in both abilities to obtain a higher score.

The main difference between compensatory and noncompensatory models is the mechanism for computing the total probability of a correct response. The compensatory model of M3PL in Equation 2.6 sums the probabilities from a series of θ values to obtain the overall probability. Unlike compensatory models, noncompensatory models use a multiplication procedure to compute the overall probability. Each dimension in the model has a separate probability, and these probabilities are multiplied to find the overall probability for responding to an item correctly (Reckase & McKinley, 1982). The noncompensatory model for dichotomous responses was described by Sympson (1978) and Whitely (1991) as follows:

$$P\left(X_{ij}=1\middle|\boldsymbol{\theta_{ip}},\mathbf{a_j},\mathbf{b_j},c_j\right)=c_j+\left(1-c_j\right)\prod_{p=1}^{\mathrm{P}}\frac{exp\left[\mathbf{a_j}(\boldsymbol{\theta_{ip}}-\mathbf{b_j})\right]}{1+exp\left[\mathbf{a_j}(\boldsymbol{\theta_{ip}}-\mathbf{b_j})\right]}. \quad (2.9)$$

In contrast to the compensatory model, this model includes separate difficulty parameters for each dimension. Instead of the scalar difficulty (d_j) in Equation 2.6, there is a vector of difficulties in Equation 2.9 ($\mathbf{b_j}=b_{j1},\ldots..,b_{jp}$) for each item. This model does not allow for compensating a low ability using a high ability. For the noncompensatory model, high probability means high ability for all dimensions (Reckase, 1997, 2009).

Figure 2.3 demonstrates a graphical illustration of multidimensional compensatory and noncompensatory models. The surface plot for the compensatory MIRT model indicates that the probability increases as both dimensions increase. A low

level of dimension 1 is compensated by dimension 2. The same relationship is true for a low level of dimension 2 as well. The contour plot also shows that low ability on dimension 1 (e.g. $\theta_1=0$) and high ability on dimension 2 (e.g. $\theta_2=2$) still lead to a high probability (i.e., above .8). In contrast to the compensatory model, the surface plot for the noncompensatory model shows that the probability increases slowly as abilities on dimension 1 and 2 increase. The contour plot also shows that a person needs to have high ability levels on both dimensions to have a high probability of responding to an item correctly. There is no compensation between the dimensions.

Compensatory models can be appropriate for items having disjunctive component processes (Maris, 1999). For example, items having multiple solution strategies (Reckase, 1997) likely possess compensatory multidimensionality as a deficiency in one ability (i.e., skill with one strategy) naturally compensates for the other (i.e., skill with a different strategy). By contrast, noncompensatory models may be appropriate for items that have conjunctive component processes (Maris, 1999). For example, a word problem on a mathematics test may require reading ability to interpret the question and then mathematics ability to solve it. For such an item, it is unlikely that either ability will be able to compensate for a lack of the other.

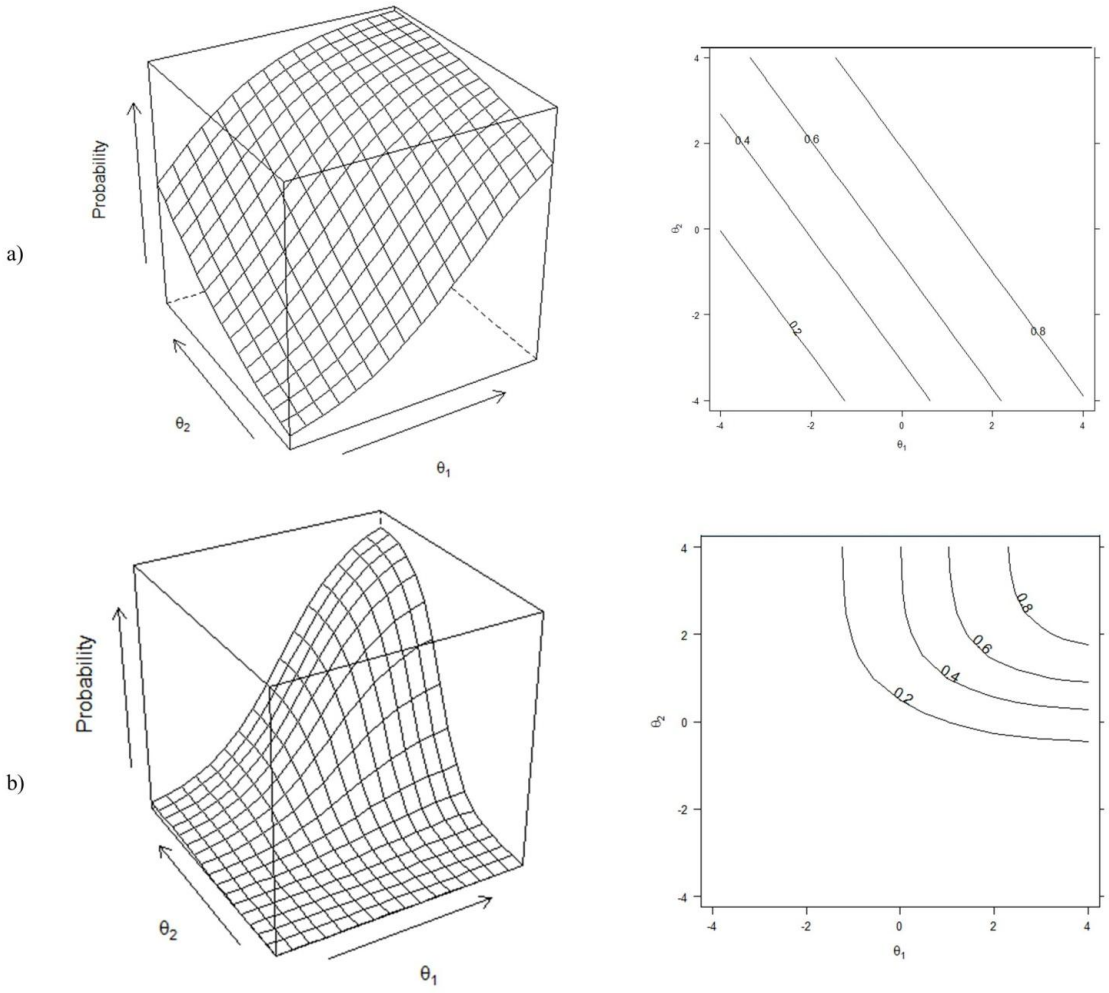

Figure 2.3. Item response surface and contour plots for two dimensional compensatory (a) and noncompensatory (b) models.

MIRT Estimation Programs

There are two well-known programs for estimating MIRT models with dichotomous data; TESTFACT (Wilson, Wood, & Gibbons, 1998) and NOHARM (Fraser, 1987). TESTFACT is a program that was designed to perform a non-linear, exploratory full information factor analysis on dichotomous item responses. As an exploratory program, TESTFACT does not allow for specifying prior restrictions on item

parameters and relating the items to predefined dimensions. However, bi-factor MIRT models can be estimated as a confirmatory model in TESTFACT. The program sets the dimensional structure of individual items based upon the number of traits, defined a priori, that contribute to their responses (McDonald, 1999). TESTFACT uses full-information marginal maximum likelihood (MML) in combination with an expectation-maximization (EM) algorithm to estimate the item parameters (Bock & Atkin, 1981; Bock, Gibbons, & Muraki, 1988; Gibbons & Hedeker, 1992). In the estimation procedure, it is assumed that persons represent a random sample from the population, and their latent trait levels come from a normal distribution with mean zero and a standard deviation of one. Because TESTFACT cannot estimate the c parameter (i.e., lower asymptote), users need to enter either a fixed value of a vector of pre-specified c parameters.

NOHARM, which stands for Normal Ogive Harmonic Analysis Robust Method, uses a polynomial approximation procedure (McDonald, 1997; 1999). The program aims to minimize the difference between observed values and expected values under the selected IRT model. In contrast to TESTFACT, NOHARM does not rely upon observed response vectors to estimate item parameters, and thus is not a full-information factor analytic procedure. It utilizes the information given by the pairwise proportions of an examinee successfully answering any two given items (Knol & Berger, 1988). An unweighted least squares (ULS) function of the difference between observed pairwise proportions and computed expected pairwise proportions is minimized through an iterative process. The item parameters that minimize the ULS function are used as the final parameter estimates. There are two important restrictions of NOHARM. Like

TESFACT, NOHARM cannot estimate the lower asymptote. The lower asymptote can be pre-specified based on a fixed value or a vector. Second, NOHARM can only estimate item parameters. So, it does not provide estimates of person abilities.

ACER ConQuest (Wu, Adams, & Wilson, 1998) is another MIRT program that can estimate multidimensional item responses and latent regression models. ConQuest can fit several unidimensional and multidimensional models such as Rasch model, the 1PL model, Andrich's (1978) Rating Scale Model, Masters' (1982) Partial Credit Model, generalized unidimensional models, and multidimensional item response models. ConQuest allows for estimating both within-item and between-item models. Users can add restrictions on the items and latent dimensions (e.g. uncorrelated latent dimensions). ConQuest can estimate both item parameters and person abilities.

A recently developed IRT program, IRTPRO (Cai, Thissen, & du Toit, 2011), is capable of fitting various IRT models to dichotomously and polytomously scored items. The program can estimate multidimensional versions of several IRT models (e.g., 1PL, 2PL, 3PL, rating scale model, partial credit model, graded response model). Users can specify different constraints in item parameter and ability estimation procedures. Differently from other IRT programs, IRTPRO provides multiple approaches to estimate structural parameters of a model. These approaches are Bock–Aitkin approach with expectation–maximization algorithm (BAEM; Bock & Aitkin, 1981), adaptive quadrature (ADQ; Schilling & Bock, 2005), and Metropolis–Hastings Robbins–Monro (MH-RM; Cai, 2010). For ability estimation, IRTPRO provides three estimation methods: EAP, maximum a posteriori (MAP; or Bayes modal estimator), and EAP for summed scores (Thissen, Nelson, Rosa, & McLeod, 2001).

BMIRT (Yao, 2003) is a relatively newer program that uses the Bayesian framework. BMIRT adopts a Markov Chain Monte Carlo (MCMC) method to estimate item and ability parameters in the multidimensional IRT framework. The program can be used for dichotomous and polytomous data that are multidimensional in nature. BMIRT supports both exploratory and confirmatory MIRT models. Multi-unidimensional models, within-item models, bi-factor models, diagnostic classification models, and higher-order IRT models can be estimated with BMIRT to obtain domain scores and overall test scores. For subscore estimation, BMIRT includes various ability estimators such as maximum likelihood (MLE), maximum a posteriori estimation (MAP), and MCMC estimation.

In addition to the programs mentioned above, Mplus (Muthén & Muthén, 1998-2011), some macros in SAS (e.g. PROC NLMIXED), and STATA (e.g. GLLAMM) can be used for estimating MIRT models. MIRT parameterization programs for tests with mixtures of dichotomous and polytomous items are also available, such as POLYFACT (Muraki, 1999), which uses marginal maximum likelihood (MML) estimation, and MicroFACT (Waller, 2002), which, like TESTFACT, employs exploratory factor analysis.

Subscore Estimation

There are several techniques for estimating subscores or domain scores from tests that consist of multiple subtests. Both CTT-based and IRT-based techniques exist to estimate and report subscores from the tests. These estimation techniques can be grouped into two categories based on whether they take the multidimensional structure of the test

into account when estimating subscores. These categories are unidimensional subscore estimation and multidimensional subscore estimation. This section provides a brief description of unidimensional and multidimensional subscore estimation techniques.

Unidimensional Estimation of Subscores

Kelley's regressed score method (1927), Wainer et al. (2001)'s multivariate empirical Bayes estimation method, and objective performance index scoring (OPI; Yen, 1987) are common methods for estimating subscores that are unidimensional in nature. In addition, other techniques based on the use of unidimensional IRT for the scoring of subscales or domains (Bock, Thissen, & Zimowski, 1997) and the subscore augmentation within the CTT framework (Haberman, 2008) have been demonstrated in the literature.

Kelley's Regressed Score Method. Kelley's regressed score method (Kelley, 1927, 1947) is based on weighting the observed subscores based on the group mean. Using the CTT notation, it can be written as follows:

$$\hat{\tau} = \rho x + (1 - \rho)\mu, \tag{2.10}$$

where $\hat{\tau}$ represents an estimate of true score (τ), x is the observed score, μ is the group mean, and ρ is the reliability of the test.

This method aims to improve the estimate of true score through the shrinkage in the observed score toward the group mean by an equal amount of reliability. Based on Equation 2.10, when the test is very reliable, the impact of the observed score becomes very dominant on the estimate of the true score. That is, a reliable observed score is assumed to be a very precise approximation of the true score. However, when the test is not highly reliable, the estimate of the true score shrinks toward the group mean to

remove the unreliable part of the observed score. So, it can be said that Kelley's regressed score method improves the precision of test scores by using the group mean as ancillary information.

Considering a test with several subtests, Kelley's regressed score method can be applied to the subscores simultaneously. Instead of using a reliability index and a group mean for a single test within a univariate design, the same estimation procedure can be generalized for multivariate cases. When Equation 2.10 is rearranged, it can be written as:

$$\hat{\tau} = \mu + \rho(x - \mu). \tag{2.11}$$

In a multivariate case with subscores from multiple subtests, Equation 2.11 can be written with a compact matrix notation as follows:

$$\hat{\tau} = \boldsymbol{\mu} + \boldsymbol{\beta}(\boldsymbol{x} - \boldsymbol{\mu}), \tag{2.12}$$

where $\boldsymbol{\mu}$ is a vector of subtest means, $\boldsymbol{x}$ is a vector of subscores, and $\boldsymbol{\beta}$ is a matrix of the reliability indices for each subtest.

Objective Performance Index (OPI). The Objective Performance Index (OPI) is an estimated true score for the items in an objective based on the performance of a given examinee (Yen, 1987). The OPI scoring method incorporates information from the total test score using an empirical Bayes procedure. For computing the OPI scores, it is assumed that each test consists of *n* items based on *J* objectives, with n_j items in objective *J*. Each item on the test is assumed to contribute to one objective at most. Assuming that X_{ij} is either the observed number-correct score or the unidimensional IRT estimate of person *i* on objective *j*, T_{ij}equals to $E(X_{ij}/n_j)$, and X_{ij} follows a binomial distribution

given T_{ij}, i.e., $p(X_{ij} = x_{ij}|T_{ij})$, the OPI can be computed as follows:

$$\hat{\mathrm{T}}_{\mathrm{ij}} = \frac{1}{n_j}\sum_{j=1}^{J} P_{ij}(\hat{\theta}), \tag{2.13}$$

where $P_{ij}(\hat{\theta})$ is the unidimensional ability estimate for person i on objective j. By using $\hat{T}_{ij}$, two additional values are computed:

$$p_{ij} = \hat{T}_{ij} n_j + x_{ij} \tag{2.14}$$

and

$$q_{ij} = [1 - \hat{T}_{ij}] n_j + n_j - x_{ij}. \tag{2.15}$$

Using these two values, the OPI is defined as:

$$\tilde{T}_{ij} = \frac{p_{ij}}{p_{ij} + q_{ij}}. \tag{2.16}$$

As previously mentioned, the OPI can use the prior information to gather more accurate estimates of subscores. For more detailed information about incorporating additional information into the prior distribution of T_{ij}, see Yen (1987).

Subscore Augmentation. The subscore augmentation procedure proposed by Wainer et al. (2001) is basically a multivariate generalization of Kelley's regressed score method. Subscore augmentation is also based on the empirical Bayes estimates of the subscores using a weighting procedure. The main distinction between the two methods is

that Wainer et al.'s (2001) subscore augmentation can use either the number-correct scores or the IRT scale score estimates as the observed score.

Subscore augmentation may be thought of as a multi-stage estimation procedure for proficiency estimates for the domains (Thissen & Edwards, 2005). In the first stage, unidimensional IRT ability estimates are obtained using one of maximum likelihood (MLE), maximum a posteriori (MAP), or expected a posteriori (EAP) methods. The values of MLE(θ), EAP(θ), and MAP(θ) in the augmentation procedure correspond to the regressed estimates in Kelley's method. Second, an IRT-based estimate of reliability is computed in conjunction with the observed covariance matrix among the unidimensional IRT ability estimates. In the final stage, the estimates of the IRT scale scores are regressed on all subscores and weighted using the IRT-based estimate of reliability.

When Equation 2.12 for Kelley's method is rearranged using the IRT scale scores, the augmented subscores are as follows:

$$\begin{aligned} \mathrm{MLE}(\hat{\theta}) &= \overline{\mathrm{MLE}(\theta)} + \rho\big(\mathrm{MLE}(\theta) - \overline{\mathrm{MLE}(\theta)}\big), \\ \mathrm{EAP}(\hat{\theta}) &= \overline{\mathrm{EAP}(\theta)} + \rho\big(\mathrm{EAP}(\theta) - \overline{\mathrm{EAP}(\theta)}\big), \\ \mathrm{MAP}\big(\hat{\theta}\big) &= \overline{\mathrm{MAP}(\theta)} + \rho\big(\mathrm{MAP}(\theta) - \overline{\mathrm{MAP}(\theta)}\big). \end{aligned} \tag{2.17}$$

It should be noted that the reliability value (ρ) in Equation 2.18 is not a CTT-based index of reliability (e.g., alpha coefficient, split-halt reliability) anymore. Rather, ρ is a marginal reliability of theta (Green et al., 1984), and it can be computed for a particular subscale as follows:

$$\hat{\rho} = 1 - \bar{\sigma}_e^2, \tag{2.18}$$

where $\bar{\sigma}_e^2$ represents the average error variance of the estimated abilities.

As an approximation of the reliability within the CTT framework, Wainer et al. (2001) suggested using the ratio of unconditional true score variance to unconditional estimated true score variance as an estimate of reliability. The unconditional true score variance for the kth subscore is the kth diagonal element of the variance-covariance matrix (S) below: (Skorupski, 2008):

$$\mathbf{A} = \mathbf{S}^{\mathbf{true}}(\mathbf{S}^{\mathbf{obs}})^{-\mathbf{1}}\mathbf{S}^{\mathbf{true}}(\mathbf{S}^{\mathbf{obs}})^{-\mathbf{1}}\mathbf{S}^{\mathbf{true}}. \quad (2.19)$$

Similarly, the unconditional estimated true score variance for the kth subscore is the kth diagonal element of the matrix:

$$\mathbf{C} = \mathbf{S}^{\mathbf{true}}(\mathbf{S}^{\mathbf{obs}})^{-\mathbf{1}}\mathbf{S}^{\mathbf{true}}. \quad (2.20)$$

Using these two matrices, the estimate of the reliability of the kth subscore can be computed as follows:

$$\rho_{kk} = \frac{a_{kk}}{c_{kk}} = \frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{true}} + \sigma^2_{\text{error}}}, \quad (2.21)$$

where σ^2_{true} is true score variance of the kth subscore, and σ^2_{error} is residual or error variance for the kth subscore.

Although the Wainer et al.'s (2001) subscore augmentation method is based on the unidimensional IRT estimates of subscores, the method becomes isomorphic to the MIRT approach when a simple structure (i.e., multi-unidimensional structure) is present (Thissen & Edwards, 2005).

Unidimensional IRT Subscoring. Bock, Thissen, and Zimowski (1997) proposed an IRT-based subscore estimation that produced more accurate estimates than the number-correct scoring method. Assuming that examinees' IRT-based subscores (θ) are estimated from an item bank with *n* items using a well-fitting IRT model, the subscores on the IRT metric can be transformed to domain expected number-correct scores as follows:

$$d(\hat{\theta}) = \frac{\sum_{j=1}^{n} w_j P_j(\hat{\theta})}{\sum_{j=1}^{n} w_j}, \tag{2.22}$$

where w_j is the sampling weight, and $P_j(\hat{\theta})$ is the response function of item j in the item bank. When $d(\hat{\theta})$ is rescaled for obtaining the domain percent-correct score, it becomes:

$$D(\hat{\theta}) = \frac{100 d(\hat{\theta})}{n}. \tag{2.23}$$

The advantage of this approach is that it does not require the test items to be a random sample of the domain, and standard errors of the domain scores can be easily computed. Bock et al. (1997) suggest that this method can be used for mastery and diagnostic classifications in the context of student qualification and accountability assessments rather than selection and ranking purposes.

Haberman Augmentation. Haberman (2008) and Haberman, Sinharay, and Puhan (2009) proposed a method based on CTT to estimate subscores and determine whether they provide any added value over total scores. Haberman and Sinharay (2010) refer to this approach as Haberman augmentation. Consider a test with $q \geq 2$ correctly answered items taken by a sample of $n \geq 2$ examinees. If examinee i ($1 \leq i \leq n$) responds

to item *j* ($1 \leq j \leq q$) correctly, his/her score for this item (X_{ij}) becomes 1, and 0 otherwise. Examinees' responses to the items are independent from each other and identically distributed. The test is also assumed to measure more than one skill based on a simple structure of the items. Based on these assumptions, the total raw score of examinee *i* is

$$S_i = \sum_{j=1}^{q} X_{ij}, \tag{2.24}$$

and the raw subscore on the subtest *k* is

$$S_{ik} = \sum_{j \in J(k)} X_{ijk,} \tag{2.25}$$

where *J(k)* is a subtest that measures skill k ($1 \leq k \leq r$). The subscore, S_{ik}, ranges from 0 to *q(k)*. The true total score corresponding to S_i is T_i, and the true subscore corresponding to S_{ik} is T_{ik}. Using the observed scores and true scores defined above, Haberman (2008) and Haberman et al. (2008) define three ways to estimate subscores through a linear combination:

a) $U_{iks} = \alpha_{ks} + \beta_{ks} S_{ik}$, based on the observed subscore S_{ik}.

b) $U_{ikx} = \alpha_{kx} + \beta_{kx} S_i$, based on the observed total score S_i.

c) $U_{ikc} = \alpha_{kc} + \beta_{k1c} S_i + \beta_{k2c} S_{ik}$, based on the total raw score S_i and the subscore S_{ik}.

For all of these subscore estimations, α refers to either subscore reliability or total score reliability. In addition to the subscore estimates shown above, Haberman and Sinharay (2010) suggest that an augmented subscore based on all the raw subscores (Wainer et al., 2001) can be computed as $U_{ika} = \alpha_{ka} + \sum_{k'=1}^{r} \beta_{kk'a} S_{ik'}$. More detailed description about

Haberman's subscore augmentation and mean squared errors for the estimated subscores can be found in Haberman (2008) and Haberman et al. (2009).

Multidimensional IRT Subscore Estimation

The subscore estimation approaches described in the previous section aim to estimate each subscore or domain score from a test one by one. As an alternative approach, MIRT models can be also employed to report subscores (e.g., Beguin & Glas, 2001; de la Torre & Patz, 2005; Reckase, 1997, 2007; Yao & Boughton, 2007). There are certain advantages of using MIRT models over other subscore estimation methods. First, unlike unidimensional subscoring methods, MIRT does not require a test based on a simple structure to estimate the subscores. MIRT models can estimate subscores from both simple and complex test structures. Second, despite its computational complexity, MIRT is more straightforward than other subscoring methods that use an empirical Bayes procedure to borrow information from external variables. The subscoring methods such as the OPI (Yen, 1987) and subscore augmentation (Wainer et al., 2001) require a multi-stage procedure in which unidimensional subscores are first estimated, and then ancillary information (e.g., group mean, test reliability, examinees' previous scores) are used to weight the estimated subscores to improve the precision. However, MIRT models can estimate accurate subscores by using ancillary information such as the correlation between the subscores within a single estimation process.

Several types of estimators can be used for estimation subscores from the MIRT models. These estimators can be either Bayesian (e.g., EAP, and MAP) or non-Bayesian (e.g., MLE) Also, MCMC techniques can be employed for estimating the subscores from a MIRT model although it is computationally intensive and inconvenient for testing

programs with large sample sizes (de la Torre & Patz, 2005; Sheng, 2005; Yao & Boughton, 2007). The MLE method estimates the subscores by maximizing the likelihood of an examinee's item responses. This method fails when an examinee responds to all items correctly or incorrectly (Embretson & Reise, 2000).

Both EAP and MAP are based on the Bayesian perspective. The MAP method estimates the subscores by maximizing a posterior distribution based on prior information about the subscores with the likelihood function. The MAP method allows for estimating scores from all possible response patterns (e.g., all-correct response pattern, all-incorrect response pattern). Similarly, EAP estimates the subscores through a posterior distribution. The EAP method aims to find the mean of the posterior distribution, which may however lead to biased estimates of the scores (Wainer & Thissen, 1987). For MAP and EAP estimation of the subscores, strong priors, standard normal priors, and non-informative priors can be applied during the estimation process. In a multidimensional context, Carlson (1987) created a joint ML method of estimating MIRT item parameters and multiple latent traits, $\theta = [\theta_1, \theta_2, \ldots, \theta_k]$. Segall (1996) developed a MAP ability estimation approach for calculating θ based on the covariance matrix of the posterior distribution of latent traits.

MIRT ability estimation has been shown to outperform number-correct scoring and OPI (Yao & Boughton, 2007), and was on par with the augmentation methods (de la Torre & Patz, 2005; Dwyer et al., 2006) in terms of recovering true values of the subscores. MIRT has also provided promising results for the estimation of subscores and composite scores (de la Torre & Hong, 2010; de la Torre & Song, 2009; Haberman &

Sinharay, 2010; Wang, Cheng, & Chen, 2004; Yao, 2011; Yao & Boughton 2009). More detailed information about these studies will be provided in the following sections.

Subscore Reliability

Although reliability has not been a well-defined concept within the IRT framework, many studies have presented alternative methods for computing the IRT-based reliability as a function of item parameters and the distribution of person ability (e.g., Bechger, Maris, Verstralen, & Béguin, 2003; Dimitrov, 2003; May & Nicewander, 1994; Samejima, 1994; Shojima & Toyoda, 2002). Some of these methods (e.g., May & Nicewander, 1994) aim to define a constant IRT-based reliability index while others (e.g. Dimitrov, 2003; Shojima and Toyoda, 2002) focus on the approximation of a reliability coefficient. Samejima (1994) proposed a reliability index by combining CTT and IRT approaches through the test information function (TIF) and the ability distribution of a target population. According to Samejima (1994), TIF provides more precise local measures of accuracy in trait estimation than are available from the reliability coefficient.

The concept of score reliability in IRT cannot be defined as a constant but rather is a function of θ. By following the same approach used for test reliability in CTT, Wang et al. (2004) described a method for obtaining IRT-based reliability. First, the test information is averaged over the θ level to obtain $\bar{T}$. The average test information is the average degree of measurement precision that the test or subtest provides for the sampled persons. Based on this fact, the IRT-based test reliability, which is also called the composite test reliability, can be defined as:

$$\rho_{IRT} = 1 - \frac{\bar{T}^{-1}}{\sigma_{\theta}^{2}} \tag{2.26}$$

where σ_θ^2 is the variance of the θ distribution. This IRT reliability is also known as marginal reliability. To simplify the computation of this reliability, Mislevy et al. (1992) suggested a simpler solution when MML estimation is used:

$$\rho_{MML} = \frac{\sigma_{EAP}^2}{\sigma_\theta^2} \tag{2.27}$$

where σ_{EAP}^2 is the variance of the EAP estimates. Wang et al. (2004) noted that the second formula of IRT test reliability is more practical in real data analysis.

Kim and Feldt (2010) also described how to estimate an IRT-based reliability coefficient using the CTT framework. In CTT, test reliability is defined as the ratio of true-score variance to observed-score variance (i.e. σ_T^2/σ_X^2). This is equivalent to the squared correlation between true score (*T*) and the observed score (*X*). From the perspective of nonlinear regression, Kim and Feldt (2010) argued that the same approach can be applied to the correlation of test score X with ability θ:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = 1 - \frac{\sigma_e^2}{\sigma_X^2} \tag{2.28}$$

where σ_e^2 is the average test error variance over an ability distribution of the conditional error variances given θ.

In addition to finding the reliability of a single test, there have been some efforts to find reliability techniques that can be applied to test batteries including several subtests. When a composite score and its associated subscores are estimated, Feldt and Brennan (1989) proposed the following method for computing test reliability:

$$\rho_{ZZ'} = 1 - \frac{\sum_{j=1}^{k} \sigma_{Xj}^2 (1 - \rho_{XjXj'})}{k^2 \sigma_Z^2} \quad (2.29)$$

where Z is the mean of k subscores on a test battery, σ_{Xj}^2 is the observed score variance for the j[th] subscore, $\rho_{\mathrm{XjXj'}}$ is the reliability of the j[th] subtest, and σ_{Z}^2 is the variance of the mean subtest score in the population of examinees. As explained earlier, if there are two sets of composite scores, reliability can be estimated using the Pearson product-moment correlation coefficient or the intraclass correlation coefficient (ICC).

Recently, Haberman (2008) has proposed a new reliability coefficient called the proportional reduction in mean squared error (PRMSE). The main purpose of PRMSE is to determine whether estimated subscores are accurate and they have added value over the total score. The three alternative subscores (i.e. U_{iks}, U_{ikx}, U_{ikc}) proposed by Haberman (2008) have been previously described under the section of unidimensional estimation of subscores. τ_{ks}^2, τ_{kx}^2, and τ_{kc}^2 are the variances of U_{iks}, U_{ikx}, and U_{ikc} respectively. Assuming τ_{k0}^2 is the variance of the true raw score (T_{ik}) of person i on the k[th] subtest, PRMSEs for the subscores become as follows:

$$\mathrm{PRMSE_{ks}} = 1 - \tau_{\mathrm{ks}}^2/\tau_{\mathrm{k0}}^2,$$

$$\mathrm{PRMSE_{kx}} = 1 - \tau_{\mathrm{kx}}^2/\tau_{\mathrm{k0}}^2,$$

$$\mathrm{PRMSE_{kc}} = 1 - \tau_{\mathrm{kc}}^2/\tau_{\mathrm{k0}}^2.$$

As most reliability coefficients, PRMSE also lies between 0 and 1. A larger PRMSE (i.e., closer to 1) is equivalent to a smaller mean squared error in estimating the true subscore. That is, the larger the PRMSE, the more accurate the corresponding

subscore estimate. Haberman (2008) recommended the following criteria to determine whether a subscore or a weighted average has added value. If $PRMSE_{ks}$ is less than $PRMSE_{kx}$, the subscore does not provide added value over the total score, indicating that the observed total score provides more accurate diagnostic information than the observed subscore. Furthermore, if $PRMSE_{kc}$ is substantially larger compared to both $PRMSE_{ks}$ and $PRMSE_{kx}$, then the weighted average has added value over the total score (Sinharay, 2010). Based on these comparisons of PRMSEs, if neither the subscore nor the weighted average has added value over the total score, they should not be reported for diagnostic purposes.

Previous MIRT Studies about Subscore Estimation

The final section of Chapter 2 provides a review of recent studies about the use of MIRT for the estimation of subscores, and the comparison of MIRT against other subscoring methods in terms of subscore reliability.

As previously mentioned, MIRT models can be applied to test batteries and tests consisting of multiple subtests. MIRT models allow for the use of correlations between subtests of test batteries to improve the measurement precision of individual ability estimates. Wang, Chen, and Cheng (2004) demonstrated two empirical examples to solve the problem of ignoring the correlations between latent traits that yields imprecise measures when tests are short. For this study, the multidimensional random coefficients multinomial logit (MRCML) model (Adams, Wilson, & Wang, 1997) was used. To compare measurement efficiency of the unidimensional and multidimensional approaches, test reliability and numbers of items needed to achieve the same measurement precision were estimated. Data from a science proficiency test and a teacher

personality inventory were analyzed using both unidimensional and multidimensional models. The results showed that the multidimensional approach improved measurement precision substantially using the correlations between latent traits, especially when tests are short and the number of tests is large. The greater the number of latent traits and the shorter the target tests, the more significant the improvements are. Wang, Chen, and Cheng (2004) suggested that if there are other kinds of collateral information available, such as persons' educational backgrounds, status on demographic variables, in-class test grades, or homework grades, they could be incorporated into the multidimensional approach to further improve measurement efficiency.

De la Torre and Patz (2005) conducted a study where they applied a hierarchical Bayesian framework to ability estimation. The authors proposed a practical method based on the availability of information from multiple tests measuring the correlated abilities given in a single test administration. They conducted a simulation study to examine the performance of the hierarchical model ability estimates under some factors such as the number of abilities (2 and 5), the number of items (10, 30, and 50) and the degree of correlation between the abilities (.0, .4, .7, and .9). The M3PL model was used for estimating the subscores. To quantify the amount of improvement attributable to simultaneous estimation of tests, Pearson correlations between estimated and true abilities and relative efficiency, which is the ratio of the mean squared error of the unidimensional ability estimates over the mean square of the multidimensional ability estimates, were computed. Correlations and ability estimates were obtained using a MCMC algorithm. The results of this study showed that ability estimates obtained from the hierarchical (i.e., multidimensional) approach were more accurate and precise

compared to ability estimates obtained using one dimension at a time. This approach becomes more efficient when tests are short and dimensions are highly correlated. The authors showed that employing multidimensional scoring can further reduce the bias and standard error of the estimates of traditional unidimensional EAP, which already have smaller bias and standard error compared to other methods of estimation (Kim & Nicewander, 1993; Thissen & Orlando, 2001).

De la Torre (2009) proposed a model that incorporates ancillary variables and correlational structure of the latent abilities in scoring the examinees using various MIRT models. The study specifically focused on the M3PL model and covariates related to examinees. Markov chain Monte Carlo parameter estimation algorithms were used for both simulated and actual data. The study also examined how the number of tests (2 or 5), the length of test (10 or 20 items), the correlation between the different abilities (.5 or .9), and the correlation between the ancillary variables and the latent trait (.25 or .5) affect the quality of the estimates. Results showed that using the different sources of information separately or simultaneously provided better ability estimates (i.e., higher correlation with the true abilities and smaller posterior variance and mean squared error). The optimal condition occurs when several short tests measuring highly correlated abilities that also correlate highly with the covariates are used.

To examine the performance of the MIRT models in polytomously scored responses, de la Torre (2008) used multidimensional scoring of abilities in ordered polytomous data. The Generalized Partial Credit (GPC) model was used for estimating subscores. The study systematically examined how improvement in ability estimates is affected by factors such as the number of score categories (2, 3, or 4), number of tests (2

or 5), test length (5, 10, or 20 items), and correlation between abilities (.0, .4, .7, or .9). Ability estimates and the correlational structure were obtained using the MCMC method. Correlation and mean squared error (MSE) between true and estimated abilities and the posterior variance of the abilities were used to examine the quality of estimates. As in previous studies, this study also indicated that greater improvement can be achieved when the abilities are very highly correlated. Results showed that more accurate estimates of the correlation between the abilities were obtained when several long tests with more score categories were used. The number of tests did not affect the accuracy of estimates substantially. Higher correlations between abilities could be accurately estimated even with only two tests with at least 10 polytomous items in each of them. In extreme cases where abilities measured by the different subtests were perfectly correlated, reporting subtest scores on top of the overall score did not provide additional information. The variability observed among the subscores did not reflect true differences in abilities, but rather differences due to measurement errors. The posterior variances representing the precision of the ability estimates showed that when longer and more tests, more score categories, and higher correlations between abilities were involved, better results were obtained

Other studies in the literature have focused on the use of MIRT for obtaining subscores under different conditions. For example, Yao and Boughton (2007) conducted a simulation study of dichotomous and polytomous MIRT for subscale score proficiency estimation using real data-derived parameters from a large-scale statewide assessment. The simulation conditions were with sample size (1000, 3000, or 6000) and correlations between subscales (.0, .1, .3, .5, .7, or .9). The study examined the recovery of a Markov

chain Monte Carlo (MCMC) estimation approach to multidimensional item and ability parameter estimation, as well as subscale proficiency and classification rates. The accuracy of subscore estimation was investigated for number-correct scores (NC), multidimensional IRT Bayesian subscale scores (BMIRTSS), multidimensional IRT Bayesian domain subscale scores (BMIRTDS), and objective performance index scoring (OPI). Results showed that to report accurate diagnostic information at the subscale level, the subscales need to be highly correlated and borrow information from other subscales or a multidimensional approach should be used. In terms of classification recovery, as the correlations increase among the dimensions, the average error rates for BMIRTSS and BMIRTDS become closer to the OPI rates. As the correlation among the dimensions decreased, the error rates for the OPI increased, and BMIRTSS and BMIRTDS classification errors decreased.

DeMars (2005) compared several IRT-based methods of subscoring. These methods included two bifactor models, one of which used each subtest as a composite score based on the primary trait measured by the set of tests and a secondary trait measured by the individual subtest; the other was a model where the traits measured by the subtests were separate but correlated. Composite scores based on unidimensional item response theory, with each subtest borrowing information from the other subtests, as well as independent unidimensional scores for each subtest, were also considered. Data from two multiple-choice assessment tests were used for this study. By using results from real data, simulations were run to assess bias and RMSE of the ability estimates. Results showed that the independent unidimensional scores showed the greatest bias and RMSE. The relative bias and RMSE for the other approaches differed on the two tests. The

bifactor and 2-factor models showed very similar levels of bias and RMSE; on one test higher than the augmented scores at the extremes, and on the other test lower. Based on these results, there is no clear advantage for any of these three methods over the others, but all produced lower bias and RMSE than the separate unidimensional models.

Yao (2010) investigated the performance of four methods [UIRT model, higher-order IRT model (HO-IRT), MIRT model, and the bifactor general model] using simulated data to demonstrate how reliable and valid the overall scores and domain scores provided by each method are. For data simulation, sample size (500, 1000, or 2000), correlations between domains (.2, .3, .4, .5, .7, or .9), and test length (20, 32, 48, or 60 items) were manipulated. Root mean squared error (RMSE), absolute bias (ABS), bias (BIAS), and reliability (squared correlation between true and estimated parameters) were used to evaluate the accuracy of overall ability and domain ability parameter recoveries. RMSE and the test response function (TRF) were used to evaluate the item parameter recovery. The findings showed that the M3PL model provided more reliable domain and overall scores in comparison to the other models used in this study. As the test length increased, as the correlation between dimensions increased, and as the sample size increased, the reliability of domain scores increased and RMSE and BIAS decreased for all the models and methods. The MIRT estimation method performed slightly better than HO-IRT for all the criteria and for all the conditions, but the differences between the two methods were minor. For the overall scores, the MIRT method performed as well as HO-IRT when the correlation was high. Although the HO-IRT model performed equally well in most cases, it may not be as useful as the MIRT model because an item can only

contribute to one domain (simple structure) in the HO-IRT method whereas the MIRT models allow for both simple and non-simple structures.

De la Torre, Song, and Hong (2011) conducted a similar study to Yao (2010), where the authors made a comparison of four subscoring methods (multidimensional scoring (MS), augmented scoring (AS), higher order IRT scoring (HO), and objective performance index scoring (OPI)) using simulated and real data. In the simulation study, test length (10, 20, or 30 items), number of subtests or domains (2 or 5), and correlation between the abilities (.0, .4, .7, or .9) were manipulated. The quality of the subscore estimates was evaluated using the correlation between the true and estimated abilities, RMSE of the estimates across the examinees, and conditional bias and conditional mean absolute deviation (MAD). Results indicated that the correlation-based methods (i.e., MS, AS, and HO) provided mostly similar results, and performed most efficiently under conditions involving multiple short subtests, more dimensions and highly correlated abilities. In most of the conditions considered, the OPI method performed poorer compared to other methods on both ability estimates and proportion correct scores. The authors argued that although HO and MS may provide better estimates than AS for extreme abilities, the AS method can be preferred because of its efficiency and lesser complexity depending on the purpose of subscoring.

Summary

The effectiveness of MIRT models for improving measurement precision and accuracy of subscores has been well established in previous studies. Researchers investigated the performance of various MIRT models under various data conditions such

as test length, number of dimensions, sample size, and correlations between dimensions in simulation studies. A majority of the research has indicated an increase in the measurement precision of subscores obtained from MIRT models compared to subscores obtained from other methods (e.g., Haberman & Sinharay, 2010; Sheng & Wikle, 2007; Tate, 2004; Wang et al., 2004; Yao, 2010; Yao & Broughton, 2007). The aim of these subscoring methods is to improve diagnostic utility of the subscores by improving the reliability of the subscores. It is assumed that the reliable subscores will help identify an examinee's relative strengths and weaknesses. However, the main concern should be whether the subscores provide reliable information about an examinee's relative strengths and weaknesses.

In the comparison of UIRT and MIRT models regarding subscore reliability, the evaluation criteria were mostly the correlation between true and estimated subscores, bias and RMSE. Although these measures indicate to what extent subscores are accurately estimated, they do not consider how the relationship among the subscore estimates varies depending on conditions such as test length, number of dimensions, or correlation structure of the subscores. Although PRMSE (Haberman, 2008) seems to be promising for the determination of subscore reliability, the availability of this approach for MIRT is still questionable because it is heavily based on CTT.

Reliability is an important psychometric characteristic of scores, and it has received great attention in the literature for many years. As the importance of diagnostic information from subscores increases in education, obtaining reliable and accurate subscores has become a more crucial task. The research on subscore estimation in MIRT is still in a development phase with many uncertainties. Considering the computational

burden and high complexity of MIRT models, the major benefits of this framework should be clearly revealed for both researchers and practitioners. Therefore, more studies are needed to determine whether MIRT should be used as an alternative method for estimating subscores with added value.

CHAPTER 3

METHODOLOGY

Chapter 3 consists of four sections. The first section discusses the details of multidimensional and unidimensional IRT models, and the method of estimation used for estimating the subscores. In the second section, the framework of between-person and within-person reliability is introduced, and between-person and within-person reliability coefficients are described. The third section explains the design of the simulation study, including simulation conditions, data generation, and subscore estimation procedures. In addition, evaluation criteria for the estimated subscores are described. In the last section, a real data study is described. The instrument, the sample, and data preparation for estimating subscores are explained.

Subscoring Procedure

The following sections describe the unidimensional and multidimensional IRT models that are used for subscore estimation in this study. Furthermore, details about the subscore estimation process are provided.

Models for Subscore Estimation

In this study, two IRT models were used for estimating the subscores. The first model was a unidimensional 3PL model. The 3PL model and its components were explained in Chapter 2 (see Equation 2.3). This model assumes that each subtest measures a unidimensional ability that is not affected by the level of abilities obtained from other subtests. That is, each subtest is a simple structure by itself. Therefore, there is

neither a compensatory nor a noncompensatory relationship between the estimated subscores from different subtests. The following example shows the item parameter structure of the 3PL model from a test including multiple subtests. Assume that there is a test consisting of three subtests based on three content domains (e.g., algebra, trigonometry, and geometry). Each subtest includes ten multiple-choice items that are scored dichotomously. The item parameters for the three subtests based on the unidimensional 3PL model are illustrated in Figure 3.1 below. Each subtest has a separate set of item parameters. There was no subscore augmentation that weighted the estimated subscores based on ancillary variables. The estimated abilities (θ_1, θ_2, and θ_3) were used as subscore estimates.

Subtest	**Item**	$\boldsymbol{\beta_{2j}}$	$\boldsymbol{\beta_{1j}}$	$\boldsymbol{\beta_{3j}}$
1	1	1.87	-1.44	0.15
1	2	2.61	-0.42	0.09
1	3	2.46	1.56	0.12
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
2	11	2.02	0.33	0.23
2	12	1.74	-1.23	0.11
2	13	2.25	1.31	0.07
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
3	21	2.62	1.22	0.15
3	22	2.13	-0.15	0.19
3	23	1.81	1.89	0.06
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Figure 3.1. An example of item parameters for three unidimensional subtests.

The second model was a multi-unidimensional 3PL (M3PL) model. This model assumes that each subtest measures a domain defined by unique items (i.e., simple structure). The overall test is assumed to be multidimensional while each subtest still remains unidimensional. The subscores from the subtests are estimated simultaneously. The estimation procedure allows for including the inter-dimension correlations as ancillary information to improve the precision of subscore estimates. Using the notation of the compensatory MIRT model in Yao and Schwartz (2006), the M3PL model (Reckase, 1997) for a dichotomous item *j* answered by person *i* with abilities $\vec{\theta}_i = (\theta_{i1}, \ldots, \theta_{iD})$ on a test including *D* subtests can be shown as:

$$P_{ij} = P\left(x_{ij} = 1 \middle| \vec{\theta}_i, \beta_{1j}, \vec{\beta}_{2j}, \beta_{3j}\right) = \beta_{3j} + \frac{1 - \beta_{3j}}{1 + e^{1.7\left(-\vec{\beta}_{2j} \odot \vec{\theta}_i^T + \beta_{1j}\right)}}, \tag{3.1}$$

where x_{ij} is the response of person *i* to item *j*, $\vec{\beta}_{2j}$ is a vector of item discrimination parameters for *D* dimensions (i.e., subtests), β_{1j} is the item difficulty parameter, β_{3j} is the lower asymptote or the guessing parameter, and $-\vec{\beta}_{2j} \odot \vec{\theta}_i^T = \sum_{l=1}^{D} \beta_{2jl}\theta_{il}$. Using the same example given earlier, the M3PL item parameter structure for a test with three unidimensional subtests is illustrated in Figure 3.2.

Subtest	Item	$\vec{\boldsymbol{\beta}}_{2j}$	$\boldsymbol{\beta}_{1j}$	$\boldsymbol{\beta}_{3j}$
1	1	[1.87, 0, 0]	-1.44	0.15
1	2	[2.61, 0, 0]	-0.42	0.09
1	3	[2.46, 0, 0]	1.56	0.12
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
2	11	[0, 2.02, 0]	0.33	0.23
2	12	[0, 1.74, 0]	-1.23	0.11
2	13	[0, 2.25, 0]	1.31	0.07
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
3	21	[0, 0, 2.62]	1.22	0.15
3	22	[0, 0, 2.13]	-0.15	0.19
3	23	[0, 0, 1.81]	1.89	0.06
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Figure 3.2. An example of a multi-unidimensional structure based on three subtests.

Subscore Estimation Method

In this study, both unidimensional and multidimensional subscores were estimated with the maximum a posteriori (MAP) method. By using the information of the prior distribution, MAP estimation can provide a lower estimated error of θ (Chen, 2006). In most cases, MAP is more feasible compared to EAP and MLE methods for estimating subscores from MIRT models because MAP can estimate the subscores for the examinees who obtained either a perfect or zero score on one of the dimensions, which is not possible in MLE. Also, since MAP does not require an iterative estimation process like EAP, the estimation time is much shorter.

Using the item parameters defined above for the M3PL model, the probability can be written as

$$P_{ij} = P_{ij}\left(X_{ij}\middle|\vec{\theta}_i, \beta_{1j}, \vec{\beta}_{2j}, \beta_{3j}\right) = P_{ij}^{(X_{ij}=1)}(1 - P_{ij})^{(X_{ij}=0)}, \tag{3.2}$$

where X_{ij} is response of examinee i to item j, $\vec{\theta}_i$ is a vector of abilities, and $\beta_{1j}, \vec{\beta}_{2j}$, and β_{3j} are item parameters. If the item parameters for the j^{th} item are $\overrightarrow{\boldsymbol{\beta}_j} = (\beta_{1j}, \vec{\beta}_{2j}, \beta_{3j})$ and the item parameters for all items on the test are expressed as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_j, \ldots, \beta_J)^T$, then the likelihood equation can be shown as follows:

$$P(X|\theta, \beta) = \prod_{i=1}^{N} P(\overrightarrow{\boldsymbol{X}_i}|\, \vec{\boldsymbol{\theta}}_i, \boldsymbol{\beta}) = \prod_{i=1}^{N}\prod_{j=1}^{J} P\left(\boldsymbol{X_{ij}}|\vec{\boldsymbol{\theta}}_i, \vec{\boldsymbol{\beta}}_j\right). \tag{3.3}$$

Boughton, Yao, and Lewis (2006) defined the posterior probability distribution for an examinee population with ability θ using the pre-defined population priors in the Bayesian framework as follows:

$$P(\theta, \beta, \lambda|X) \propto P(X|\theta, \beta, \lambda)P(X|\theta, \beta)P(\lambda)P(\beta) \tag{3.4}$$

$$= P(X|\theta, \beta)P(\theta|\lambda)P(\lambda)P(\beta) \tag{3.5}$$

where $\lambda = (\vec{\mu}, \boldsymbol{\Sigma})$, and λ is defined by $\vec{\mu}$ and $\boldsymbol{\Sigma}$ that are the vector of population means and the variance-covariance matrix of the abilities (i.e., population priors), respectively. For instance, the DxD correlation matrix, the vector of the population means, and the vector of the population variances for a D-dimensional test can be written as

$$\sum = \begin{bmatrix} 1 & \cdots & r_{1.D} \\ \vdots & \ddots & \vdots \\ r_{D.1} & \cdots & 1 \end{bmatrix}_{DxD} \qquad \vec{\mu} = [\mu_1 \quad \cdots\cdots \quad \mu_D] \qquad \vec{\sigma} = [\sigma_1 \quad \cdots\cdots \quad \sigma_D].$$

Equation 3.5 represents the posterior likelihood distribution of $\vec{\theta}$ that can be shown as $f(\vec{\theta}|\vec{X})$. The elements of the 1st and 2nd derivatives of the posterior density function for MAP estimation using variance-covariance matrix, **Σ**, are expressed as

$$\frac{\partial \log f(\vec{\theta}|\vec{X})}{\partial \vec{\theta}} = \frac{\partial \log L(\vec{X}|\vec{\theta})}{\partial \vec{\theta}} - \frac{\partial(\vec{\theta} - \vec{\mu})}{\partial \vec{\theta}} \Sigma^{-1}(\vec{\theta} - \vec{\mu}), \tag{3.6}$$

and

$$\frac{\partial^2 \log f(\vec{\theta}|\vec{X})}{\partial \vec{\theta}^2} = \frac{\partial^2 \log L(\vec{X}|\vec{\theta})}{\partial \vec{\theta}^2} - \Sigma^{-1} = J(\vec{\theta}) - \Sigma^{-1} \tag{3.7}$$

where $\boldsymbol{J}(\vec{\boldsymbol{\theta}})$ is the matrix of the second partial derivative.

When using Bayesian MAP estimation, MIRT-based abilities are estimated by finding the mode that maximizes the posterior likelihood function, $f(\vec{\theta}|\vec{X})$, using the Newton-Raphson method (Yao, 2013), which can be expressed as

$$\frac{\partial \log f(\vec{\theta}|\vec{X})}{\partial \vec{\theta}} | \vec{\theta} = 0, \tag{3.8}$$

and the m[th] approximation that maximizes the posterior likelihood function becomes

$$\vec{\theta}^{m+1} = \vec{\theta}^{m} - \vec{\delta}^{m}, \tag{3.9}$$

where

$$\vec{\delta}^{m} = \left[J(\vec{\theta}^{m})\right]^{-1} \text{x} \frac{\partial \log f(\vec{\theta}|\vec{X})}{\partial \vec{\theta}}. \tag{3.10}$$

In the MAP subscoring procedure, using standard normal or no informative priors would ignore the correlated information between domains that would yield similar results

as those from MLE. However, using strong priors would allow the information to be borrowed from each dimension and increase the precision, especially when the test is short (Yao, 2013; Yao & Boughton, 2007). The same estimation procedure can be applied to each subtest separately so that unidimensional MAP estimates are obtained. For unidimensional MAP, the only prior information would be the population mean and variance. Because there is only one dimension, it is not possible to use inter-dimension correlations during the estimation.

Subscore Reliability

The aim of an assessment is to obtain reliable scores that can be used to evaluate examinees' skills for diagnostic, classification or selection purposes. The higher the reliability of a test, the better examinees are evaluated based on their test scores. Brennan (2005) described three types of reliability based on classical test theory. These are parallel-form reliability, canonical reliability, and internal consistency. The third reliability approach, internal consistency, is the most common way to evaluate the reliability of scores from tests and subtests with dichotomously scored items. Coefficient alpha (Cronbach, 1951) and the Kuder–Richardson Formula 20 (KR-20; Kuder & Richardson, 1937) are typical examples of reliability coefficients that examine the variation in test scores across the examinees within a single test. In such reliability coefficients, obtaining consistent scores across all examinees on the same test or subtest is highly desirable. Because this type of reliability focuses on the variation between the examinees' scores on the test, it can be seen as a measure of between-person reliability. Between-person reliability coefficients can be particularly useful for selection assessments because this type of assessment requires reliable and differentiating scores so

that the examinees with high performances can be separated from those with low performance. For tests with multiple subtests or test batteries, internal consistency is computed for each subtest separately, ignoring the relationship between the subtests.

When the main purpose of an assessment is to determine the strengths and weaknesses of examinees in particular domains, the variation among the subscores for each individual is more important than the variation within each subscore across all examinees. Multiple subscores derived from an assessment can be considered as a test score profile. The term "test score profile" can be described as a collection of test scores attained by a particular student. An examinee's test score profile provides information about his/her strengths and need for improvement of the knowledge, skills, and abilities relevant to the assessments in the profile (Arce-Ferrer, 2010). Individual test reports based on test score profiles often include recommendations to improve students' academic achievement and classroom teaching, or to select the most effective intervention for students. The analysis of test score profiles can provide information about an examinee's performance in either broad content areas (e.g., reading, mathematics, and science) or narrow content domains (e.g., algebra, geometry, and calculus).

Test score profiles can be used for both for inter-individual and intra-individual interpretations. Through the use of profile analysis techniques, a person's strengths and weaknesses can evaluated based upon their ipsatized scores (i.e. pattern vectors), which are obtained by subtracting an examinee's average score on the domains from each score in the test score profile (Davison, 1996; Davison, Kim, & Close, 2009). Figure 3.3 shows the test score profiles of six individuals on three domains. For each person, the three

scores in the person's profile vector are shown above the profile while each person's ipsatized scores are shown below the profile. Individual differences in profile can be seen in the comparison of the top and bottom three profiles. Also, from left to right, variation in the ipsatized scores can be seen. The first two test score profiles display a linearly increasing pattern; the second two display an inverted V shape pattern; and the last two display a linearly decreasing pattern.

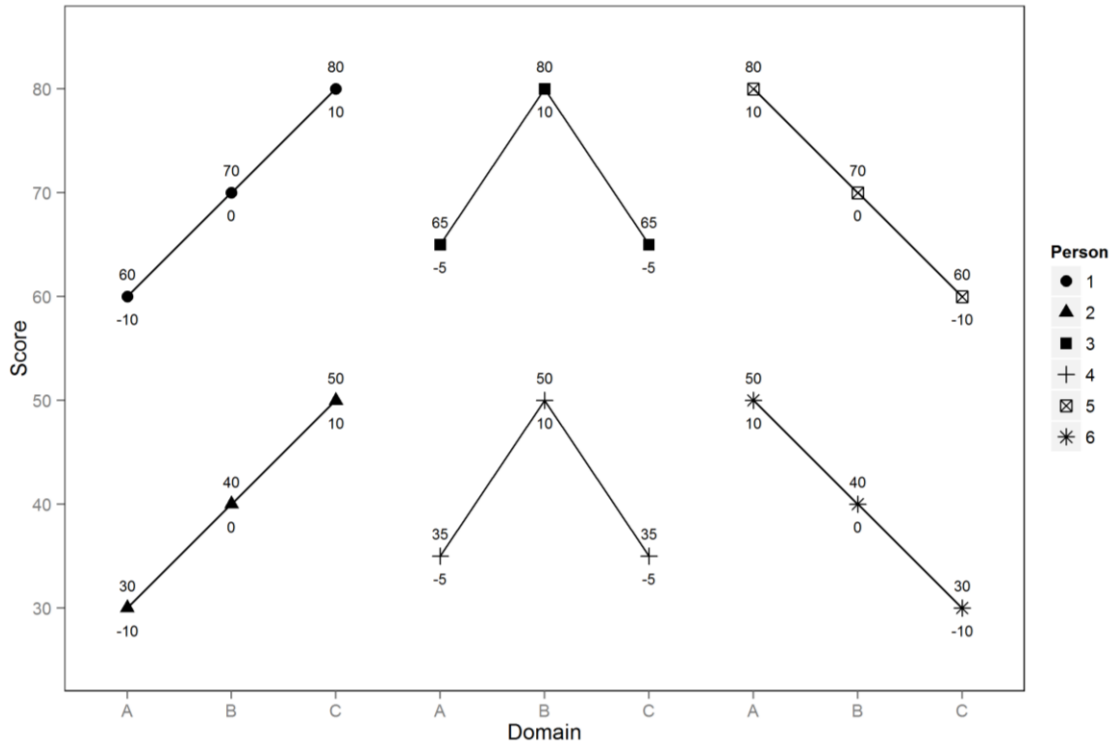

Figure 3.3. A hypothetical example of test score profiles of six persons on three domains. Adapted from "Factor Analytic Modeling of Within Person Variation in Score Profiles" by M. L. Davison, S. Kim, and C. Close, 2009, Multivariate Behavioral Research, 44, p. 669.

Examining an individual's strengths and weaknesses from ipsatized scores has brought some uncertainties due to lack of evidence for subscore reliability and validity (Watkins, Glutting & Youngstrom, 2005). In order to estimate the precision of unique

patterns of test score profiles, Davison, Chang, and Davenport (2012) proposed an approach for estimating the reliability of individual differences in test score profiles based upon the total variation, the variation among individuals, and the variation among the subscores. This approach is mainly an extension of canonical test reliability as proposed by Conger and Lipshitz (1973). Canonical reliability (Conger & Lipshitz, 1973) for test score profiles is the multivariate version of the traditional univariate reliability as the ratio of the variance of true scores to the variance of observed scores. Conger and Lipshitz (1973) define the observed difference vector as $(\vec{X}_i - \vec{X}_i.)$, where $\vec{X}_i$ is the vector of subscores for person *i*, and $\vec{X}_i.$ is the average of subscores that person *i* obtained. Using the observed and true difference vectors, canonical reliability can be written for any distance function as

$$\rho = \frac{\left(\vec{T}_i - \vec{T}_i.\right)' \mathrm{A}\, (\vec{T}_i - \vec{T}_i.)}{\left(\vec{X}_i - \vec{X}_i.\right)' \mathrm{A}\, (\vec{X}_i - \vec{X}_i.)}, \tag{3.11}$$

where $(\vec{T}_i - \vec{T}_i.)$ is the true difference vector, and A is a square matrix used for weighting the reliability. The square matrix A can be either a correlation matrix of the subscores on the test score profiles or an identity matrix if weighting is not desired (Conger & Lipshitz, 1973).

Using the profile reliability framework defined by Conger & Lipshitz (1973), Davison et al. (2012) has described a profile reliability approach that makes use of both the vector of difference scores (i.e. pattern) and the vector of the average subscore (i.e. level) in a test score profile. For a given person, the level of a test score profile is the mean of the subscores in the profile, which can be expressed as

$$\bar{X}_i = \frac{1}{D}\sum_{d=1}^{D} X_{id}, \tag{3.12}$$

where X_{id} is the score of person *i* (*i* = 1, ……, *I*) on subtest *d* (*d* = 1, ….., *D*), and $\bar{X}_i$ is the average scores from *D* subtests for person *i*. When Equation 3.12 is applied to each person in the sample, a level vector that consists of *D* level scores is obtained (see Figure 3.4)

Person	Subtests		
	1	…. ….	*D*
1	X_{11}		X_{1D}
2			
.			
.			
.			
.			
.			
I	X_{I1}	….	X_{ID}

$\bar{\boldsymbol{X}}_{\boldsymbol{1}} = (X_{11} + \cdots + X_{1D})/D$
$\bar{\boldsymbol{X}}_{\boldsymbol{2}} = (X_{21} + \cdots + X_{2D})/D$
$\bar{\boldsymbol{X}}_{\boldsymbol{3}} = (X_{31} + \cdots + X_{3D})/D$
.
.
.
.
$\bar{\boldsymbol{X}}_{\boldsymbol{I}} = (X_{I1} + \cdots + X_{ID})/D$

Figure 3.4. Obtaining the level scores for each examinee.

The pattern of a test score profile is a vector of the score differences (i.e. ipsatized scores) between each subscore and the mean of the subscores for a given person. A pattern vector of a test score profile can be shown as $\{(X_{i1} - \bar{X}_i), \ldots\ldots, (X_{id} - \bar{X}_i)\}$, where X_{pv} is the subscore for person *i* on the subtest *d*, and $\bar{X}_i$ is the mean of the subscores that person *i* obtained from *D* subtests. Therefore, for each person, the number of difference scores in the pattern vector is the same as the number of the subtests in the profile. Figure 3.5 illustrates the pattern vectors of the examinees on *D* subtests.

Person	Pattern Scores		
	1	…. ….	D
1	$(X_{11} - \bar{X}_1)$		$(X_{1D} - \bar{X}_1)$
2			
.			
.			
.			
.			
.			
P	$(X_{I1} - \bar{X}_i)$	….	$(X_{ID} - \bar{X}_i)$

Figure 3.5. Obtaining the vector of pattern scores for each examinee.

Using the terms described above, the total score variation (*T*) of a test score profile can be defined as the sum of the variances for the *D* subtests. The total variance of a test score profile can be indicated as follows:

$$T = \sum_{d=1}^{D} \sigma_d^2 \tag{3.13}$$

Davison et al. (2012) stated that the total score variation can be divided into two orthogonal components: $T = B + W$; where B is the between-person variation referred to as profile level, and W is the within-person variation referred to as profile pattern. Essentially, B is the between-person variation due to individual differences in profile level, and W is the within-person variation due to individual differences in profile patterns. The following sections show the derivation of between-person and within-person reliability coefficients based on the pattern and level variances. Because level is the indicator of between-person variation (*B*), and pattern is the indicator of within-person variation (*W*), these terms will be used interchangeably throughout this study.

Between-person Subscore Reliability

To characterize the between-person reliability in the test score profiles, the relationship between observed and true test scores in the CTT framework is used. Based on CTT, if observed test scores (X_{id}) are expressed as a sum of true scores (T_{id}) and error (E_{id}), then each person has a profile of observed scores$\{X_{i1}, X_{i2}, \cdots, X_{iD}\}$ as well as the profile of true scores $\{T_{i1}, T_{i2}, \cdots, T_{iD}\}$. Therefore, the same approach for computing level and pattern scores can also be applied to true scores. If Equation 3.12 is applied to the true scores, the level of true scores becomes $\bar{T}_i = (\sum_{d=1}^{D} T_{id})/\mathrm{D}$.

Since reliability in CTT is defined as the proportion of observed total variation in profiles that is attributable to true scores, the total variation in the observed and true level scores should be computed for obtaining between-person reliability. If the total observed score variation is defined as the sum of the variances for *D* subtests, then the observed total level variance becomes $B = D * \sigma_{\bar{X}_i}^2$, where *B* is the observed total level variance and *D* is the number of subtests. Similarly, the true total level variance based on the true level values becomes $B_T = D * \sigma_{\bar{T}_i}^2$; where $\sigma_{\bar{T}_i}^2$ is the variance of true level scores, and so B_T becomes the total true level variance. Using the observed and true level variances, between-person reliability can be defined as the ratio of true level variation to observed level variation:

$$\rho_B = \frac{\sigma_{\bar{T}_i}^2}{\sigma_{\bar{X}_i}^2} = \frac{B_T}{B}. \tag{3.14}$$

Based on Equation 3.14, between-person reliability can be interpreted as the proportion of variation in observed profile levels that can be attributable to true level

variation in a test score profile. So, the more consistent the variation among the examinees becomes, the higher between-person reliability is.

Within-person Subscore Reliability

Within-person reliability can be defined in a similar fashion. In a test score profile, the total observed pattern variance is $W = \sum_{d=1}^{D}\left[\frac{1}{I}\sum_{i=1}^{I}(X_{id} - \bar{X}_i)^2\right]$, where W represents the total observed within-person variation due to individual differences in the subscores. Similarly, the total true pattern variance can be shown as $W_T = \sum_{d=1}^{D}\left[\frac{1}{I}\sum_{i=1}^{I}(T_{id} - \bar{T}_i)^2\right]$, where W_T is the total true within-person variation in the test score profile. By using the same approach with the ratio of observed and true scores, within-person reliability can be defined as the ratio of true pattern variation to observed pattern variation as follows:

$$\rho_W = \frac{\sum_{d=1}^{D}\left[\frac{1}{I}\sum_{i=1}^{I}(T_{id} - \bar{T}_i)^2\right]}{\sum_{d=1}^{D}\left[\frac{1}{I}\sum_{i=1}^{I}(X_{id} - \bar{X}_i)^2\right]} = \frac{W_T}{W}. \quad (3.15)$$

The within-person reliability coefficient can be interpreted as the proportion of variation in observed profile patterns that can be attributed to true pattern variation in the test score profile. Within-person reliability can also be interpreted as a weighted average of the within-person reliability for each subtest, and as a weighted average of the person profile reliabilities (see Davison et al., 2012).

Overall Profile Reliability

As with between-person and within-person reliability, the overall profile reliability is also the proportion of observed total variation in the test score profile that is attributable to true scores. The total observed and true score variances can be defined as

the sum of the observed score variance ($T = \sum_{d=1}^{D} \sigma^2(X_{id})$) and the sum of the true score variance ($T_T = \sum_{i=1}^{I} \sigma^2(T_i)$) in the separate measures respectively. To find the overall profile reliability in a test score profile, the ratio of true total variation to observed total variation can be computed as follows:

$$\rho_T = \frac{\sum_{d=1}^{D} \sigma^2(T_{id})}{\sum_{d=1}^{D} \sigma^2(X_{id})} = \frac{T_T}{T}. \tag{3.16}$$

The overall profile reliability is directly related to both the between-person and within-person reliability because it is a weighted average of the between-person and within-person reliability. Following the fact that the total variation is the sum of the between-person and within-person variation, Equation 3.16 can be rewritten as follows:

$$\rho_T = \frac{T_T}{T} = \frac{B_T + W_T}{T} = \frac{B}{T} * \frac{B_T}{B} + \frac{W}{T} * \frac{W_T}{W} = \frac{1}{B + W}(B * \rho_B + W * \rho_W). \tag{3.17}$$

According to Davison et al. (2012), in some cases, most or all of the variation in a test score profile is due to level; in other cases, it is due to pattern. Therefore, as a weighted average of between-person and within-person variation, the total profile reliability always lies within a range between within-person reliability and between-person reliability. All of the between-person reliability, the within-person reliability, and the overall profile reliability coefficients range from 0 to 1, where a higher value indicates higher reliability. For a test in which subscores are reliable and have added value over the total score, within-person reliability should be higher and more dominant than between-person reliability in the profile of test scores.

Estimating Total, Between-person, and Within-person Reliabilities

As explained above, within-person and between-person reliability coefficients are based on the relationship between true and observed subscores in a test score profile. It is assumed in the classical test theory that true scores are unknown, and observed scores are the approximations of true scores. When there are parallel forms of a test, the covariance of the two forms provides an estimate of the true score variation. Holland and Hoskens (2003) noted that if true scores from two tests are perfectly correlated (i.e., congeneric) and equally reliable, then the correlation between the observed scores provides an estimate of the proportion of true score variance to observed score variance. Brennan (2005) explained the derivation of true score variance from two parallel test profiles as follows:

$$\rho(X, X') = \frac{\sigma(X, X')}{\sigma(x)\sigma(X')}, \tag{3.18}$$

where X and X' are two parallel test profiles, and each profile consists of D subtests. When the expectations for the correlation and covariance are taken for all possible pairs of parallel test profiles, the expected value of the covariance becomes

$$\boldsymbol{E}\ \rho(X, X') = \frac{\boldsymbol{E}\ \sigma(X, X')}{\boldsymbol{E}\ \sigma^2(X)}, \tag{3.19}$$

and the expected value of the covariance of the parallel forms is

$$\begin{aligned} \boldsymbol{E}\ \sigma(X, X') &= \boldsymbol{E}\ \left[\frac{1}{D}\sum_{d=1}^{D}(X_d - \bar{X})(X'_d - \bar{X}')\right] \\ &= \frac{1}{D}\sum_{d=1}^{D} \boldsymbol{E}\ [(X_d - \bar{X})\ (X'_d - \bar{X}')] \end{aligned} \tag{3.20}$$

$$= \frac{1}{D}\sum_{d=1}^{D} \boldsymbol{E}\,(X_d - \bar{X})\;\boldsymbol{E}\,(X'_d - \bar{X}')$$

$$= \frac{1}{D}\sum_{d=1}^{D} (T_d - \mu)^2$$

$$= \sigma^2(T).$$

Following the reasoning in Equation 3.20, after obtaining an estimate of the covariance between every possible pair of parallel tests d and d′, the true score variation becomes equal to the average of all possible covariances because the tests d and d′ are assumed to have equal variances. Based on this fact, the proportion of total profile variation due to true scores can be estimated as follows:

$$\hat{\rho}_T = \frac{\sum_{d=1}^{D=1} \hat{\sigma}(X_{id} X_{id'})}{\sqrt{\sum_{d=1}^{D=1} \hat{\sigma}(X_{id}) * \sum_{d=1}^{D=1} \hat{\sigma}(X_{id'})}} = \frac{\sum_{d=1}^{D=1} \hat{\sigma}(X_{id'})}{\sum_{d=1}^{D=1} \hat{\sigma}(X_{id})} \tag{3.21}$$

Using the same approach, within-person and between-person reliability coefficients based on the test scores from parallel test forms can be formulated as follows:

$$\hat{\rho}_W = \frac{\sum_{i=1}^{I}\left(\sum_{d=1}^{D}(X_{id} - \bar{X}_i)(X_{id'} - \bar{X}_{i'})\right)}{\sum_{i=1}^{I}\left(\sum_{d=1}^{D}(X_{id} - \bar{X}_i)^2\right)} \tag{3.22}$$

$$\hat{\rho}_B = \frac{\hat{\sigma}(\bar{X}_i \bar{X}_{i'})}{\sqrt{\hat{\sigma}^2(\bar{X}_i)\hat{\sigma}^2(\bar{X}_{i'})}} = \frac{\hat{\sigma}(\bar{X}_i \bar{X}_{i'})}{\hat{\sigma}^2(\bar{X}_i)}. \tag{3.23}$$

It should be noted that Equation 3.22 indicates an overall within-person reliability coefficients that is a weighted average of within-person reliability coefficients from all persons. Without averaging over the persons, within-person reliability coefficients can also be used to evaluate reliability for each individual in the sample.

Simulation Study

To address the research questions stated in Chapter 1, a simulation study was designed with various simulation conditions. The aim of the simulation study was to compare UIRT and MIRT ability estimates (i.e. subscores) in terms of between-person and within-person reliability, and to investigate how factors such as test length, number of subtests, etc., affect the between-person and within-person reliability of ability estimates obtained from UIRT and MIRT models. As pointed out in Chapter 2, previous studies have indicated that the subscore estimates from MIRT models tend to be more reliable and precise than the subscore estimates from UIRT models when the subtests are short and the abilities are highly correlated (de la Torre, 2008, 2009; Wang, Chen, & Cheng, 2004; Yao, 2010; Yao & Boughton, 2007). This simulation study examined whether MIRT subscore estimation is still favorable over unidimensional subscore estimation in terms of between-person and within-person reliability under various test conditions. The following section explains the details of simulation conditions used in this study.

Simulation Conditions

In this study, there were three simulation conditions chosen based on the suggestions from previous MIRT studies. These conditions were test length, number of subtests, and correlation between dimensions.

a) *Test length:* In earlier studies, Tate (2004) used 12 items as the lowest number and 30 as the highest number of items for each subtest. Yao (2010) designed a simulation study with minimum 20 and maximum 60 items for each subtest. De la Torre, Song and Hong (2011) used 10, 20, and 30 items as test length. In this study, the number of items for each subtest was 10, 20, and 40, representing short, moderate, and long subtests. Each subtest had the same number of items.

b) *Number of subtests:* In previous simulation studies, the number of subtests (i.e., dimensions) ranged from 2 to 5 (De la Torre, 2008; De la Torre, Song & Hong, 2011; Tate, 2004; Yao, 2010). Considering the number of subtests in test batteries and similar tools (e.g., personality scales), the number of subtests was chosen to be 3, 5, or 7 in this study.

c) *Correlation between dimensions:* The size of correlations between the dimensions (i.e., subscores from different subtests) was .3, .5, or .8, representing low, moderate, and high subscore correlations. Correlations between the dimensions were the same. For instance, with three subtests, the correlations between dimensions one and two, dimensions one and three, and dimensions two and three were the same.

The three simulation conditions (test length, number of subtests, and correlation between dimensions) yielded 27 crossed conditions in total. A summary of the simulation

conditions in this study is presented in Figure 3.6. Because, sample size was found to have little effect on multidimensional and unidimensional ability estimation procedures (de la Torre & Patz, 2005; de la Torre & Song, 2009), a fixed sample size, *N*=1500, was used for all conditions in the simulation study. This sample size is considered a sufficient number of examinees to obtain stable estimates of subscores.

# of subtests	Test length	Correlation between dimensions		
		.3	.5	.8
3	10			
	20			
	40			
5	10			
	20			
	40			
7	10			
	20			
	40			

Figure 3.6. Simulation conditions of the study

Data Generation

As explained earlier, a simple structure was chosen for item parameters of the simulated tests to make a direct comparison of unidimensional and multidimensional models in terms of subscore reliability. Each subtest was assumed to be unidimensional while the overall test was multi-unidimensional. Each test item had a single difficulty parameter across all subtests and multiple discrimination parameters. The number of item discrimination parameters for each item was equal to the number of subtests. Because of the simple structure design, each item had a vector of item discriminations in which only one component was nonzero, as shown in Figure 3.2 above. Item discrimination

parameters were drawn from a uniform distribution, $a_i \sim$ U [0.8, 2.5]; item difficulty parameters were drawn from a normal distribution, $b_i \sim N$ (0, 1); and c parameters (i.e. lower asymptote) were drawn from a uniform distribution, $c_i \sim$ U [0, 0.25]. Different levels of the discrimination, difficulty, and c parameters reflect test items with low to high discrimination, low to high difficulty, and low to high guessing.

The true subscores were drawn from a multivariate normal distribution with a pre-specified variance-covariance matrix, $\theta_i \sim MVN$ (0, Σ). Based on the correlations between the dimensions explained under simulation conditions, mean vectors, variance vectors, and correlation matrices for the three-dimensional, five-dimensional, and seven-dimensional MIRT models were as follows:

a) Three-dimensional model: $\mu = \{0,0,0\} \quad \sigma = \{1,1,1\}$

$$\sum = \begin{bmatrix} 1 & .3 & .3 \\ .3 & 1 & .3 \\ .3 & .3 & 1 \end{bmatrix}_{3x3} \quad \text{or} \quad \sum = \begin{bmatrix} 1 & .5 & .5 \\ .5 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}_{3x3} \quad \text{or} \quad \sum = \begin{bmatrix} 1 & .8 & .8 \\ .8 & 1 & .8 \\ .8 & .8 & 1 \end{bmatrix}_{3x3}$$

b) Five-dimensional model: $\mu = \{0,0,0,0,0\} \quad \sigma = \{1,1,1,1,1\}$

$$\sum = \begin{bmatrix} 1 & .3 & .3 & .3 & .3 \\ .3 & 1 & .3 & .3 & .3 \\ .3 & .3 & 1 & .3 & .3 \\ .3 & .3 & .3 & 1 & .3 \\ .3 & .3 & .3 & .3 & 1 \end{bmatrix}_{5x5} \quad \sum = \begin{bmatrix} 1 & .5 & .5 & .5 & .5 \\ .5 & 1 & .5 & .5 & .5 \\ .5 & .5 & 1 & .5 & .5 \\ .5 & .5 & .5 & 1 & .5 \\ .5 & .5 & .5 & .5 & 1 \end{bmatrix}_{5x5} \quad \sum = \begin{bmatrix} 1 & .8 & .8 & .8 & .8 \\ .8 & 1 & .8 & .8 & .8 \\ .8 & .8 & 1 & .8 & .8 \\ .8 & .8 & .8 & 1 & .8 \\ .8 & .8 & .8 & .8 & 1 \end{bmatrix}_{5x5}$$

c) Seven-dimensional model: $\mu = \{0,0,0,0,0,0,0\}$ $\sigma = \{1,1,1,1,1,1,1\}$

$$\Sigma = \begin{bmatrix} 1 & .3 & .3 & .3 & .3 & .3 \\ .3 & 1 & .3 & .3 & .3 & .3 \\ .3 & .3 & 1 & .3 & .3 & .3 \\ .3 & .3 & .3 & 1 & .3 & .3 \\ .3 & .3 & .3 & .3 & 1 & .3 \\ .3 & .3 & .3 & .3 & .3 & 1 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & .5 & .5 & .5 & .5 & .5 \\ .5 & 1 & .5 & .5 & .5 & .5 \\ .5 & .5 & 1 & .5 & .5 & .5 \\ .5 & .5 & .5 & 1 & .5 & .5 \\ .5 & .5 & .5 & .5 & 1 & .5 \\ .5 & .5 & .5 & .5 & .5 & 1 \end{bmatrix} \Sigma = \begin{bmatrix} 1 & .8 & .8 & .8 & .8 & .8 \\ .8 & 1 & .8 & .8 & .8 & .8 \\ .8 & .8 & 1 & .8 & .8 & .8 \\ .8 & .8 & .8 & 1 & .8 & .8 \\ .8 & .8 & .8 & .8 & 1 & .8 \\ .8 & .8 & .8 & .8 & .8 & 1 \end{bmatrix}$$

To make the simulation results more comparable, a 10-item parameter set was duplicated twice to get parameters for the 20-item tests, and they were duplicated four times to get item parameters for the 40-item tests. Using the item parameters and true subscores described above, dichotomous item responses were generated in SimuMIRT (Yao, 2003) as follows:

1- First, item parameters were generated in R (R Development Core Team, 2010) based on the predefined characteristics.

2- The correct response probabilities for the items were computed for each person using the probability function of the M3PL model (see Equation 3.1).

3- A random number from a uniform distribution, U [0, 1], was drawn for each item. If the random number is smaller than the probability of correct response, then the item response was equal to 1 indicating a correct response, and if the random number was larger than the probability of correct response, the item response became 0, indicating an incorrect response.

4- For three subtests, 1500x30, 1500x60, and 1500x120 data matrices were generated; for five subtests, 1500x50, 1500x100, and 1500x200 data matrices were generated; and for seven subtests, 1500x70, 1500x140, and 1500x280 data matrices generated.

5- For the MIRT model, the data, including all subtests, were used for simultaneous estimation of subscores. For the UIRT model, the simulated response datasets were separated for each subtest.

600 replications (300 x 2 parallel forms) were carried out for each cell in Figure 3.6. Both item and person parameters were redrawn in each replication. As explained above, to estimate between-person and within-person reliability indices for each test, two parallel test forms are required. Therefore, step 3 in the data generation procedure was repeated twice, which yielded two response datasets based on the same item parameters and true subscores. These response datasets were used as parallel forms. As a result of the data generation process, 600 response datasets (300 replications x 2 parallel forms) were simulated for each crossed condition in Figure 3.7. An example syntax file for SimuMIRT (Yao, 2003) is shown in Appendix A1.

Subscore Estimation Procedure

For the estimation of subscores, true item parameters were used for both MIRT and UIRT models. The purpose of using the true item parameters was to eliminate additional errors in the subscore estimates due to the estimation error of the item parameters. Using the true item parameters and response datasets for parallel test forms, the subscores were estimated with BMIRT (Yao, 2003). BMIRT is a Bayesian software program that allows the estimation of item parameters and person abilities for both unidimensional and multidimensional IRT models. When estimating subscores, the multidimensional MAP estimation procedure in BMIRT allows for providing standard normal, noninformative, or strong priors to improve the precision of the ability estimates.

The syntax files for multidimensional and unidimensional subscore estimation in BMIRT are illustrated in Appendix A2.

The subscores from the M3PL model (see Equation 3.1) were estimated for each test form using the true item parameters based on a simple structure and strong priors. The strong priors were the mean vector for the subscore estimates and the variance-covariance matrix of the subscores. The strong priors were based on the generating distribution of the true subscores. Using strong priors for the population mean, variance, and inter-dimensional correlations would allow the information to be borrowed from one dimension to improve the precision of another dimension (Yao, 2013).

The subscores from the unidimensional 3PL model were also estimated using the true item parameters. For the unidimensional subtests, each subtest had a separate set of item parameters (see Figure 3.7). In addition to item parameters, each subtest in the simulated response datasets was saved as a separate data file to be used in the unidimensional subscore estimation in BMIRT. Noninformative priors were applied by entering prior values for the mean and variance of the unidimensional subscores. Because each subtest was analyzed separately, the correlated information between subtests was ignored.

Item	Subtest 1			Subtest 2			Subtest 3		
	a_j	b_j	c_j	a_j	b_j	c_j	a_j	b_j	c_j
1	1.47	-1.12	0.16	1.23	1.26	0.13	1.23	0.96	0.23
2	1.67	0.85	0.21	0.98	0.58	0.17	1.35	0.85	0.19
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
18	2.14	-0.36	0.07	1.44	-0.67	0.08	1.89	1.26	0.03
19	2.11	-1.45	0.12	1.87	-0.05	0.05	1.96	-0.79	0.09
20	1.76	0.66	0.15	1.77	1.48	0.11	1.48	0.08	0.12

Figure 3.7. A sample set of item parameters from three unidimensional subtests with 20 items.

The estimation of unidimensional and multidimensional subscores in BMIRT was implemented using the computers of the Minnesota Supercomputing Institute (MSI). The computer system was a Dell PowerEdge R710s with 2 quad-core 2.66 GHz processors and 48 GB memory, running 64-bit Windows Server 2008 R2. While the unidimensional subscore estimation was very quick regardless of test length, the multidimensional subscore estimation, especially for five- and seven-dimensional models, was computationally very intensive and it required more estimation time.

Evaluation Criteria

This section explains the criteria used for the evaluation of the data simulation procedure and the reliability of subscores estimated from the MIRT and UIRT models. First, the simple structure of simulated response datasets and the descriptive statistics of the true subscores (i.e., mean, variance, and the correlations among the true subscores) were checked to verify the accuracy of the data simulation procedure. Then, the methods for evaluating the between-person and within-person reliability of the estimated subscores are discussed.

Evaluation of Simulation Design. Several researchers have conducted factor analyses to determine whether the subtests of a test are distinct enough to estimate subscores (e.g., Stone, Ye, Zhu, & Lane, 2010; Wainer et al., 2001; Sinharay, Haberman, & Puhan, 2007). In terms of the subscores, the purpose of factor analytic approaches is to discover the underlying factor structure of a test including several subtests. Exploratory factor analysis (EFA) can be used for the evaluation of subscores in terms of distinctness based on the eigenvalues from the correlation matrix of the subscores (Sinharay, Puhan, & Haberman, 2011). In addition to EFA, confirmatory factor analysis (CFA) can be implemented when the number of dimensions is assumed to be known. Although CFA has been derived from the CTT framework, the parameterization of factor loadings, thresholds, and factor scores in non-linear factor analysis software programs, such as Mplus (Muthén & Muthén, 1998-2011), is very similar to the parameterization of item discrimination, item difficulty, and person abilities in IRT. CFA uses the factor loadings to indicate the relationship between the indicator variable (i.e., item) and the latent variable (i.e., subscore) across all levels of the latent variable (Osteen, 2010). Also, CFA provides a variety of fit indices for evaluating model-data fit, which may provide some insight regarding the test structure.

In this study, CFA was used for the examination of whether the simulated response datasets display a simple structure as specified in the data simulation procedure. A CFA model in which the items of each subtest were loaded on a single dimension was implemented for each dataset using Mplus (Muthén & Muthén, 1998-2011). CFI, TLI, and RMSEA fit indices were used to evaluate the fit of the simple structure model. TLI and CFI values greater than 0.90 are considered acceptable, and values greater than 0.95

are considered a good fit (Browne and Cudeck, 1993; Hu & Bentler, 1999; Kline, 2005). RMSEA values smaller than 0.07 are considered a close fit (Steiger, 2007).

After the confirmation of test structure, means and variances of the true subscores from the simulated response data were checked. Also, the correlations among the true subscores were obtained using Pearson's correlation coefficient. It is important to examine whether the true subscores follow the characteristics of the data generation distributions because these three components were used as the priors in the subscore estimation procedure. Descriptive statistics for the parallel forms were averaged over 300 replications for each simulation condition.

Evaluation of Subscore Reliability. As explained earlier, this study focused on the examination of subscore reliability based on the between-person and within-person variations in the estimated subscores. To make a comparison of the UIRT and MIRT models in terms of between-person and within-person subscore reliability under various simulation conditions, several evaluation criteria were considered. For the evaluation of subscore reliability, the evaluation criteria were as follows:

a) *The correlation of the subscores from two parallel forms*: After the subscores were estimated from the parallel test forms for both MIRT and UIRT models, the correlation of the subscores from the first test form and the second test form was examined using the Pearson correlation coefficient as follows:

$$r_{1.2} = \frac{1}{K}\sum_{k=1}^{K}\frac{cov(\theta_1, \theta_2)}{\sigma_{\theta_1}\sigma_{\theta_2}} = \frac{1}{K}\sum_{k=1}^{K} cor(\theta_1, \theta_2), \tag{3.24}$$

where θ_1 and θ_2 are the subscore estimates of the same subtest from test form 1 and test form 2, K is the number of replications, and $r_{1.2}$ is the average correlation over 300 replications. A higher correlation indicates that the subscore estimates from the parallel forms are similar.

b) *The correlation among the subscores within each test form*: In addition to the correlation of the subscores across test forms, the correlations among the subscore estimates within each test form were also computed for the UIRT and MIRT models. As in Equation 3.24, the correlations were averaged over 300 replications. The magnitude of these correlations indicated to what extent the resulting subscores had a meaningful relationship based on the factors they measure.

c) *The average total profile reliability, within-person reliability, and between-person reliability*: Using Equations 3.21, 3.22, and 3.23, total profile, within-person, and between-person reliability coefficients were computed and averaged over 300 replications for each simulation condition in R (R Development Core Team, 2012). The R code for computing total profile, between-person, and within-person reliability coefficients are presented in Appendix A3. Reliability estimation procedure was repeated for both unidimensional and multidimensional subscore estimates. Then, the magnitudes of the reliability estimates were compared across the two models.

d) *Sampling distributions of within-person, between-person, and total reliability coefficients*: Graphical illustrations of the sampling distributions of the reliability coefficients from 300 replications are presented. These graphical illustrations

show the differences in the UIRT and MIRT models in terms of within-person and between-person reliability under various simulation conditions.

e) *Repeated measures multivariate analysis of variance model (MANOVA)*: A repeated measures MANOVA was used to evaluate the effects of simulation conditions (test length, number of subtests, and true correlations among subscores) as between-factor variables and the type of estimation method (i.e., UIRT vs. MIRT) as a within-factor variable on between-person, within-person, and total profile reliability estimates. The main advantage of the repeated measures MANOVA is that it does not require equal variances and covariances among dependent measures (i.e., sphericity assumption) as opposed to univariate repeated measures analysis. For each predictor, partial eta squared (η^2) was reported as a measure of effect size. η^2 is the proportion of the total variance accounted for by each independent variable. For the within-subject factor (i.e., subscore estimation method), η^2 was computed based on the method described by Tabachnick and Fidell (2007). Using Wilk's lambda (Λ), the partial eta squared can be found as follows:

$$\eta^2 = 1 - \sqrt{\Lambda} \tag{3.25}$$

For the between-subject factors, the partial eta squared is the ratio of the sum of squares for the main effect of the factor (SS_{effect}) to the sum of squares for the total variance. Based on this definition, the partial eta squared becomes as follows:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}, \tag{3.26}$$

where SS_{total} is equal to $SS_{effect} + SS_{error}$ (Yon, 2006).

Real Data Study

In the last section of this study, to illustrate the between-person and within-person reliability procedures, a subset of the Entrance Examination for Graduate Studies (EEGS) was used. EEGS is a nationwide test that is used for the selection of students for graduate programs at the universities in Turkey. EEGS consists of three subtests: Quantitative 1, Quantitative 2 and Verbal. Quantitative 1 and Quantitative 2 subtests include 40 multiple-choice items that measure the mathematical and logical reasoning abilities of the examinees. The items in Quantitative 2 are designed to be more advanced than the items in Quantitative 1. The verbal subtest includes 80 multiple-choice items that measure the verbal reasoning ability. All items in EEGS have five response options, and they are scored dichotomously.

The data used in this study were from the 2008 administration of EEGS. A random sample of 10,000 examinees (5000 male, 5000 female) was selected from the full dataset. The sample includes examinees from 123 universities in Turkey and outside of Turkey. Examinees' ages ranged from 18 to 61. Table 3.1 shows the descriptive statistics of the raw subscores from EEGS.

Table 3.1

Summary Statistics for the Raw Subscores in the Three Subtests of EEGS

Subtest	# of items	*M*	*SD*	Min	Max	α
Quantitative 1	40	23.28	11.92	0	40	.96
Quantitative 2	40	18.36	13.31	0	40	.97
Verbal	80	59.72	16.66	0	80	.96

Note: α: Alpha reliability coefficient

In order to compute between-person and within-person reliability coefficients from the subtests of EEGS, two parallel forms were created from each subtest based on the test information function (TIF). TIF is basically the sum of the information functions of the items on a test, which can be computed as

$$I(\theta) = \sum_{j=1}^{J} I_j(\theta), \tag{3.27}$$

where $I(\theta)$ is the amount of test information at an ability level of θ, $I_j(\theta)$ is the amount of information for item j at ability level θ, and J is the number of items in the test (Baker, 2001). TIF can be used to design similar test forms and also to control measurement error very precisely within a test form.

To create two parallel test forms from each of the subtests of EEGS, item parameters for the three subtests were estimated using the unidimensional 3PL model. After obtaining the item parameters, test items with similar item difficulties and item information functions were placed into separate test forms. Each parallel form had the same number of items. Quantitative 1 and Quantitative 2 had two parallel test forms with 20 items, and the Verbal subtest had two parallel test forms with 40 items. The similarity of the resulting parallel forms was examined by plotting the TIF for each test form. Figure 3.9 shows test information functions for the parallel test forms based on the Quantitative 1, Quantitative 2, and Verbal subtests of EEGS. As seen in Figure 3.8, the parallel test forms indicated very similar test information functions.

In the subscore estimation procedure, the subscores were estimated for each of the parallel forms using the unidimensional 3PL model and the M3PL model. First, the

unidimensional subscore estimates were obtained for each parallel test form of Quantitative 1, Quantitative 2, and Verbal. Then, Pearson correlations among the unidimensional subscore estimates were found. In the next step, these correlations were used as strong priors in the multidimensional estimation of subscores. As in the simulation study, a simple structure was assumed again. The items of each subtest were loaded on a single dimension. Both unidimensional and multidimensional subscore estimates were obtained using BMIRT. Between-person, within-person, and total profile reliability coefficients were computed for the subscore estimates from each estimation method. Then, the magnitudes of these reliability coefficients were compared.

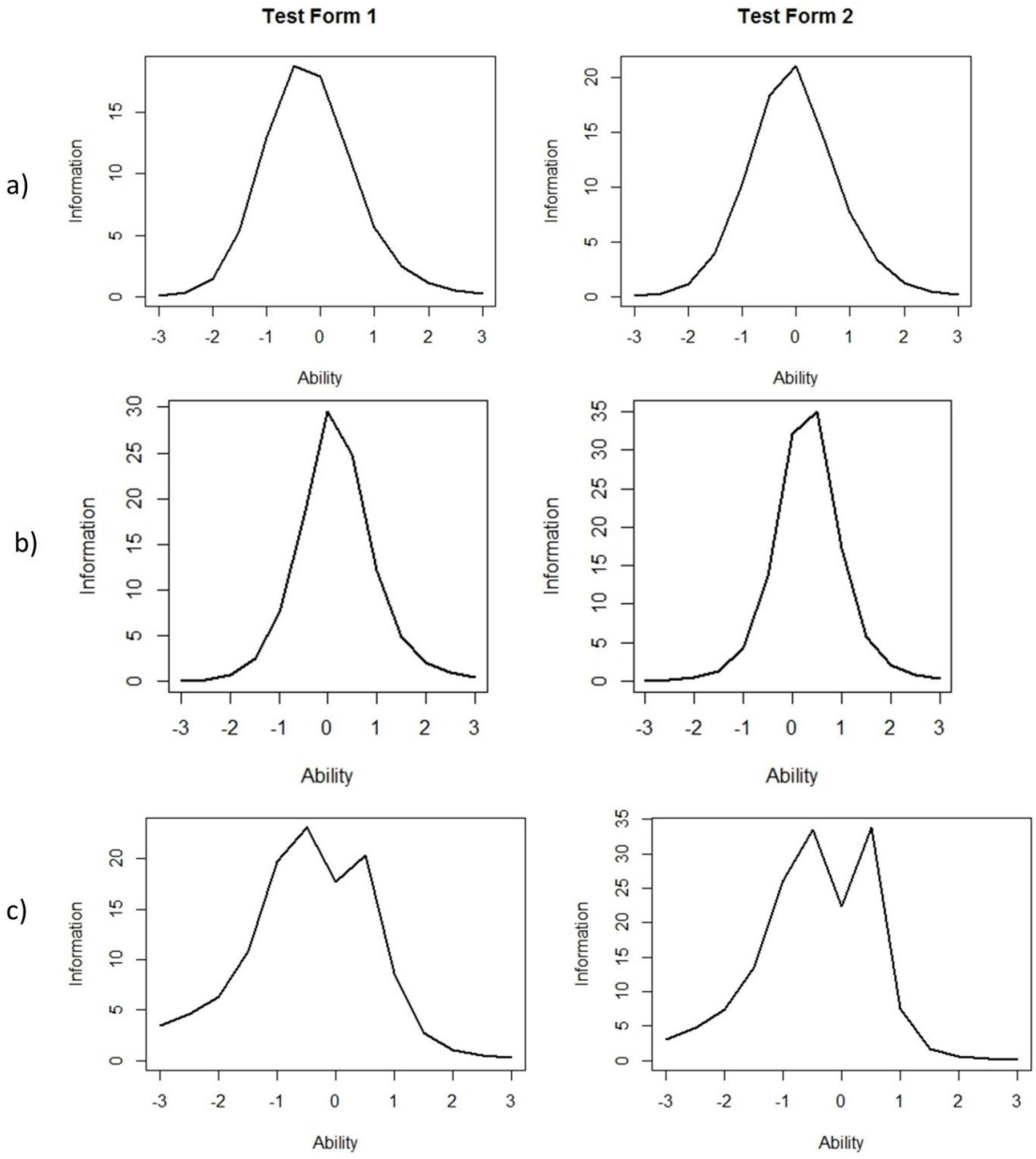

Figure 3.8. Test information functions of parallel test forms from Quantitative 1 (a), Quantitative 2 (b), and Verbal (c) subtests of EEGS

CHAPTER 4

RESULTS

This chapter presents the results from the simulation and real data studies described in the previous chapter. The first section summarizes the results from the simulation study in which the subscores from parallel test forms were estimated using the unidimensional and multidimensional IRT models. The simulated response datasets were examined in terms of their generating distributions. Then, the estimated subscores from the UIRT and MIRT models were compared based on between-person, within-person, and total profile reliability coefficients to address the research questions of this study. In the second section of this chapter, the results from a real data study are presented. Model comparisons are carried out to examine the differences between unidimensional and multidimensional subscore estimates of EEGS regarding subscore reliability.

Results of the Simulation Study

As explained in Chapter 3, several evaluation criteria were used for the inspection of the simulation design and the comparison of the UIRT and MIRT models. The simulated response datasets were examined based on the following criteria: (a) Correlations between the true subscores generated in SimuMIRT; and (b) the fit of CFA models based on simple structure. After checking the accuracy of the data simulation process, the subscore estimates from the unidimensional 3PL model and the M3PL model were compared based on between-person, within-person, and total profile reliability coefficients. The following section explains how the data generation procedure was

evaluated based on the underlying distributions of true subscores and the structure of simulated response datasets.

Means, Variances, and Correlations of True Subscores

The simulation study assumed three different correlations between the true subscores (.3, .5, and .8) to represent low, medium, and high correlations of the subscores. The correlation matrices for the three-dimensional, the five-dimensional, and the seven-dimensional models were as follows:

$$R = \begin{pmatrix} 1 & .3 & .3 \\ .3 & 1 & .3 \\ .3 & .3 & 1 \end{pmatrix} \quad R = \begin{pmatrix} 1 & .3 & .3 & .3 & .3 \\ .3 & 1 & .3 & .3 & .3 \\ .3 & .3 & 1 & .3 & .3 \\ .3 & .3 & .3 & 1 & .3 \\ .3 & .3 & .3 & .3 & 1 \end{pmatrix} \quad R = \begin{pmatrix} 1 & .3 & .3 & .3 & .3 & .3 & .3 \\ .3 & 1 & .3 & .3 & .3 & .3 & .3 \\ .3 & .3 & 1 & .3 & .3 & .3 & .3 \\ .3 & .3 & .3 & 1 & .3 & .3 & .3 \\ .3 & .3 & .3 & .3 & 1 & .3 & .3 \\ .3 & .3 & .3 & .3 & .3 & 1 & .3 \\ .3 & .3 & .3 & .3 & .3 & .3 & 1 \end{pmatrix}$$

For the other two correlations, the value of .3 was replaced with either .5 or .8. In addition, the correlations between the true subscores, means, and variances of the true subscores were specified in the data generation process. For all dimensions, the subscores followed a normal distribution with a mean of 0 and variance of 1.

In order to check if the true subscores were accurately generated based on the generating distributions, the correlations between the true subscores as well as the means and variances of the true subscores were computed under each simulation condition, and the average values for 300 replications were reported. Since the parallel test forms were generated based on the same true subscores, the data inspection procedure was implemented once. Tables 4.1, 4.2, and 4.3 show the average correlations between the true subscores when the test consisted of three, five, and seven subtests.

Results indicated that the correlations among the true subscores were very close to the hypothesized correlations (i.e., .3, .5, and .8) used for generating the multivariate subscores. In addition to the correlations, the means and variances of the true subscores were examined. These two statistics were particularly important because they were entered as the population priors in the subscore estimation procedure. As explained earlier, the subscores were designed to have a multivariate normal distribution. Results indicated that the average mean and variance values over 300 replications were very similar to the values from a multivariate normal distribution. The off-diagonal elements of Tables 4.1, 4.2, and 4.3 show the variance of each dimension (i.e., subscore). These values were very close to 1 in all conditions. The mean of the true subscores ranged between -.004 and .003 across all crossed conditions, indicating that the true subscores followed the generating distributions closely. Also, testing of the generating mean values on a random set of datasets indicated that the mean values were statistically no different than the generating parameters.

Evaluation of Simple Structure

To evaluate if the simulated response datasets display a simple structure, CFA models were fitted to the simulated datasets in Mplus (Muthén & Muthén, 1998-2011). CFI, TLI, and RMSEA fit indices were used to evaluate whether there was an adequate model-data fit. In the CFA models, the subtest items were defined as categorical variables, and each item was loaded on a single dimension to define a simple structure (see Figure 4.1). The CFA models were estimated with a mean- and variance-adjusted weighted least squares (WLSMV) estimator that uses a robust weighted least squares approach for categorical variables in Mplus.

Table 4.1

Correlation Matrix of the True Subscores from Three Subtests

ρ		S1	S2	S3
	S1	0.999		
.3	S2	.301	0.997	
	S3	.300	.300	0.998
	S1	0.999		
.5	S2	.502	0.997	
	S3	.501	.500	0.998
	S1	0.998		
.8	S2	.800	0.999	
	S3	.801	.800	0.998

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3. Diagonal elements are variances and off-diagonal elements are correlations over 300 replications. ρ: True correlation between subscores used in the data generation.

Table 4.2

Correlation Matrix of the True Subscores from Five Subtests

ρ		S1	S2	S3	S4	S5
	S1	.988				
	S2	.290	1.004			
.3	S3	.290	.290	1.002		
	S4	.300	.310	.300	.998	
	S5	.290	.310	.300	.300	1.003
	S1	0.988				
	S2	.490	1.001			
.5	S3	.490	.490	0.989		
	S4	.500	.500	.500	0.997	
	S5	.500	.510	.500	.500	1.002
	S1	0.988				
	S2	.790	0.993			
.8	S3	.790	.790	0.997		
	S4	.800	.800	.800	0.993	
	S5	.800	.800	.800	.800	0.993

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5. Diagonal elements are variances and off-diagonal elements are correlations over 300 replications. ρ: True correlation between subscores used in the data generation.

Table 4.3

Correlation Matrix of the True Subscores from Seven Subtests

ρ		S1	S2	S3	S4	S5	S6	S7
	S1	0.998						
	S2	.290	0.995					
	S3	.300	.300	1.001				
.3	S4	.290	.300	.290	1.001			
	S5	.300	.290	.290	.300	0.999		
	S6	.300	.300	.300	.300	.300	0.996	
	S7	.310	.300	.300	.300	.300	.300	1.002
	S1	0.999						
	S2	.490	0.996					
	S3	.500	.500	1.000				
.5	S4	.490	.500	.490	0.999			
	S5	.500	.490	.490	.500	0.999		
	S6	.500	.500	.500	.500	.500	0.998	
	S7	.500	.490	.500	.500	.500	.500	1.001
	S1	0.999						
	S2	.790	0.996					
	S3	.800	.800	1.000				
.8	S4	.800	.800	.790	0.999			
	S5	.800	.790	.800	.800	0.999		
	S6	.800	.800	.800	.800	.800	0.998	
	S7	.800	.800	.800	.790	.800	.800	1.001

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore 7. Diagonal elements are variances and off-diagonal elements are correlations over 300 replications. ρ: True correlation between subscores used in the data generation.

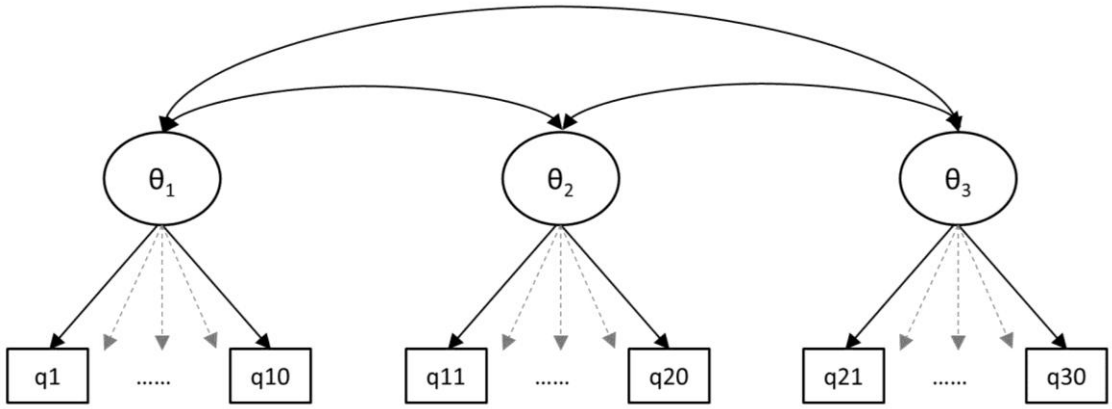

Figure 4.1. A simple-structure CFA model based on three subtests and thirty test items.

To evaluate model-data fit, the following criteria were used: CFI > .95; TLI > .95; RMSEA < .07. The model fit indices were averaged over 300 replications. Because the parallel test forms were based on the same true subscores and the distribution characteristics, only one set of the forms was evaluated in terms of the factor structure. Table 4.4 presents the findings from the CFA models.

The results from the CFA models show that the response datasets indicated adequate fit based on the fit indices. Under all conditions, the simulated datasets provided CFI and TLI values higher than .95, and RMSEA smaller than .05. CFI and TLI values were very close across different levels of subtest length and inter-dimension correlations. However, as the number of subtests (i.e., dimensions) increased, CFI and TLI decreased because of the increasing complexity of the CFA models. As the correlation among the dimensions increased, RMSEA also increased slightly, indicating that the CFA models with low correlated dimensions (r = .3) provided a better fit compared to the CFA models with moderately (r = .5) and highly (r = .8) correlated dimensions.

Table 4.4

Summary of Model-Fit Statistics from CFA Models Used for Testing Simple Structure

N of Subtests	Subtest Length	Correlation	CFI	TLI	RMSEA
3	10	.3	.998	.999	.003
3	10	.5	.998	.998	.004
3	10	.8	.998	.999	.005
3	20	.3	.998	.999	.002
3	20	.5	.998	.998	.003
3	20	.8	.997	.997	.004
3	40	.3	.988	.986	.001
3	40	.5	.985	.986	.002
3	40	.8	.983	.981	.003
5	10	.3	.978	.976	.011
5	10	.5	.977	.976	.013
5	10	.8	.975	.977	.016
5	20	.3	.973	.976	.009
5	20	.5	.974	.975	.011
5	20	.8	.975	.974	.014
5	40	.3	.973	.971	.008
5	40	.5	.971	.971	.011
5	40	.8	.972	.971	.012
7	10	.3	.964	.963	.021
7	10	.5	.963	.963	.023
7	10	.8	.963	.962	.026
7	20	.3	.959	.958	.019
7	20	.5	.958	.957	.021
7	20	.8	.958	.957	.022
7	40	.3	.953	.951	.017
7	40	.5	.953	.952	.019
7	40	.8	.952	.952	.020

Note: CFI, TLI, and RMSEA values in the table are the average of 300 replications.

The second step of the simulation study was the comparison of between-person, within-person, and total profile reliability estimates obtained from the UIRT and MIRT models. The subscores for each response dataset were estimated using multidimensional and unidimensional IRT scoring methods. The following sections present the findings of subscore reliability estimates from the two methods, and explain how the subscore reliability estimates are influenced by the simulation conditions. In addition, the results for the recovery of correlations between the subscores and the correlation between the parallel test forms are presented.

Correlations between Parallel Test Forms

As explained in Chapter 3, between-person, within-person, and total profile reliability coefficients are estimated using parallel test forms. Assuming that subscores from two parallel test forms have equal variances, the true score variation becomes equal to the covariances of subscores from the parallel test forms. True score variations among the individual subscore estimates can be used for estimating an overall profile reliability coefficient as well as between-person and within-person subscore reliability coefficients.

In this study, to create parallel test forms, the same item parameters and true subscores were used for generating two datasets. The random seed option in SimuMIRT was modified between the first and second data generation procedures to simulate different response files based on the same item and person parameters. After the subscore estimates were obtained for each response dataset using the UIRT and MIRT models, correlations between the subscore estimates from the two test forms were computed. Table 4.5 and Table 4.6 present the correlations of the subscores across the parallel test forms from the multidimensional and unidimensional scoring, respectively.

Table 4.5

Correlations of the Subscore Estimates from MIRT across Two Parallel Test Forms

Subtests	Subtest Length	ρ	S1	S2	S3	S4	S5	S6	S7
3		.3	.77	.77	.77				
	10	.5	.79	.79	.79				
		.8	.85	.85	.85				
		.3	.86	.86	.86				
	20	.5	.87	.87	.87				
		.8	.90	.90	.90				
		.3	.92	.92	.92				
	40	.5	.93	.92	.92				
		.8	.94	.94	.94				
5		.3	.78	.77	.78	.78	.78		
	10	.5	.81	.81	.81	.81	.81		
		.8	.88	.88	.88	.88	.88		
		.3	.86	.86	.87	.86	.87		
	20	.5	.88	.88	.88	.88	.88		
		.8	.92	.92	.92	.92	.92		
		.3	.92	.92	.92	.92	.92		
	40	.5	.92	.93	.92	.93	.93		
		.8	.94	.95	.94	.94	.95		
7		.3	.79	.79	.79	.79	.79	.79	.79
	10	.5	.82	.83	.82	.82	.82	.82	.83
		.8	.89	.89	.89	.89	.89	.89	.89
		.3	.88	.88	.88	.88	.87	.88	.88
	20	.5	.91	.91	.91	.92	.91	.91	.91
		.8	.92	.92	.92	.92	.92	.92	.92
		.3	.93	.93	.93	.93	.93	.93	.93
	40	.5	.94	.94	.94	.94	.94	.94	.94
		.8	.95	.95	.94	.95	.95	.95	.95

Note: ρ: Correlation between the subscores. S1: Subscore 1; S2: Subscore2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore 7.

Table 4.6

Correlations of the Subscore Estimates from UIRT across Two Parallel Test Forms

Subtests	Subtest Length	ρ	S1	S2	S3	S4	S5	S6	S7
3		.3	.74	.74	.74				
	10	.5	.74	.74	.74				
		.8	.74	.74	.74				
		.3	.85	.85	.85				
	20	.5	.85	.85	.85				
		.8	.85	.85	.85				
		.3	.92	.92	.92				
	40	.5	.92	.91	.92				
		.8	.92	.92	.92				
5		.3	.74	.73	.74	.74	.75		
	10	.5	.74	.73	.74	.74	.75		
		.8	.74	.73	.73	.74	.74		
		.3	.85	.85	.85	.85	.85		
	20	.5	.85	.85	.85	.85	.85		
		.8	.85	.85	.85	.85	.85		
		.3	.91	.92	.91	.92	.92		
	40	.5	.91	.92	.91	.91	.92		
		.8	.91	.92	.91	.91	.92		
7		.3	.74	.74	.74	.74	.74	.74	.74
	10	.5	.74	.74	.74	.74	.74	.74	.74
		.8	.74	.74	.74	.74	.74	.74	.74
		.3	.85	.85	.85	.85	.85	.85	.85
	20	.5	.85	.85	.85	.85	.85	.85	.85
		.8	.85	.85	.85	.85	.85	.85	.85
		.3	.92	.91	.92	.92	.92	.92	.92
	40	.5	.92	.91	.92	.92	.91	.92	.92
		.8	.92	.91	.92	.92	.91	.92	.92

Note: ρ: Correlation between the subscores. S1: Subscore 1; S2: Subscore2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore 7.

The results indicated that the multidimensional subscore estimates from the parallel test forms had higher correlations than the unidimensional subscore estimates under all simulation conditions. The correlation between the parallel forms ranged from .77 to .95 for the multidimensional subscore estimation, and from .74 to .92 for the unidimensional subscore estimation. The difference between the two methods was the largest when the number of items was small and the correlations between the subscores were high. The number of test items had a positive effect on the correlations between the parallel forms for both methods. As the number of items increased from 10 to 40 in each subtest, the correlation between the parallel test forms dramatically increased. The number of subtests had no impact on the magnitude of the correlations between the parallel forms for both unidimensional and multidimensional scoring methods. Figure 4.2 shows the relationship between subtest length, the true correlation between the subscores, and the correlation between the parallel forms when the multidimensional subscore estimation is used.

The results also suggest that the correlation between the subscores had an impact only when the multidimensional subscore estimation was used. As the correlation between the subscores increased from .3 to .8 within each test form, the correlation of the subscore estimates between the parallel forms increased substantially. In contrast to the multidimensional subscore estimation, the correlations between the parallel test forms remained constant across different levels of the correlations among the subscores within each test form when unidimensional scoring was used.

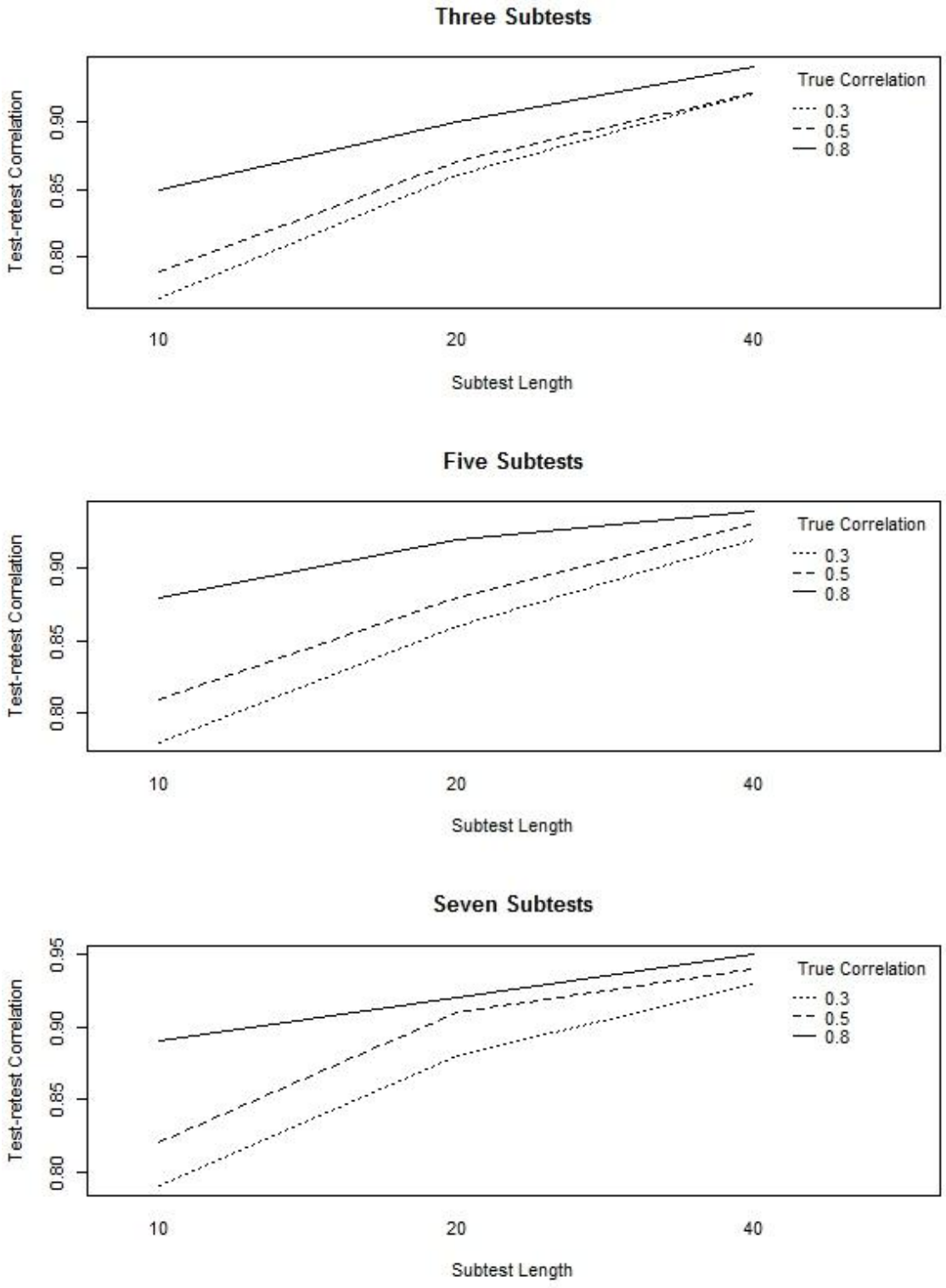

Figure 4.2. The average test-retest correlations of the multidimensional subscores across three levels of subtest length and true correlations between the subscores.

The main reason for the difference between the two estimating procedures is that the multidimensional subscore estimation takes the correlations among the subscores into account as strong population priors, whereas the unidimensional subscore estimation estimates the subscores from each subtest one at a time by ignoring the correlation between the subscores. Incorporating the correlational structure in the subscore estimation provided moderately large to large improvements (.08–.10) when tests were short and $\rho = .8$, and small improvements (.02–.04) when tests were longer and $\rho = .3$. Despite its positive effects, the number of items had still greater impact than the size of the correlation between the subscores when multidimensional scoring was used.

Correlations of Estimated Subscores

The correlations of the multidimensional subscore estimates across all simulation conditions are presented in Table B1.1 through Table B1.6 in Appendix B1. The results indicated that correlations among the multidimensional subscores were overestimated under all simulation conditions. The estimated correlations were higher than the true correlations used for generating the subscores. The estimated correlations did not change across different ability dimensions (i.e., subscores), indicating that recovery of the true correlations based on the MIRT model was the same across all dimensions. The number of items (i.e., subtest length) had an impact on the recovery of true correlations. The largest discrepancy between the true correlations and the estimated correlations occurred when the subtests were short (e.g., 10 items per subtest). As the subtest length increased, the correlations of the multidimensional subscore estimates became closer to the true correlations among the subscores. In addition to subtest length, recovery of the true correlations also depends on the true correlations among the subscores. When the true

correlation among the subscores was small (ρ=.3), the discrepancy between the true correlations and the estimated correlations were the smallest. As the true correlations increased from .3 to .8, the correlations of the multidimensional subscore estimates indicated larger deviations from the true correlations. The number of subtests (i.e., dimensions) did not seem to affect the observed correlations among the subscores. Across the 3-dimensional, 5-dimensional, and 7-dimensional models, correlations among the multidimensional subscore estimates did not change for all ability dimensions.

The correlations of the unidimensional subscore estimates across all simulation conditions are presented in Table B2.1 through Table B2.6 in Appendix B2. In contrast to the multidimensional subscore estimates, correlations among the unidimensional subscore estimates were always smaller than the true correlations among the subscores under all simulation conditions. Also, the discrepancy between the true subscore correlations and the observed subscore correlations was larger for the unidimensional subscore estimates than the multidimensional subscore estimates, indicating that the MIRT model recovered the true correlations better than the UIRT model.

As for the multidimensional subscore estimation, subtest length was influential on the recovery of the true subscore correlations. Especially when the number of subtest items was small (i.e., 10 items for each subtest) and the true correlation among the subscores was high (ρ=.8), the correlations among the unidimensional subscore estimates were fairly low compared with the true correlations. When subtest length increased from 10 to 40, the recovery of the true correlations substantially improved and the discrepancy between the estimated and true correlations became smaller. The best approximations of the true correlations were observed when subtest length was long (i.e., 40 items for each

subtest) and the true correlation among the subscores was small (ρ=.3). The number of subtests (i.e., dimensions) was again not an important factor for the recovery of true correlations among the subscores. For both multidimensional and unidimensional subscore estimates, the recovery of the true correlations was very similar across the parallel test forms.

Between-person, Within-person, and Total Profile Reliability Estimates

Descriptive summaries of subscore reliability estimates based on the multidimensional and unidimensional estimation procedures are presented in Table 4.7 and Table 4.8, respectively. For each crossed simulation condition, subscore reliability estimates included a sample of 300 estimates resulting from 300 replications.

The results in Table 4.7 and Table 4.8 indicated that between-person subscore reliability estimates were higher than within-person subscore reliability estimates across all simulation conditions for both multidimensional and unidimensional subscores. The average between-person subscore reliability ranged from .83 to .98 for the multidimensional subscore estimates and it ranged from .82 to .98 for the unidimensional subscore estimates. For both unidimensional and multidimensional subscores, the magnitude of between-person reliability estimates was found to be positively associated with subtest length, number of subtests, and correlations among the subscores. As subtest length, number of subtests, and correlations among the subscores increased, between-person reliability substantially improved. The highest between-person reliability ($\rho_B =$ .98) in the multidimensional subscore estimates was obtained when there were seven subtests, each of which included 40 items, and the correlations between the seven subtests were high (ρ=.8). Similarly, the highest between-person reliability estimate was

.98 for the unidimensional subscore estimates when the subtest length was 40 and the true correlations among the subscore estimates were .8.

Table 4.7

Descriptive Statistics for Between-Person, Within-Person, and Total Profile Reliability Estimates from the Multidimensional Subscore Estimates across Simulation Conditions.

Subtests	Subtest Length	ρ	ρ_B		ρ_W		ρ_T	
			M	*SD*	*M*	*SD*	*M*	*SD*
3	10	.3	.831	.015	.689	.021	.771	.016
		.5	.859	.012	.617	.023	.792	.014
		.8	.885	.010	.406	.026	.851	.010
	20	.3	.905	.008	.809	.014	.862	.011
		.5	.921	.007	.754	.017	.872	.009
		.8	.936	.005	.567	.025	.904	.006
	40	.3	.947	.004	.888	.008	.921	.005
		.5	.957	.003	.852	.011	.924	.005
		.8	.965	.002	.711	.019	.939	.003
5	10	.3	.868	.025	.689	.029	.776	.021
		.5	.898	.017	.619	.036	.806	.021
		.8	.922	.014	.410	.045	.878	.017
	20	.3	.928	.013	.808	.019	.865	.013
		.5	.946	.009	.776	.023	.879	.013
		.8	.958	.007	.571	.037	.918	.011
	40	.3	.961	.008	.887	.012	.921	.008
		.5	.971	.005	.852	.016	.926	.008
		.8	.977	.004	.709	.031	.944	.021
7	10	.3	.878	.031	.689	.037	.786	.025
		.5	.908	.022	.619	.049	.821	.027
		.8	.932	.018	.410	.059	.888	.024
	20	.3	.938	.018	.808	.024	.901	.018
		.5	.949	.014	.759	.029	.913	.016
		.8	.958	.012	.578	.046	.921	.014
	40	.3	.958	.011	.893	.016	.922	.011
		.5	.967	.007	.851	.021	.931	.011
		.8	.979	.006	.712	.043	.941	.036

Note: ρ: True correlation between the subscores. ρ_B: Between-person reliability; ρ_W: Within-person reliability; ρ_T: Total profile reliability.

Table 4.8

Descriptive Statistics for Between-Person, Within-Person, and Total Profile Reliability Estimates from the Unidimensional Subscore Estimates across Simulation Conditions.

Subtests	Subtest Length	ρ	ρ_B		ρ_W		ρ_T	
			M	*SD*	*M*	*SD*	*M*	*SD*
3	10	.3	.817	.020	.669	.026	.739	.023
		.5	.848	.017	.594	.028	.739	.023
		.8	.879	.014	.379	.031	.739	.023
	20	.3	.899	.009	.799	.014	.849	.011
		.5	.918	.007	.741	.018	.849	.011
		.8	.936	.006	.538	.024	.849	.011
	40	.3	.946	.004	.884	.009	.916	.006
		.5	.956	.003	.845	.012	.915	.006
		.8	.965	.003	.688	.019	.915	.006
5	10	.3	.858	.026	.667	.035	.737	.027
		.5	.890	.018	.594	.041	.737	.027
		.8	.918	.016	.382	.054	.737	.027
	20	.3	.926	.015	.798	.021	.849	.015
		.5	.943	.010	.741	.025	.849	.015
		.8	.960	.007	.541	.041	.847	.017
	40	.3	.959	.008	.883	.013	.915	.009
		.5	.969	.005	.845	.016	.915	.009
		.8	.978	.004	.687	.032	.914	.009
7	10	.3	.873	.029	.673	.043	.741	.031
		.5	.901	.022	.594	.045	.741	.031
		.8	.931	.019	.389	.058	.741	.031
	20	.3	.931	.021	.801	.028	.852	.019
		.5	.939	.016	.743	.032	.851	.019
		.8	.958	.013	.542	.045	.851	.019
	40	.3	.971	.013	.879	.019	.919	.012
		.5	.983	.011	.847	.022	.920	.012
		.8	.981	.009	.689	.039	.919	.012

Note: ρ: True correlation between the subscores. ρ_B: Between-person reliability; ρ_W: Within-person reliability; ρ_T: Total profile reliability.

The average within-person subscore reliability estimates ranged from .41 to .89 for the MIRT model and from .38 to .88 for the UIRT model. In contrast to between-person subscore reliability, within-person subscore reliability was negatively associated with the true correlations among both the subscore estimates. As the true correlations increased from .3 to .8, within-person reliability estimates from the MIRT and UIRT models became fairly small. Within-person reliability estimates were the lowest for both multidimensional and unidimensional subscore estimates especially when the true correlation among the subscores was .8. This finding implies that high correlations among the subscore estimates leads to low within-person reliability as a result of smaller variations among the subscores. Therefore, when subtests are very similar (i.e., highly correlated), the estimated subscores may not be a reliable indicator of the within-person variation among the subscores.

Within-person reliability was also dependent on the number of items in the subtests. As subtest length increased from 10 to 40, within-person reliability improved substantially for both multidimensional and unidimensional subscore estimates. The highest average within-person reliability ($\rho_B = .89$) in the multidimensional subscore estimates was obtained when subtest length was 40 for each subtest and the true correlation among the subscores was low (ρ=.3). Similarly, the highest within-person reliability estimate ($\rho_B = .88$) was obtained for the unidimensional subscore estimates when subtest length was 40 and the true correlations among the subscore estimates were .3. Number of subtests did not seem to have an impact on within-person reliability. Within-person reliability estimates obtained from the MIRT and UIRT models did not change when the number of subtests increased from 3 to 7.

As mentioned earlier, total profile reliability is a weighted average of between-person and within-person reliability coefficients. Results of the simulation study showed that total profile reliability estimates always lie between the values of within-person reliability and between-person reliability. For the UIRT model, estimates of total profile reliability were only dependent on subtest length. Correlations among the subscores and the number of subtests did not affect total profile reliability. In contrast to the UIRT model, total profile reliability from the MIRT model was positively associated with both subtest length and correlations among the subscores. As in the UIRT model, total profile reliability estimates remained constant across different numbers of subtests in the MIRT model. Figures 4.3, 4.4, and 4.5 show the interaction between between-person reliability, within-person reliability, and total profile and simulation conditions for three subtests, five subtests, and seven subtests, respectively.

Density plots in Appendix C1 show the sampling distributions of reliability estimates from the MIRT and UIRT models across the simulation conditions. These plots showed that reliability estimates had a larger variation when subtest length was short (i.e., 10 items) and the true correlations among the subscores were low (ρ=.3). Under these conditions, between-person, within-person, and total profile reliability estimates showed very similar distributions. However, as subtest length and true correlations among the subscores increased, variation in between-person and total profile reliability estimates decreased and the distributions became narrower. Especially when the correlations among the subscores were high, the distributions of between-person reliability estimates indicated high kurtosis with a sharper peak and fatter tails while the distributions of within-person reliability estimates indicated smaller kurtosis with longer tails.

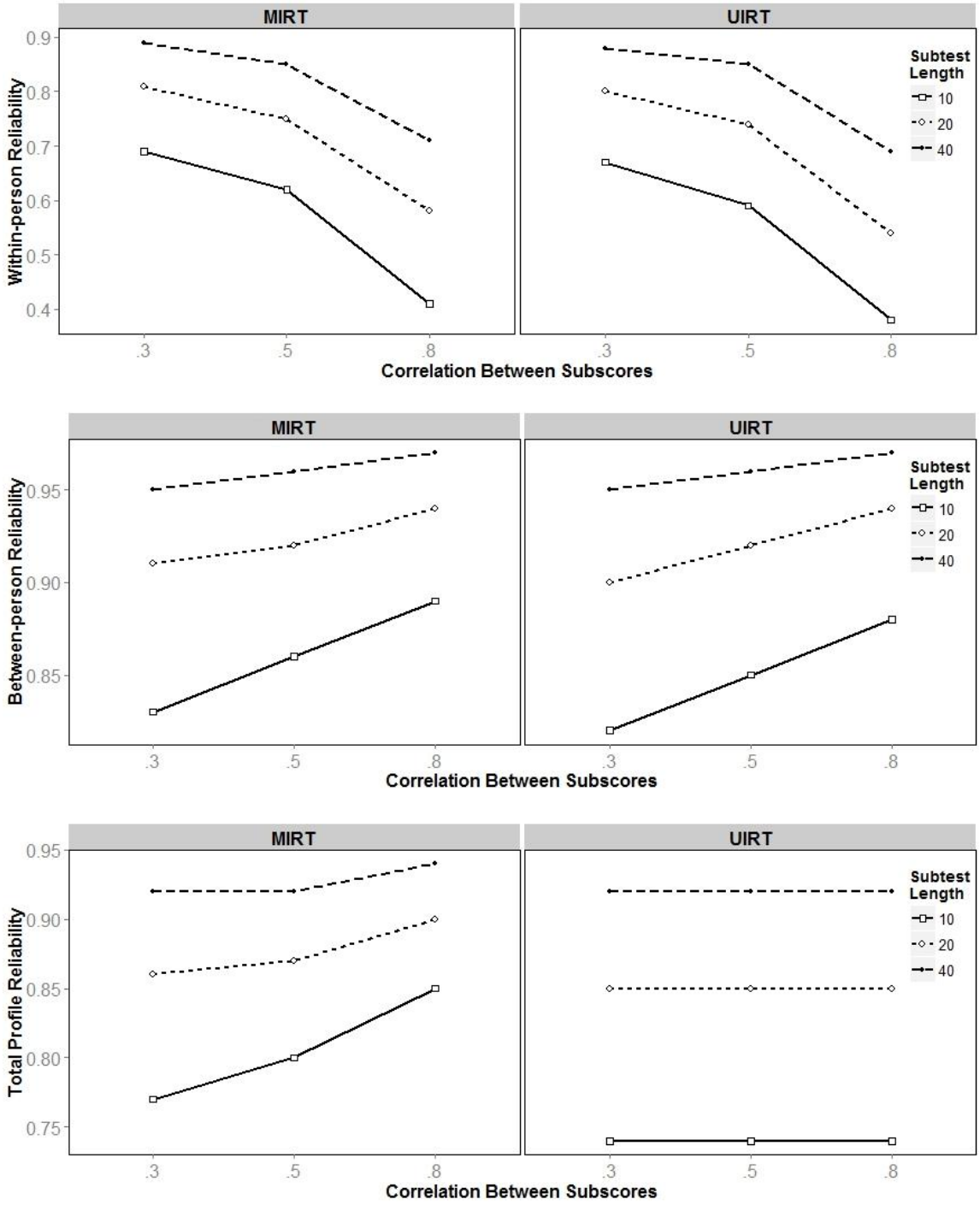

Figure 4.3. Interaction between estimated subscore reliability coefficients, correlations among subscores, and subtest length for three subtests.

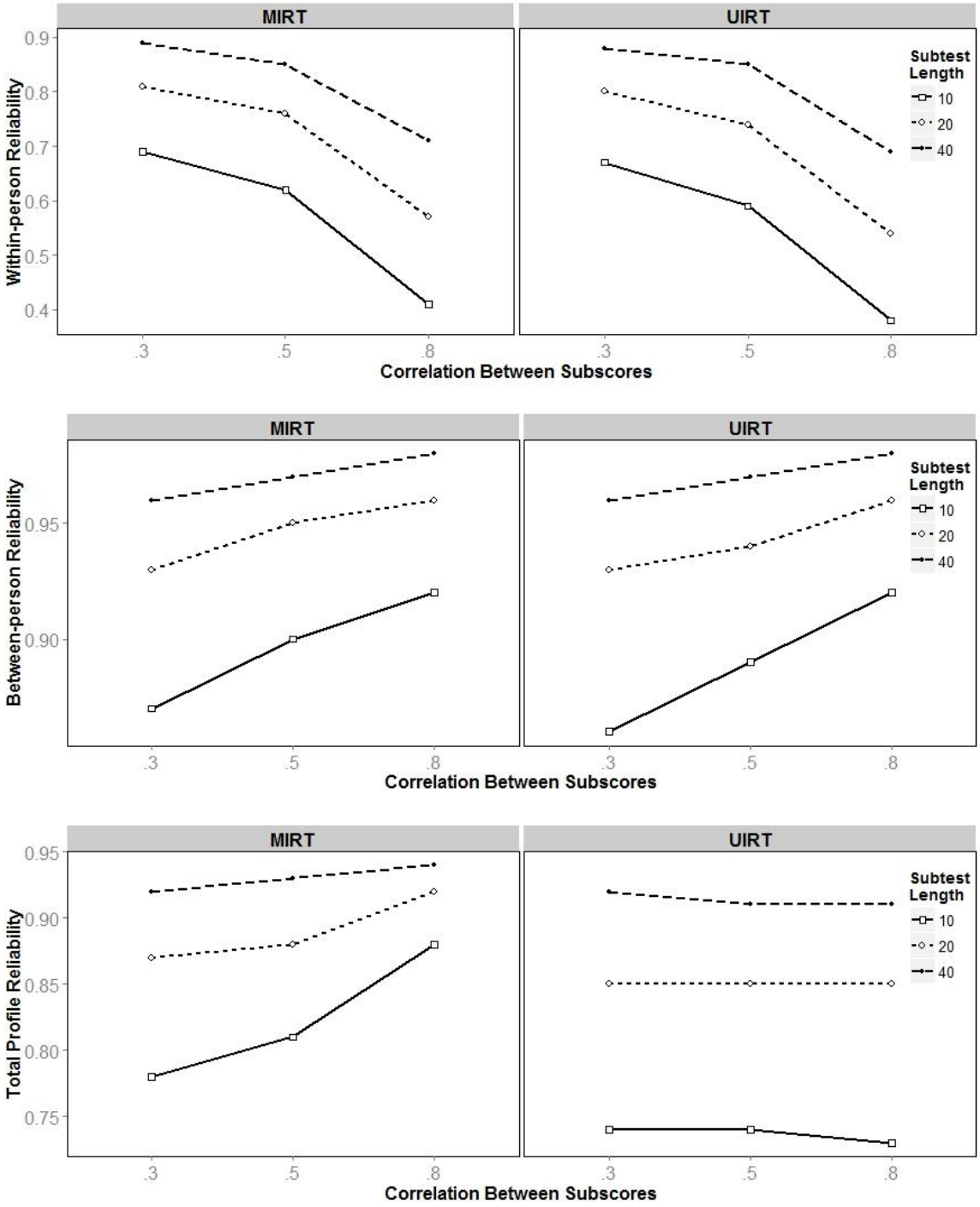

Figure 4.4. Interaction between estimated subscore reliability coefficients, correlations among subscores, and subtest length for five subtests.

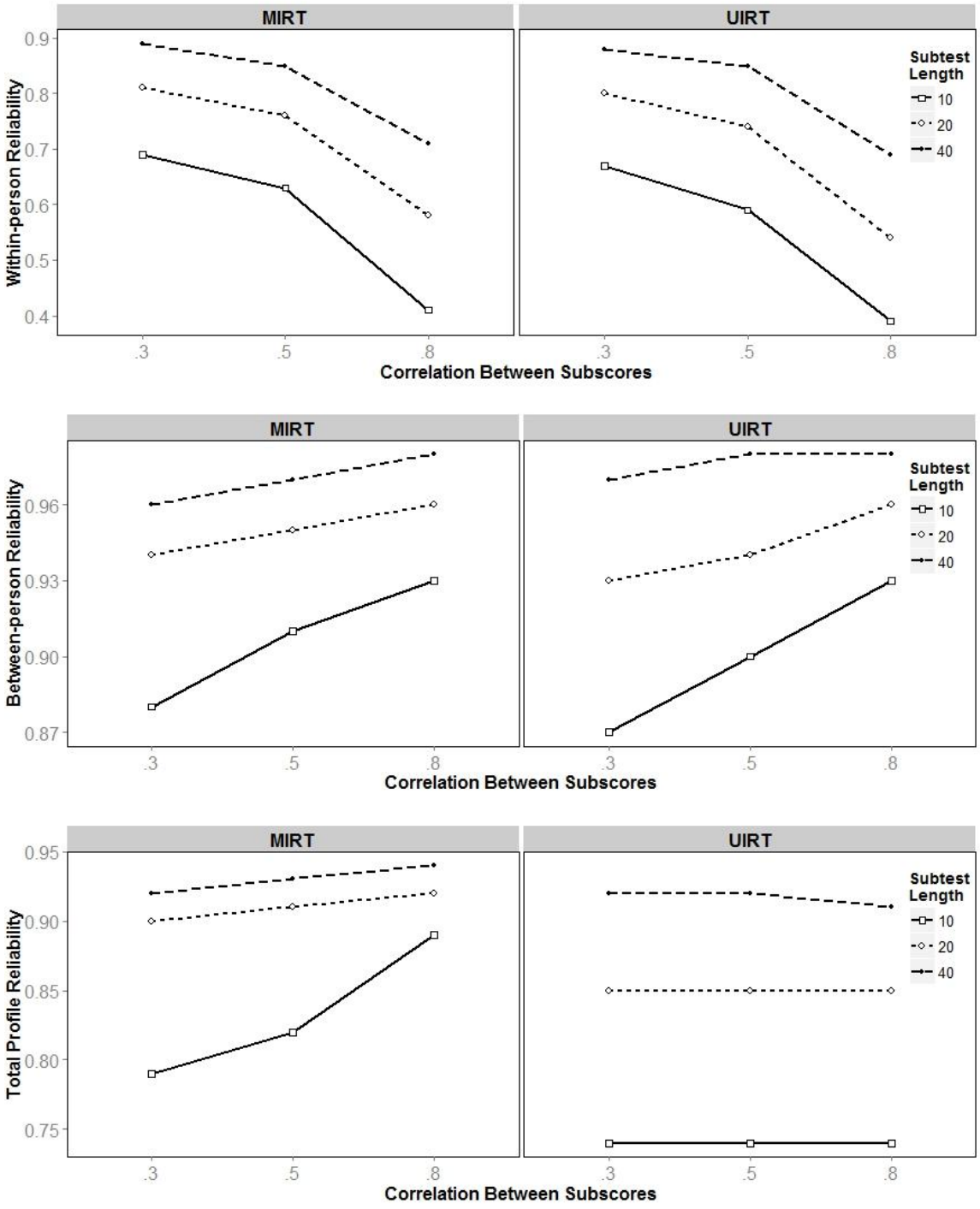

Figure 4.5. Interaction between estimated subscore reliability coefficients, correlations among subscores, and subtest length for seven subtests.

Reliability estimates from both estimation methods (i.e., MIRT and UIRT) indicated similar sampling distributions under all simulation conditions. However, the sampling distributions of reliability estimates obtained from the UIRT model were more robust against subtest length and the correlations among the subscores. The sampling distributions of reliability estimates from the MIRT model were more heavily affected by the changes in subtest length and the correlations among the subscores. As explained earlier, the number of subtests did not seem to affect reliability estimates from either the MIRT or the UIRT models. When the number of subtests (i.e., dimensions) changed from 3 to 7, the sampling distributions remained very similar.

Repeated Measures Analysis

Tables 4.9, 4.10, and 4.11 show the results from repeated measure MANOVA analyses for between-person, within-person, and total profile reliability estimates, respectively. In addition to the results of statistical significance tests, partial eta squared (η^2) are also provided as a measure of effect size for the conditions. As explained earlier, in the repeated measures MANOVA analyses, the between-subject factors were true correlations among subscores, subtest length, and number of subtests. The within-subject factor was the main effect of the subscore estimation method (i.e., MIRT vs. UIRT) and all of the two-way interactions between the estimation method and the between-subject factors. The repeated measures MANOVA analysis was repeated for between-person, within-person, and total profile reliability coefficients by using them as the dependent variable.

Table 4.9

Results of Repeated Measures Analyses for Between-Person Reliability

Factors		*SS*	*df*	*MS*	η^2
Within	Model	0.062	1	0.062	0.14
	Model x Correlation	0.015	2	0.007	0.04
	Model x Subtest Length	0.046	2	0.023	0.11
	Model x Subtest Number	0.002	2	0.001	0.01
	Error	0.182	8093		
Between	Correlation	3.537	2	1.768	0.50
	Subtest Length	20.329	2	10.164	0.85
	Subtest Number	3.621	2	1.811	0.51
	Error	3.517	8093		

Note: Model: Type of IRT model (MIRT=1, UIRT=0); SS: Sums of squares; MS: Mean square; η^2: Effect size.

Table 4.10

Results of Repeated Measures Analyses for Within-Person Total Reliability

Factors		*SS*	*df*	*MS*	η^2
Within	Model	1.313	1	1.313	0.36
	Model x Correlation	0.166	2	0.083	0.08
	Model x Subtest Length	0.120	2	0.060	0.06
	Model x Subtest Number	0.001	2	<.001	< .01
	Error	0.935	8093		
Between	Correlation	171.359	2	85.680	0.91
	Subtest Length	171.114	2	85.557	0.91
	Subtest Number	0.001	2	<.001	< .01
	Error	16.752	8093		

Note: Model: Type of IRT model (MIRT=1, UIRT=0); SS: Sums of squares; MS: Mean square; η^2: Effect size.

Table 4.11

Results of Repeated Measures Analyses for Total Reliability

Factors		*SS*	*df*	*MS*	η^2
Within	Model	8.605	1	8.605	0.64
	Model x Correlation	2.310	2	1.155	0.41
	Model x Subtest Length	2.802	2	1.401	0.44
	Model x Subtest Number	0.270	2	0.135	0.09
	Error	1.281	8093		
Between	Correlation	2.113	2	1.056	0.34
	Subtest Length	58.758	2	29.379	0.93
	Subtest Number	0.185	2	0.093	0.04
	Error	4.157	8093		

Note: Model: Type of IRT model (MIRT=1, UIRT=0); SS: Sums of squares; MS: Mean square; η^2: Effect size.

Within-subject Factors. The results indicated that the within-subject factor (i.e., type of IRT model) was an important predictor for all of the reliability coefficients, indicating that the MIRT model provides significantly higher between-person, within-person, and total profile reliability estimates than the UIRT model under all simulation conditions. The effect sizes for the estimation method were 0.14, 0.36, and 0.64 for between-item reliability, within-item reliability, and total profile reliability, respectively. The difference between the MIRT and UIRT models was higher for within-person and total profile reliabilities than between-person reliability.

All interactions between the estimation method and between-subject factors were also large in terms of sums of squares except the interaction of estimation method and number of subtests for within-person reliability, implying that the difference between the

MIRT and UIRT models depends on subtest length and correlations among subscores but not the number of subtests (i.e., dimensions). The effect sizes for the interaction of the estimation method and between-subject factors were very small when comparing between-person and within-person reliability; however, the effect sizes of the interactions between estimation method and between-subject factors were larger for total profile reliability. Figures 4.6 and Figure 4.7 show box plots of between-person, within-person, and total profile reliability estimates across different levels of correlation between subscores and subtest length. Figure 4.6 shows that the multidimensional scoring (i.e., MIRT) performed better than the unidimensional scoring (i.e., UIRT) in terms of within-person and total profile reliabilities. The difference in within-person reliability between the two methods was similar across the three levels of correlation, while the difference in total profile reliability became larger as the correlation increased. The two estimation methods performed very similarly in terms of between-person reliability when the correlations among subscores were low, medium, and high.

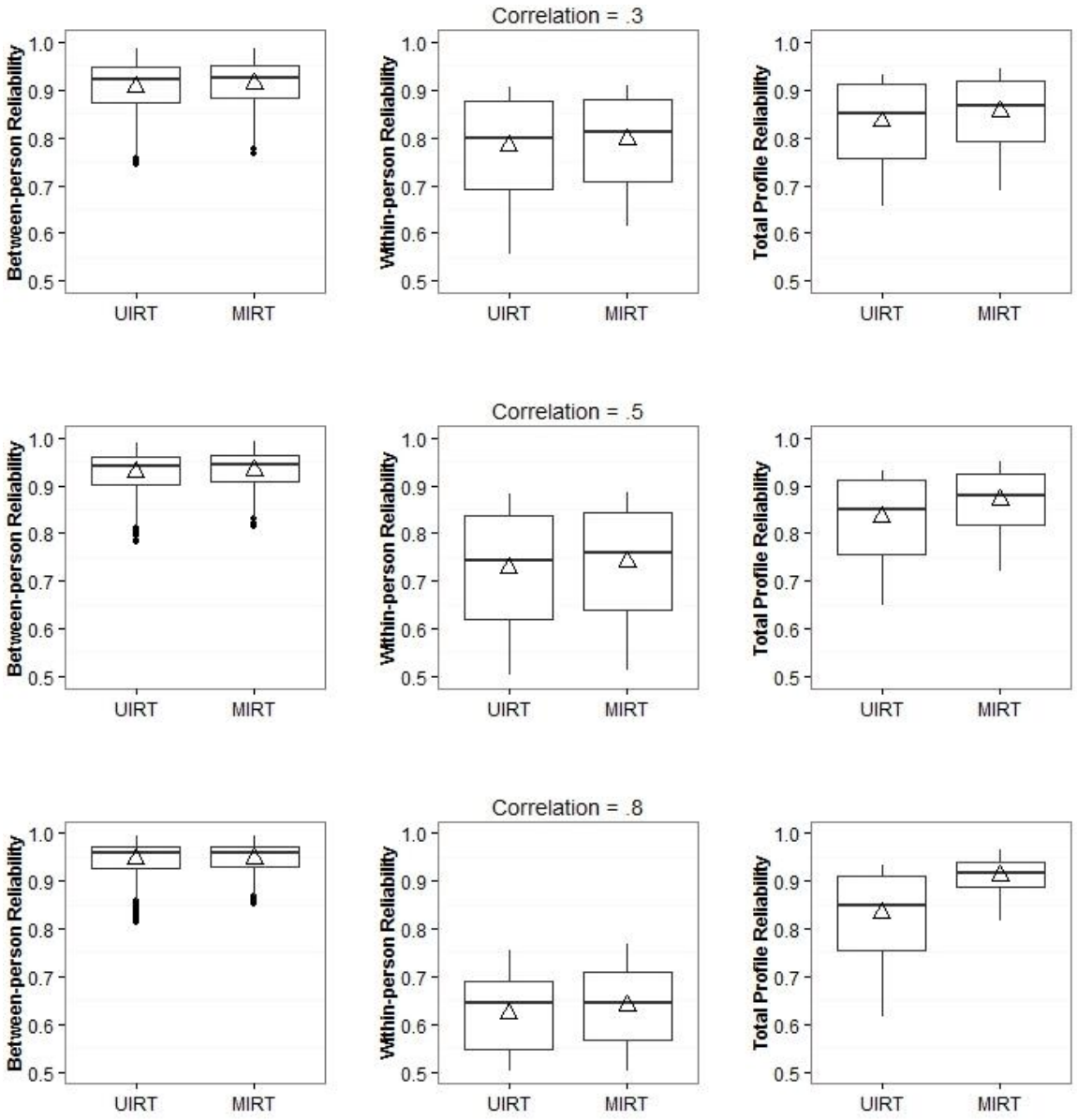

Figure 4.6. Box plots of between-person, within-person, and total profile reliability estimates from the MIRT and UIRT models across three levels of subscore correlations.

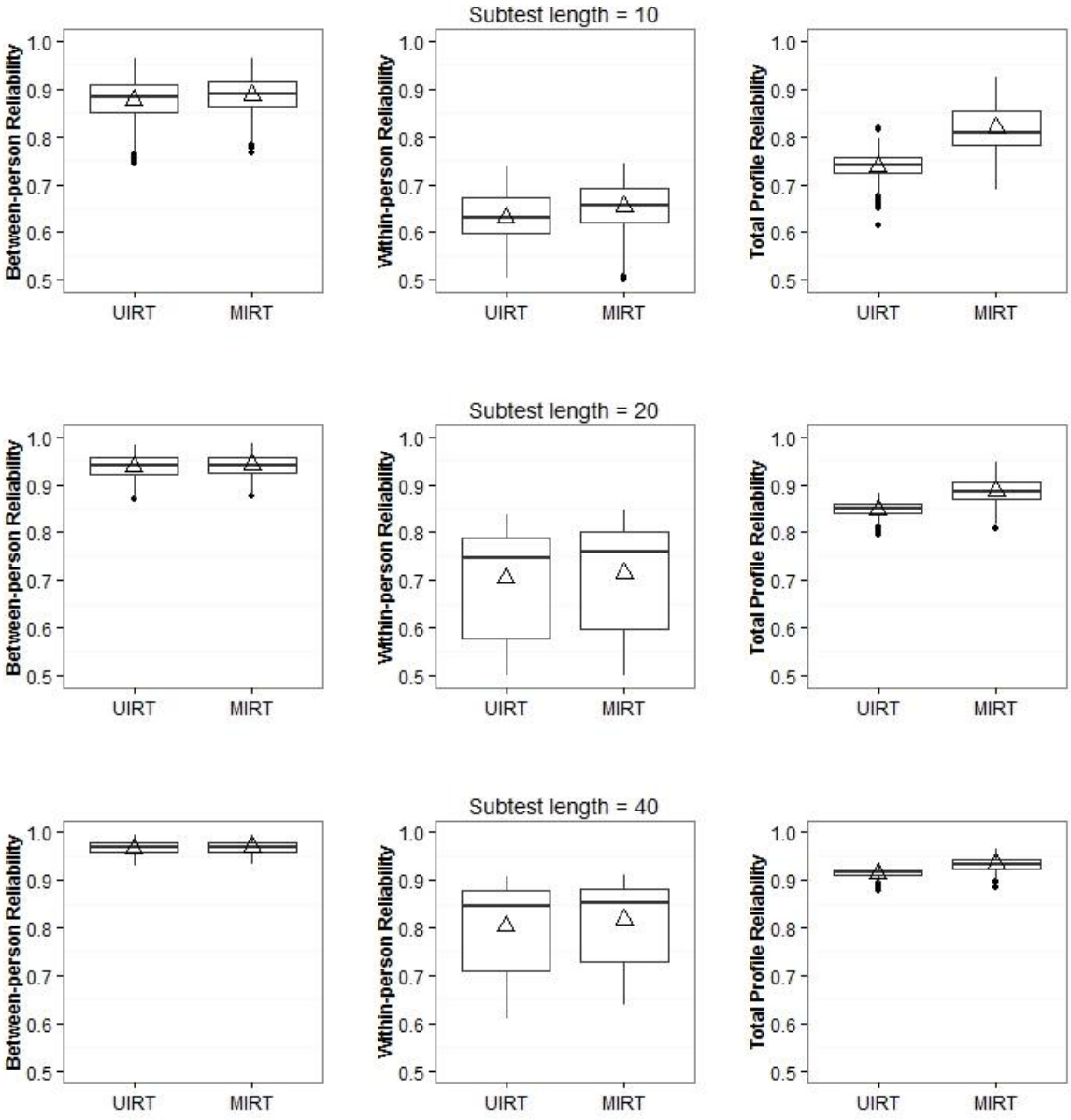

Figure 4.7. Box plots of between-person, within-person, and total profile reliability estimates from the MIRT and UIRT models across three levels of subtest length.

Figure 4.7 indicates that as the number of items in the subtests (i.e., subtest length) increased from 10 to 40, the distributions of between-person and total profile reliability estimates became narrower for both MIRT and UIRT, whereas the distributions of within-person reliability estimates became wider. As subtest length increased, the average between-person, within-person, and total reliability estimates improved for both MIRT and UIRT. Also, when subtest length was 40, the difference between the MIRT and UIRT models became negligible, suggesting that the MIRT and UIRT models' performances were not very different from each other in terms of subscore reliability when the subtest length was long.

Between-subject Factors. Results showed that for between-person, within-person, and total profile reliabilities, all of the between-subject factors (i.e., subtest length, correlations among subscores, and number of subtests) had large sums of squares except the effect of number of subtests for within-person reliability. The effect sizes of between-subject factors varied widely across the three types of reliability estimates. For between-person reliability, subtest length showed the largest effect size (η^2=.85). The effect sizes for correlations among subscores and number of subtests were equal (η^2=.55) for between-person reliability. For within-person reliability, both correlations among subscores and subtest length indicated very high effect sizes (η^2=.91), suggesting that within-person reliability heavily depends on correlations among subscores and subtest length. As subtest length increased and correlations among subscores decreased, within-person reliability improved substantially. Results also indicated that number of subtests did not have any impact on within-person reliability for both the MIRT and UIRT models.

For total profile reliability, the largest effect size for between-subject factors was subtest length (η^2=.93) followed by correlations among subscores (η^2=.34) and number of subtests (η^2=.04). Although the effect of number of subtests was large, the impact of this factor was found to be very small based on its effect size. As explained earlier, total profile reliability is a weighted combination of within-person and between-person reliability coefficients. In this study, between-person variation was larger and more dominant, and so total profile reliability estimates were closer to between-person reliability estimates than within-person reliability estimates. Therefore, the relationship between total profile reliability and the simulation conditions was similar to the relationship between between-person reliability and the simulation conditions.

Summary of the Simulation Study

The simulation study was conducted to investigate whether multidimensional and unidimensional subscore estimates differ in terms of between-person and within-person reliability under various real test situations. The two estimation methods were compared in their recovery of the relationship between subscores as well as in the subscore reliability of subscore estimates. From the simulation results summarized above, it is suggested that multidimensional and unidimensional estimation procedures perform differently in terms of subscore reliability and recovery of the relationship between subscores.

First, multidimensional subscore estimation recovers true correlations among subscores and the overall subscore structure better than unidimensional subscore estimation. MIRT tends to overestimate correlations among subscores while UIRT underestimates the same correlations. As subtest length increases, both estimation

procedures estimate correlations among subscores more accurately. Furthermore, both methods recover low correlations (ρ=.3) better than high correlations (ρ=.8). Number of subtests is not a part of subscore estimation process, and hence it does not affect the recovery of correlations among subscores. In addition to correlations among subscores, MIRT yields more similar subscores across parallel test forms than UIRT, indicating that simultaneous estimation of subscores in MIRT recovers the actual structure of subscores more consistently than separate estimation of subscores in UIRT. The effects of simulation conditions on parallel form correlations are again very similar to their effects on recovery of correlations among subscores. For MIRT, increasing subtest length and correlations among subscores improve parallel form correlations. Subtest length improve parallel form correlations of unidimensional subscore estimates as well. However, correlations among subscores have no effect on parallel form correlations. Neither MIRT nor UIRT is affected by number of subtests regarding parallel form correlations.

Second, between-person reliability estimates in the simulation study are higher than within-person reliability estimates regardless of which estimation method is used for estimating subscores. Between-person reliability improves as subtest length, correlations among subscores, and number of subtests increase. Because between-person subscore reliability is a result of between-person variation, it can be said that the higher between-person reliability, the better the test differentiates examinees. Within-person subscore reliability is heavily affected by subtest length and correlations among subscores. For short and highly correlated subtests, within-person reliability is fairly low for both estimation methods. On the contrary, for long and low correlated subtests, within-person reliability improves substantially. This finding suggests that long subtests with low

interdimensional correlations can provide more reliable subscores than short tests with highly correlated dimensions.

Finally, compared to unidimensional subscore estimates, multidimensional subscore estimates show higher within-person and between-person reliability. Results of repeated measures analysis indicate that MIRT performs better than UIRT in terms of between-person, within-person reliability, and total profile reliability. The main effect of estimation method is statistically significant for all of the three reliability indices. Estimation method (i.e., model) indicate the largest effect size in total profile reliability. Although almost all interactions between the estimation methods and the simulation conditions (i.e., subtest length, correlations among subscores, and number of subtests) are significant for all three reliability indices, the effect sizes are very small with respect to between-person and within-person reliability. The relationship between the unidimensional and multidimensional estimation methods is largely influenced by subtest length and correlations among subscores, but little by number of subtests. The interaction between estimation method and subtest length has the largest effect size among all interactions across the three reliability indices.

In conclusion, the results of the simulation study are in favor of multidimensional subscore estimation with regard to between-person and within-person subscore reliability. Although there is not a massive difference between the two estimation methods in terms of reliability estimates, the benefits of multidimensional estimation method are clearly evident. In order to examine these two estimation methods in an actual test, the same subscore and reliability estimation procedures used in the simulation study were also applied to a real dataset. The following section provides the findings from the real data

study regarding the performances of multidimensional and unidimensional subscore estimation procedures in between-person, within-person, and total profile reliability.

Results of the Real Data Study

In the real data study, unidimensional and multidimensional estimation methods were used for estimating the subscores from Quantitative 1, Quantitative 2, and Verbal subtests of EEGS. First, item parameters were estimated for the parallel test forms of each subtest. Then, unidimensional and multidimensional subscore estimates were obtained for each subtest. Lastly, between-person, within-person, and total profile reliability coefficients were estimated for EEGS.

Estimation of Item Parameters

Item parameters for the parallel test forms of Quantitative 1, Quantitative 2, and Verbal subtests of EEGS were obtained using the M3PL model in BMIRT (Yao, 2003). Table 4.12 shows the estimated item parameters for each subtest across the two parallel forms. As explained earlier, the parallel test forms for each subtest were constructed based on the item information functions obtained from the concurrent estimation of item parameters. To obtain between-person, within-person, and total profile reliabilities, subscores from the subtests were estimated using the same models in the simulation study. Unidimensional and multidimensional subscores were estimated based on the 3PL and M3PL models, respectively.

Table 4.12

Estimated Item Parameters for the Three Subtests of EEGS

Subtest	Item	Test Form 1					Test Form 2				
			β_{2j}		β_{1j}	β_{3j}		β_{2j}		β_{1j}	β_{3j}
Q1	1	1.076	0	0	0.262	0.15	1.089	0	0	-0.022	0.10
Q1	2	1.602	0	0	-0.114	0.08	1.777	0	0	-0.582	0.12
Q1	3	1.753	0	0	0.336	0.05	1.545	0	0	-0.843	0.13
Q1	4	1.618	0	0	-0.187	0.09	1.369	0	0	-0.647	0.12
Q1	5	1.161	0	0	0.387	0.07	1.463	0	0	0.272	0.05
Q1	6	2.189	0	0	-0.803	0.12	1.703	0	0	-0.250	0.07
Q1	7	0.752	0	0	0.114	0.11	1.343	0	0	0.171	0.06
Q1	8	1.564	0	0	-0.454	0.12	1.626	0	0	-0.259	0.10
Q1	9	1.857	0	0	-0.573	0.11	1.094	0	0	-0.126	0.10
Q1	10	1.147	0	0	-0.314	0.09	1.056	0	0	-0.590	0.17
Q1	11	1.596	0	0	0.088	0.10	0.788	0	0	-0.926	0.18
Q1	12	1.566	0	0	-0.239	0.08	3.473	0	0	0.704	0.11
Q1	13	1.708	0	0	0.167	0.06	3.488	0	0	0.704	0.11
Q1	14	1.931	0	0	0.217	0.09	1.654	0	0	0.758	0.10
Q1	15	1.468	0	0	0.091	0.13	0.785	0	0	-0.971	0.18
Q1	16	1.806	0	0	-0.508	0.16	1.902	0	0	-0.183	0.11
Q1	17	2.263	0	0	-0.322	0.09	1.336	0	0	0.265	0.06
Q1	18	0.948	0	0	-0.200	0.21	1.570	0	0	-0.003	0.06
Q1	19	0.856	0	0	1.153	0.16	2.069	0	0	0.043	0.05
Q1	20	1.657	0	0	0.127	0.10	1.794	0	0	0.069	0.05
Q2	21	0	1.239	0	0.320	0.06	0	1.574	0	-0.667	0.11
Q2	22	0	1.702	0	0.166	0.06	0	1.371	0	-0.665	0.11
Q2	23	0	1.793	0	0.362	0.05	0	1.239	0	-0.110	0.08
Q2	24	0	1.501	0	-0.508	0.09	0	1.376	0	-0.073	0.06
Q2	25	0	1.088	0	0.111	0.10	0	2.014	0	-0.217	0.06
Q2	26	0	1.630	0	-0.003	0.06	0	1.754	0	0.105	0.05
Q2	27	0	0.923	0	1.402	0.05	0	1.245	0	0.134	0.06
Q2	28	0	1.929	0	0.326	0.04	0	1.712	0	0.173	0.05
Q2	29	0	1.538	0	0.565	0.04	0	1.797	0	0.659	0.03
Q2	30	0	1.978	0	0.185	0.04	0	3.191	0	0.208	0.06
Q2	31	0	2.025	0	-0.104	0.05	0	2.893	0	0.298	0.04
Q2	32	0	2.305	0	-0.039	0.05	0	3.528	0	-0.020	0.06
Q2	33	0	2.074	0	0.374	0.03	0	3.009	0	0.300	0.03
Q2	34	0	2.070	0	-0.196	0.09	0	3.528	0	0.083	0.04
Q2	35	0	2.573	0	0.100	0.08	0	3.121	0	0.526	0.04
Q2	36	0	2.602	0	0.232	0.04	0	2.935	0	0.701	0.03
Q2	37	0	1.900	0	0.210	0.08	0	2.178	0	-0.288	0.12
Q2	38	0	2.016	0	0.611	0.04	0	2.147	0	-0.149	0.09
Q2	39	0	2.252	0	0.183	0.08	0	1.433	0	1.604	0.03
Q2	40	0	1.133	0	1.060	0.05	0	2.138	0	0.235	0.05

Note: Q1: Quantitative 1; Q2: Quantitative 2; V: Verbal. β_{1j}: Item difficulty; β_{2j}: Item discrimination, β_{3j}: Lower asymptote.

Table 4.12

Estimated Item Parameters for the Three Subtests of EEGS (Cont.)

Subtest	Item	Test Form 1					Test Form 2				
			β_{2j}		β_{1j}	β_{3j}		β_{2j}		β_{1j}	β_{3j}
V	41	0	0	0.425	-3.019	0.18	0	0	0.616	-2.941	0.18
V	42	0	0	0.477	-1.980	0.19	0	0	0.429	-2.287	0.18
V	43	0	0	0.461	-2.528	0.18	0	0	0.467	-2.717	0.18
V	44	0	0	0.671	-2.747	0.18	0	0	0.595	-2.559	0.19
V	45	0	0	0.686	-2.758	0.20	0	0	0.358	-1.072	0.19
V	46	0	0	0.393	-2.400	0.18	0	0	0.496	-2.329	0.17
V	47	0	0	0.475	-0.914	0.18	0	0	0.367	-2.925	0.19
V	48	0	0	0.558	-1.757	0.17	0	0	0.395	-1.984	0.18
V	49	0	0	0.625	-1.345	0.17	0	0	0.607	-2.377	0.19
V	50	0	0	0.819	-1.620	0.17	0	0	0.261	-1.251	0.17
V	51	0	0	0.639	-0.319	0.15	0	0	0.979	-2.186	0.16
V	52	0	0	0.602	-2.327	0.18	0	0	0.656	-1.092	0.15
V	53	0	0	1.187	-2.292	0.19	0	0	0.525	-1.733	0.17
V	54	0	0	0.642	-1.205	0.16	0	0	1.480	-1.741	0.15
V	55	0	0	0.417	0.139	0.18	0	0	0.758	-1.135	0.14
V	56	0	0	0.679	-0.980	0.14	0	0	0.558	-0.050	0.13
V	57	0	0	0.876	-0.314	0.12	0	0	0.803	-0.548	0.12
V	58	0	0	0.665	-1.762	0.16	0	0	1.580	-1.135	0.13
V	59	0	0	1.075	-1.604	0.15	0	0	1.627	-0.826	0.10
V	60	0	0	0.453	-0.399	0.15	0	0	0.986	-0.296	0.10
V	61	0	0	1.228	-1.003	0.13	0	0	2.363	-0.644	0.10
V	62	0	0	1.209	-0.647	0.12	0	0	1.756	-0.421	0.08
V	63	0	0	1.063	-0.158	0.09	0	0	1.790	-0.394	0.08
V	64	0	0	2.268	-0.453	0.09	0	0	2.876	-0.317	0.08
V	65	0	0	1.339	-0.415	0.09	0	0	1.387	-0.588	0.11
V	66	0	0	2.291	-0.480	0.08	0	0	1.574	-0.505	0.11
V	67	0	0	1.884	-0.309	0.08	0	0	2.141	-0.300	0.11
V	68	0	0	1.401	-0.510	0.10	0	0	0.775	0.330	0.10
V	69	0	0	2.383	-0.676	0.14	0	0	1.754	-0.100	0.08
V	70	0	0	0.928	0.325	0.08	0	0	1.281	-0.329	0.11
V	71	0	0	1.880	-0.507	0.14	0	0	2.324	-0.472	0.18
V	72	0	0	2.354	-0.253	0.09	0	0	1.967	-0.289	0.11
V	73	0	0	1.155	-0.328	0.12	0	0	1.998	-0.147	0.11
V	74	0	0	1.530	-0.147	0.11	0	0	3.423	-0.056	0.08
V	75	0	0	2.159	-0.139	0.09	0	0	3.125	-0.008	0.08
V	76	0	0	1.697	0.513	0.15	0	0	2.111	0.582	0.15
V	77	0	0	1.616	0.403	0.15	0	0	2.446	0.466	0.15
V	78	0	0	2.050	0.677	0.12	0	0	3.528	0.607	0.14
V	79	0	0	2.092	0.808	0.11	0	0	3.529	0.598	0.16
V	80	0	0	1.737	0.587	0.15	0	0	2.833	0.853	0.11

Note: Q1: Quantitative 1; Q2: Quantitative 2; V: Verbal. β_{1j}: Item difficulty; β_{2j}: Item discrimination, β_{3j}: Lower asymptote.

Subscore Estimation

Subscore estimation consisted of two steps. First, unidimensional subscore estimates were obtained for each of the subtests of EEGS separately. Then, multidimensional subscore estimates were estimated simultaneously by using correlations among unidimensional subscore estimates as population priors, as suggested by Yao (2013). As in the simulation study, unidimensional and multidimensional MAP methods were used to estimate the subscores. Figures 4.8 and 4.9 show the distributions of Quantitative 1, Quantitative 2, and Verbal subscores for the UIRT and MIRT models, respectively. The figures show that both unidimensional and multidimensional subscores are slightly negatively-skewed. The skewness is more evident in the Verbal subtest compared to Quantitative 1 and Quantitative 2. Also, the Verbal subtest has smaller variation than the other two subtests. There are a few outliers in the Quantitative 1 and 2 subtests while the number of outliers is larger in the Verbal subtest.

Table 4.13 shows the correlations among the subscores from the UIRT and MIRT models. There is a very small correlation between Quantitative 1 and Quantitative 2 in both forms although these two subtests measure the similar constructs (i.e., mathematical reasoning). The main reason of this outcome is that most examinees respond the Quantitative 1 and Verbal subtests but only the examinees from the science programs (e.g., engineering, math, etc.) tend to complete the Quantitative 2 subtest. Quantitative 2 had a higher correlation with the Verbal subtest, which also includes several items about graphical interpretation and other items requiring higher-order thinking skills. The correlations among the estimated subscores were similar across the two estimation methods.

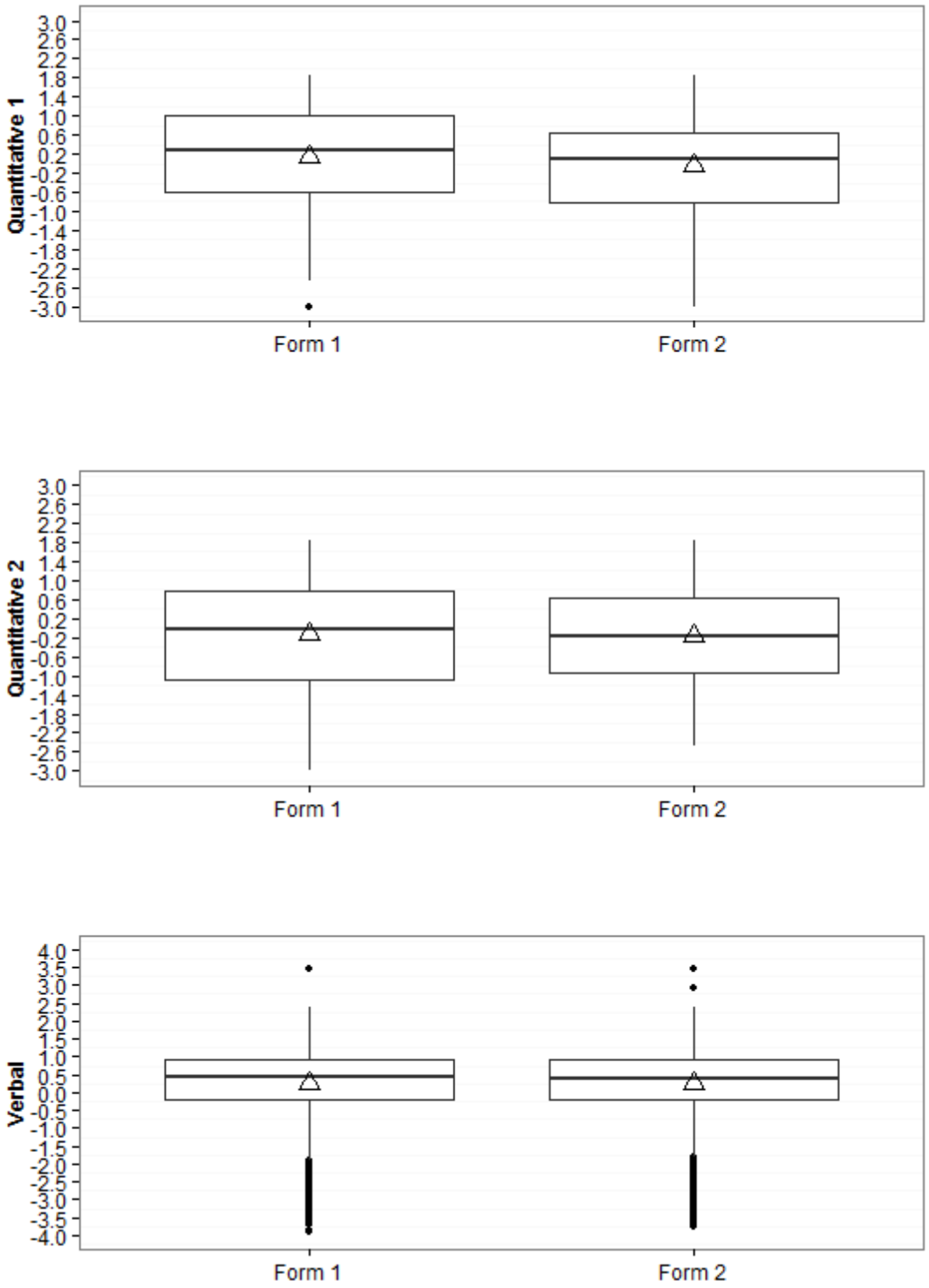

Figure 4.8. Distributions of Quantitative 1, Quantitative 2, and Verbal subscores from the UIRT model.

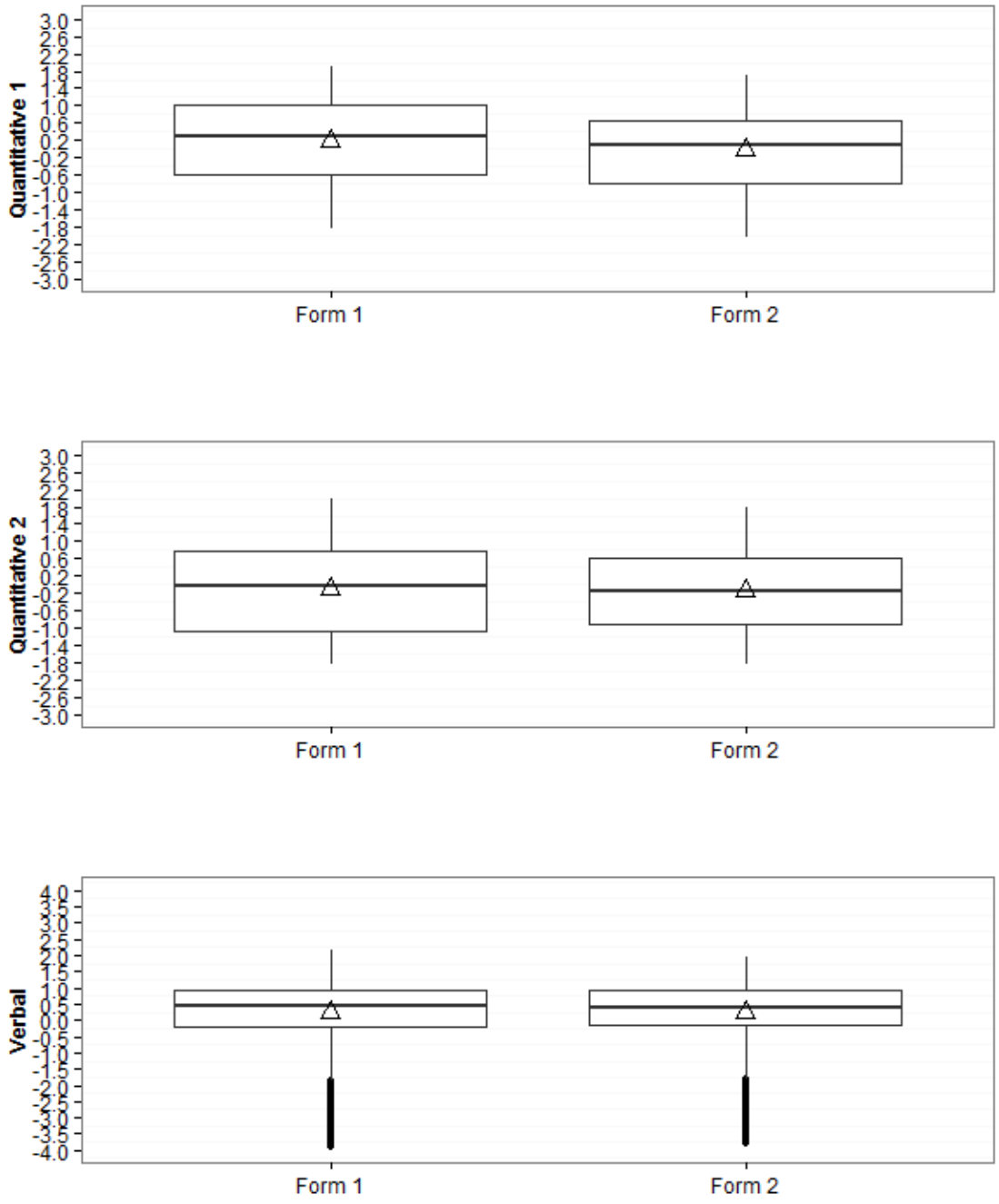

Figure 4.9. Distributions of Quantitative 1, Quantitative 2, and Verbal subscores from the MIRT model

Table 4.13

Correlation Matrices of the Estimated Subscores from Three Subtests of EEGS

Model	Parallel Forms	Subtests	Q1	Q2	V
UIRT	1	Q1	1		
		Q2	.04	1	
		V	.06	.23	1
	2	Q1	1		
		Q2	.04	1	
		V	.06	.22	1
MIRT	1	Q1	1		
		Q2	.03	1	
		V	.03	.18	1
	2	Q1	1		
		Q2	.03	1	
		V	.03	.18	1

Note: Q1: Quantitative 1; Q2: Quantitative 2; V: Verbal.

In addition to correlations among the estimated subscores, correlations among the parallel test forms were also obtained. For the unidimensional subscores, correlations between the two parallel forms of Quantitative 1, Quantitative 2, and Verbal subtests were .87, .89, and .92, respectively. Compared to the unidimensional subscore estimates, multidimensional subscore estimates indicated higher correlations between the parallel forms. The correlations for Quantitative 1, Quantitative 2, and Verbal subtests were .89, .92, and .93, respectively. As in the simulation study, MIRT performed better than UIRT when obtaining subscores from the parallel test forms. Although MIRT provided slightly better correlations among the parallel test forms than UIRT, the subscore estimates from both models indicated a strong association. Figure 4.10 shows the scatterplots of the unidimensional and multidimensional subscore estimates for each of the three subtests in EEGS. For all of the subtests, the estimates from MIRT and UIRT models were very

similar except for very high and low subscore estimates. Furthermore, the relationship between the unidimensional and multidimensional subscore estimates was very similar across the parallel test forms.

Estimating Subscore Reliability

Between-person, within-person, and total profile reliability coefficients were computed for the UIRT and MIRT models using the same approach employed in the simulation study. Results indicated that MIRT provided slightly higher reliability estimates than UIRT. Between-person, within-person, and total profile reliability estimates for the UIRT model were .88, .90, and .89, respectively. For the MIRT subscore estimates, between-person, within-person, and total profile reliabilities were .90, .92, and .91.

Both between-person and within-person reliability coefficients were fairly high for the subtests of EEGS, indicating that EEGS provides reliable subscores that can be used as measures of between-person variation as well as within-person variation. Results showed that for the estimation of subscores from EEGS, MIRT was a better method than UIRT because MIRT allowed the simultaneous estimation of subscores by using correlations among the subscore estimates as population priors. Although a simple structure was assumed for the subtests of EEGS, the MIRT procedure improved the reliability of subscore estimates by borrowing correlational information from the subtests. As indicated in the simulation study, subtest length and correlation among subscores are the main factors that affect between-person and within-person reliability. Because the subtests of EEGS were fairly long and the subscores had low correlations, subscore reliabilities were high for both unidimensional and multidimensional estimation

procedures. This finding may imply that when subtests that are long but have low correlation with each are used, both estimation procedures perform almost equally well.

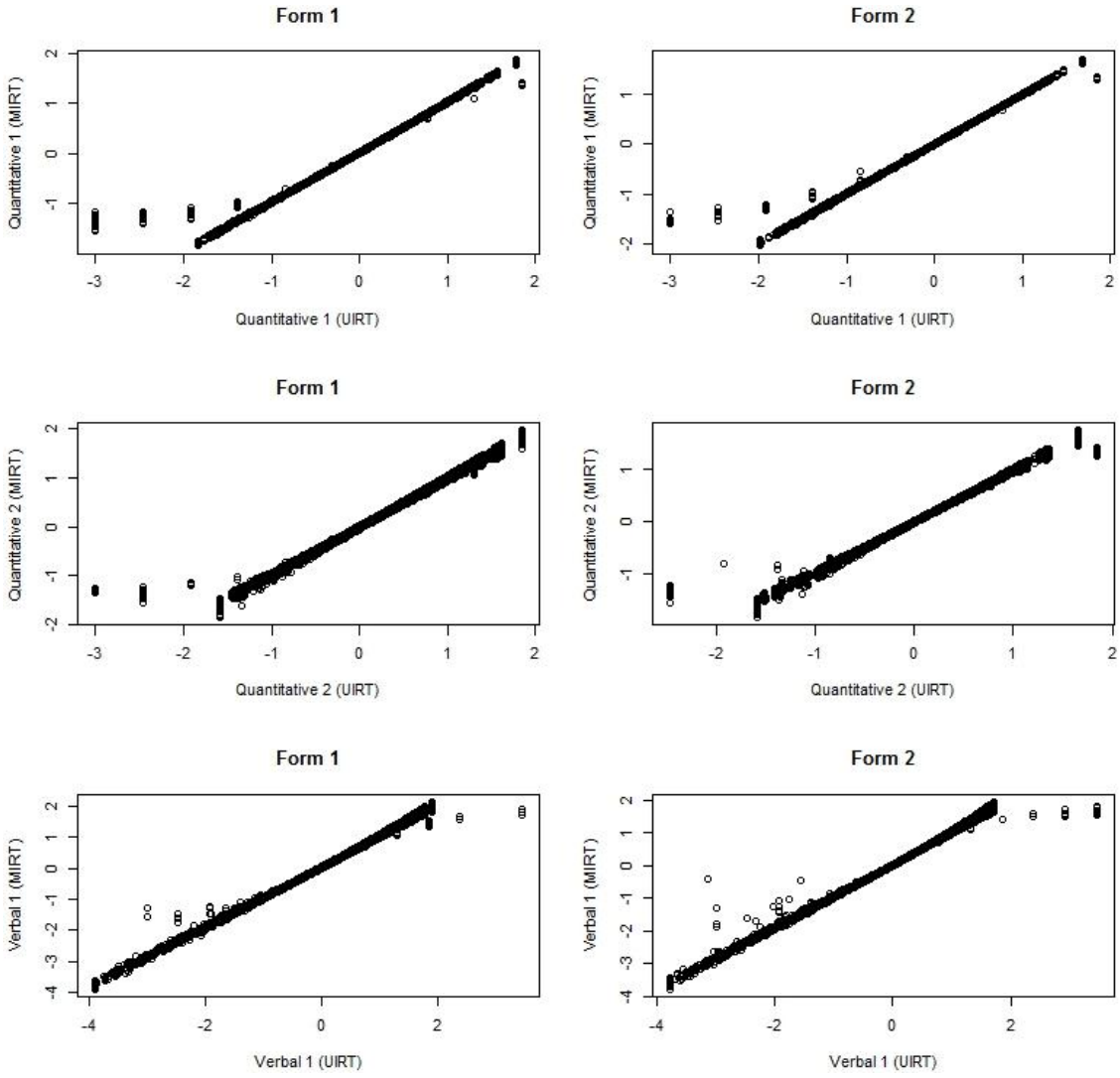

Figure 4.10. Scatterplots of unidimensional and multidimensional subscore estimates from Quantitative 1, Quantitative 2, and Verbal subtests.

CHAPTER 5

DISCUSSION and CONCLUSION

Recently, there have been several research studies about statistical procedures to improve subscore reliability. Some studies have focused on methods — such as subscore augmentation — to weight subscore estimates based on reliability indices or other subscores (e.g., Haberman, 2008; Haberman, et al., 2009; Wainer, et al., 2001; Yen, 1987). Other studies have proposed alternative estimation methods — such as Bayesian and MCMC methodologies — (de la Torre & Song, 2009; Yao & Boughton, 2007) or alternative models — such as the bi-factor model — for subscore estimation (Md Desa, 2012; DeMars, 2006). This study introduces a new profile reliability approach based on variations between examinees and within examinees. Following this approach, the simulation and real data studies were used to compare multidimensional and unidimensional subscore estimation methods in terms of subscore reliability.

The unidimensional 3PL and the M3PL IRT models were chosen for estimating subscores from simulated datasets and real data based on a multi-unidimensional structure (i.e., simple structure). In contrast with previous studies in the literature, this study employed a profile reliability approach that allows the evaluation of subscore variation across examinees as well as the variation among subscores of each single examinee. Between-person, within-person, and total profile reliability coefficients were used for the comparison of unidimensional and multidimensional subscore estimation procedures. This chapter summarizes and discusses the results of simulation and real data

studies. Each of the research questions indicated in Chapter 1 is revisited based on the findings from the simulation and real data studies. Afterwards, the implications of this study for subscore estimation and subscore reliability are discussed. Lastly, the limitations of the study and suggestions for future research are presented.

Summary of Findings

The first part of this study contains a simulation study in which unidimensional and multidimensional subscore estimation methods were applied to simulated response datasets under various simulation conditions (subtest length, correlations among subscores, number of subtests), and performances of the two methods were compared in terms of between-person and within-person subscore reliability. The simulation study addressed the first two research questions presented in Chapter 1. The second part of the study was based on an actual standardized assessment from Turkey. As in the simulation study, the difference between the unidimensional and multidimensional subscore estimation methods were demonstrated in terms of between-person and within-person subscore reliability. The real data study addressed the third research question presented in Chapter 1. The following section explains the results of these, and discusses the findings to address the research questions of this study.

Simulation Study

In the simulation study, there were two models used for the estimation of subscores from simulated data. These models were the 3PL model and the M3PL model. The M3PL model is a multidimensional version of the 3PL IRT model that can estimate item parameters and persons' abilities from multiple dimensions simultaneously. The

M3PL model is a compensatory MIRT model in which high ability in one dimension can compensate for low ability in another dimension. However, this feature is no longer in use if a test battery displays a simple structure. When each item on the test defines a single subtest or domain, the additive probability procedure in the M3PL model does not lead to a compensatory solution for estimating the probability of responding to items correctly. This type of test structure is known as a multi-unidimensional structure, where items of a subtest define a single domain but the domains may still be related to each other. The M3PL model benefits from associations (i.e., correlations) among subtests or domains by using these pieces of information as population priors in the estimation process. Previous studies indicated that this feature of multidimensional subscore estimation can help to improve the reliability of subscore estimates; therefore, MIRT is a better way of estimating subscores than UIRT (de la Torre & Hong, 2010; Haberman & Sinharay, 2010; Wang, Cheng, & Chen, 2004; Yao & Boughton 2009). In light of the findings of previous studies, the simulation study investigated the subscore reliability of unidimensional and multidimensional subscore estimates using Davison et al.'s (2012) profile reliability approach.

Evaluation of subscore estimates. Before the comparison of the MIRT and UIRT models in terms of subscore reliability, these models were evaluated based on two criteria: (a) recovery of true correlations among subscores, and (b) correlations between parallel test forms. Accurate recovery of true correlations among subscores was essential because true correlations were used as population priors in the subscore estimation procedure. Correlations between parallel test forms are also important because the profile

reliability approach employed in this study is based on the assumption that the correlation of two parallel test forms yields true score variance.

Results showed that the MIRT model tends to overestimate true correlations among subscores, whereas the UIRT model mostly underestimates the same correlations. For both models, discrepancies between true correlations and estimated correlations substantially increased as true correlations among subscores became higher. In addition to true correlations among subscores, subtest length was another factor that had an impact on the recovery of true correlations. As subtest length increased from 10 items to 40 items, both models provided better results in terms of recovery of true subscore correlations. For both MIRT and UIRT, the most accurate results were obtained when subtest length was 40 and true correlations among subscores were .3. The number of subtests affected neither the MIRT model nor the UIRT model in terms of recovery of true correlations among subscores.

The second criterion for the evaluation of subscore estimates was the magnitude of correlations between parallel test forms. As explained earlier, for each crossed condition, two response datasets were simulated using the same item parameters and subscores. The evaluation of correlations between the parallel forms was particularly important because the profile reliability approach employed in this study obtains true within-person and between-person variations based on covariances of the parallel test forms. Results indicated that parallel form correlations from the MIRT model were higher than parallel form correlations from the UIRT model under all simulation conditions. This finding suggests that the simultaneous estimation of subscores with MIRT helps to reduce the discrepancies between parallel test scores because it applies the

same population priors to both test forms. Multidimensional subscore estimates from the parallel test forms become more similar to each other whereas unidimensional subscore estimates from the parallel test forms tend to differ more from each other due to estimation errors emerging from separate estimations of subscores. As subtest length and correlations among subscores increased, parallel form correlations of multidimensional subscore estimates substantially improved. For unidimensional subscore estimates, parallel form correlations increased only when subtest length increased. Correlations among subscores within each test form did not affect parallel form correlations. For both estimation methods, the number of subtests did not have any impact on parallel form correlations.

Research Question 1. The first research question focused on the comparison of unidimensional and multidimensional IRT models in terms of within-person and between-person subscore reliability. The first research question is as follows: Does the MIRT model perform better than the UIRT model in terms of within-person and between-person subscore reliability?

As described in Chapters 2 and 3, MIRT allows the simultaneous estimation of multiple ability dimensions (i.e., subscores) whereas unidimensional item response theory (UIRT) modeling treats each subtest as a standalone test and ignores the relationship between the subtests. Another major difference between the MIRT and UIRT approaches is that MIRT takes prior information (e.g. correlations among subscores) into account when estimating subscores to improve the precision of subscore estimates, while UIRT does not allow any prior information unless an augmentation procedure is applied to subscore estimates. MIRT adjusts likelihood functions for ability estimation depending

on pre-specified population priors (i.e., mean, variance, and correlations) and allows dimensions to borrow information from each other, while UIRT deals with a single dimension or subtest each time. Through this feature, MIRT seems to provide more accurate estimates of correlations among the subscores, and it also minimizes differences between parallel test forms due to estimation errors.

Results of the simulation study suggest that multidimensional estimation generally provides more reliable subscore estimates than unidimensional estimation. For both estimation methods, within-person reliability is smaller than between-person reliability no matter which estimation method was used in the simulations. Also, total profile reliability estimates are mostly closer to between-person reliability estimates. Compared to between-person reliability, the utility of the multidimensional estimation method seems more evident in within-person reliability based on effect sizes of the estimation method in repeated measures analyses. The use of the multidimensional estimation method improves within-person subscore reliability more than between-person subscore reliability. This finding suggests that when subscores are intended to be used for making inferences regarding variation among a person's subscores (i.e., within-person variation), multidimensional subscore estimation can be more useful than unidimensional subscore estimation.

Interactions between the estimation methods and simulation conditions differ across the three reliability coefficients. Tests of interaction effects in repeated measures analyses indicate that for between-person reliability, the impact of the estimation method highly depends on the level of subtest length. The longer the subtest length, the greater the difference between the multidimensional estimation and the unidimensional

estimation. In contrast to between-person reliability, the impact of estimation method mostly depends on correlations among subscores for within-person reliability. The more correlated the subscore estimates, the less reliable subscores become in terms of within-person reliability. The number of subtests has a weak interaction with between-person reliability and it has no effect at all for within-person reliability.

Research Question 2. The second research question involved the effects of simulation conditions on between-person, within-person, and total profile reliability of subscore estimates. Simulation conditions were subtest length, correlations among subscores, and number of subtests. As discussed earlier, subtest length and correlations among subscores had an impact on the recovery of true correlations among subscores and parallel form correlations, while the number of subtests did not have an impact on the results. The second research question specifically focused on the impact of the simulation conditions on subscore reliability estimates. The second research question is as follows: How are within-person and between-person subscore reliabilities from the UIRT and MIRT models affected by varying data conditions (test length, number of subtests, and correlations between subtests)?

Subtest length has a large impact on both between-person and within-person reliability estimates. It has a very large effect as a main effect and a relatively small effect as an interaction with the estimation method. In the simulation study, three levels (10, 20, or 40 items) of subtest length are considered. As subtest length increases, both between-person and within-person reliability improve substantially. The relationship among estimation methods seems to differ depending on subtest length for both within-person and between-person reliability. Regardless of subtest length, multidimensional subscore

estimation performs slightly better than unidimensional subscore estimation with respect to between-person and within-person reliability. Multidimensional estimation seems more robust than unidimensional estimation against changes in subtest length. For both methods, when subtest length is short, between-person subscore reliability is slightly greater than within-person subscore reliability. This finding implies that short subtests may not be appropriate for evaluating the variation among each examinee's subscores in a test because the estimated subscores may not be reliable indicators of within-person variation. Rather, the use of short subtests seems more appropriate when one intends to evaluate the overall variation between the examinees. In the case of multiple short subtests, multidimensional estimation seems more advantageous over unidimensional subscore estimation.

Correlations between subscores are another factor that largely affects between-person and within-person subscore reliability. In the simulation study, three levels (.3, .5, or .8) of correlations representing low, medium, and high correlations were considered. Correlations among subscores have a positive relationship with between-person reliability, while they are negatively associated with within-person reliability. This is because highly correlated subscores lead to smaller variation between subscores but larger variation among the examinees. Consequently, high correlations among subscores increase between-person reliability whereas they substantially reduce within-person reliability. Correlations among subscores affect multidimensional and unidimensional subscore estimations in the same way. Repeated measures analyses indicate that the relationship among estimation methods with respect to subscore reliability is largely influenced by correlations among subscores. The main effect of correlations has a large

effect size while the effect size for the interaction of correlations with the estimation method is relatively small. Both the main effect and the interaction displayed larger effect sizes for within-person reliability than between-person reliability.

The number of subtests seems to influence between-person reliability but it has no effect on within-person reliability. As the number of subtests increases from three to seven, within-person reliability remains almost constant for both the unidimensional and multidimensional estimation methods. Repeated measures analyses also show that neither the main effect of the number of subtests nor its interaction with the estimation method is statistically significant. However, for between-person reliability, the number of subtests seems to have a positive impact although this effect is still very small. With more subtests, between-person reliability seems to increase while within-person reliability remains unaffected. This is because all subscores are assumed to have the same relationship with each other in the simulation study; therefore adding more subtests into the test does not affect the total within-person variation. Rather, it increases the average variation between the examinees, and so between-person reliability also increases. If the subscores from a test have different correlations with each other, the number of subtests may influence within-person reliability as well.

Real Data Study

The real data study differs from the simulation study in terms of some data characteristics. First, the sample size in the real data study is larger (N=10000) than the sample size (N=1500) in the simulation study. Second, the subtest length in the real data study is different across the subtests (20 items for Quantitative 1 and 2, 40 items for Verbal) whereas the subtest length is fixed (10, 20, or 40 items) across the subtests in the

simulation study. Lastly, correlations among the subscores also differ across the subtests in the real data study while it is fixed (.3, .5, or .8) across the subtests in the simulation study. Through these differences, the real data study provides a good example of test characteristics that are more likely to occur in real testing programs.

Research Question 3. The third research question involved the comparison of multidimensional and unidimensional subscore estimates from EEGS in terms of within-person and between-person subscore reliability. The third research question is as follows: How do the MIRT and the UIRT models perform in terms of within-person and between-person subscore reliability in real data?

Results of the real data study resemble the findings from the simulation study. In EEGS, the lengths of the three subtests are fairly long. Also, correlations between Quantitative 1 and other subtests (i.e., Quantitative 2 and Verbal) are close to zero, and Verbal and Quantitative 2 subtests have a small correlation. Under similar conditions, the simulation study suggests that both unidimensional and multidimensional subscore estimates should have high between-person and within-person reliability. Findings from the real data study show that between-person and within-person reliability for the three subtests of EEGS are fairly high. Within-person subscore reliability is greater than between-person subscore reliability, and the estimate of total profile reliability is between between-person and within-person reliability estimates. According to between-person and within-person subscore estimates, subscores from EEGS seem highly reliable for the evaluation of variation between the examinees as well as variation within each examinee's subscores.

In EEGS, multidimensional subscore estimation provides slightly higher subscore reliability than unidimensional subscore estimation. It should be noted that multidimensional estimation uses means, variances, and interdimensional correlations of unidimensional subscores as population priors. Although correlations among the subscores are quite small, using this information along with means and variances in the multidimensional estimation procedure seems to improve both between-person and within-person reliability. This finding implies that, for EEGS, multidimensional estimation can benefit from unidimensional subscore estimates to determine population priors, and this may help to improve subscore reliability.

Conclusions

The use of subscores in educational and psychological assessments is important. Subscores yield valuable diagnostic information that can be used for the evaluation of examinees' strengths and weaknesses in different domains as well as providing feedback or planning future remedial studies. However, these benefits of subscores do not necessarily mean that subscores obtained from tests are always useful. Considering that subscores are sometimes estimated from short subtests, the utility and psychometric quality of subscores may be highly questionable. One of the most common criticisms regarding reporting and using subscores is subscore reliability. Reliability is particularly a significant issue for subscores because subtests usually tend to be shorter than typical tests and subscores tend to be more highly correlated since they are components of a larger test or battery of tests. If subscores are not reliable indicators of the construct being measured, then the information they yield is not trustworthy, and any decisions based on those subscores are likely to be misleading. Therefore, it is vitally important to

investigate the reliability of subscores before using them for any purpose or reporting them to examinees.

In this study, a profile reliability approach is introduced for the evaluation of subscore reliability between persons and within persons. This approach is applied to subscores from multidimensional and unidimensional IRT models and the performances of these models are compared with simulated and real data sets. The findings of this study suggest that simultaneous estimation of subscores in MIRT benefits from the correlational information among subscores and improves both between-person and within-person subscore reliability. Unidimensional estimation in UIRT ignores the relationships among subtests, and so it fails to use this additional information to improve subscore reliability. Increasing subtest length helps to improve both between-person and within-person reliability. Higher correlations among subscores increase between-person reliability while they substantially reduce within-person reliability. Although both models underperform when subtest lengths are short and subscores are highly correlated with each other, the performance of MIRT still seems to be better than UIRT in terms of subscore reliability. Another implication of this study is that subscores with low within-person but high between-person reliability should be used for between-person comparisons rather than within-person comparisons. When subscores of a test indicate high within-person reliability, subscores can be used for interpreting examinees' strength and weaknesses in the content domains measured in the test.

The findings of this study have important implications in terms of test design. First, the use of profile scores or subscores should be determined depending on the purpose of the test. If the purpose of a test is to evaluate examinees' strength and

weaknesses in multiple domains or strands, then profile scores or subscores should be distinct enough to make inferences regarding examinees' performances in each domain. Findings of this study indicate that within-person reliability can be a good indicator of the extent which subscores provide distinct information. As correlations among subscores increase and subscores become more similar to each other, within-person reliability is getting smaller, suggesting that highly correlated subscores may not be reliable enough to make inferences or decisions about examinees' performance in each domain. However, if the purpose of the test is to compare examinees' overall test performance rather than evaluating strengths and weaknesses in each domain, high correlations among subscores do not constitute a treat in terms of subscore reliability. Highly correlated subscores imply that the subtests measure a similar construct. Therefore, a composite test score base on the multiple subscores can be more useful and reliable than using individual subscores in terms of evaluating examinees' performance.

Second, the decision of which estimation method to use should be made based on the design and purpose of the test rather than the statistical properties of the estimation methods. It should be noted that the use of MIRT over UIRT models cannot solve the issues in subscores due to test design but it may help to reduce the negative effects of test design. For instance, if a test was assumed to measure multiple distinct dimensions but the estimated subscores were unexpectedly highly correlated, then within-person reliability of subscores would be low regardless of what estimation method was used. The use of MIRT can help to improve within-person reliability of subscores by using the information among subscores. However, the degree of precise information that subscores provide would still remain questionable. Therefore, the selection of estimation method

should be done in conjunction with the relationship among subscores and the purpose of the test.

Limitations of the Study and Future Research

There are several limitations of this study. First, the simulation study in this study used true item parameters to eliminate additional errors in subscore estimation due to item parameter estimation. However, real testing programs require both item parameters and abilities to be estimated. Therefore, in addition to subtest length, correlations among subscores, and number of subtests, other factors that are likely to affect item parameter estimation (e.g., sample size, distribution of item parameters) should be taken into consideration before evaluating subscore reliability. Further studies are needed to understand the joint effects of item parameter and subscore estimation procedures on subscore reliability.

Second, subscore estimates in this study were obtained using the multidimensional MAP approach in BMIRT (Yao, 2003). The MAP method is usually more feasible than the MLE method because it allows ability estimation from all response patterns including zero or perfect scores. However, the use of the MAP method could be disadvantageous when prior mean and variance are not correctly specified. Also, ability estimates from MAP are heavily regressed towards the prior mean assumed for the ability distribution (Bock & Mislevy, 1982; Mislevy & Bock, 1997). Therefore, MAP may yield higher estimation errors than other estimation methods such as EAP and MLE. To minimize these problems with MAP, population priors should be carefully chosen, and the effects of different ability distributions on the subscore estimation should be

examined. Despite its computational complexity, MCMC could also be considered as an alternative method for subscore estimation.

Third, the simulation conditions used in this study included subtest length, correlations among subscores, and number of subtests. To facilitate the interpretation of results, the same subtest length and inter-correlations were used across subtests. However, subtest length and especially correlations among subscores may not be fixed across subtests in real testing applications. Further study is needed to examine subscore reliability when the subtests differ in length and inter-correlations.

Lastly, in order to make a direct comparison of the MIRT and UIRT models, this study considered a simple test structure in which each item measures only one dimension. However, MIRT can also estimate subscores from a non-simple test structure (i.e., complex test structure) in which items measure multiple abilities. Future studies can consider the evaluation of subscore reliability in MIRT models based on a non-simple test structure.

References

Ackerman, T.A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67–91.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education, 20*, 309–310.

Ackerman, T.A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement,* 20, 4, 311–330.

Ackerman, T. A., & Evans, J. A. (1994). The influence of conditioning scores in performing DIF analysis. *Applied Psychological Measurement,* 18, 257–275.

Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing*. Washington DC: Author.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, 9, 37–48.

Andrich, D.A. 1978. A rating formulation for ordered response categories. *Psychometrika, 43,* 561–73.

Arce-Ferrer, A. J. (2010). *Derivation of a profile reliability index for an individual: A multi-factor congeneric approach with Guttmann error type structures*. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika, 72*(2), 141–157.

Bechger, T. M., Maris, G., B´eguin, A., & Verstralen, H. H. F. M. (2003). *Combining classical test theory and item response theory*. R&D Report 2003-4, Cito, Arnhem, The Netherlands.

Beguin, A., & Glas, C. A. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*(4), 541–561.

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431–444.

Boughton, K. A., Yao, L., & Lewis, D. M. (2006). *Reporting diagnostic subscale scores for tests composed of complex structure*. Paper presented at the annual meeting of National Council on Measurement in Education, San Francisco, CA.

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

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full information item factor analysis. *Applied Psychological Measurement, 12*, 261–280.

Bock, R. D., & Liberman, M. (1970). Fitting a response model for n dichotomously scored items. *Psychometrika*, *35*, 283–319.

Bock, R. D., Thissen, D., & Zimowski, M. F. (1997). IRT estimation of domain scores. *Journal of Educational Measurement, 34*(3), 197–211.

Brennan, R. L. (2005). *Some test theory for the reliability of individual profiles*. (Research Report 12). Center for Advanced Studies in Measurement and Assessment.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In Bollen, K.A., Long, J.S. (Eds.), *Testing Structural Equation Models* (pp. 136-162). Sage, Newbury Park, CA.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika, 75*, 33–57.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). *IRTPRO for Windows* [Software]. Lincolnwood, IL: Scientific Software International.

Carlson, J. E. (1987). *Multidimensional item response theory estimation: A computer program* (Research Report ONR 87-2). Iowa City, IA: The American College Testing Program.

Chen, P.-H. (2006). The influences of the estimation methods on the precision of ability estimation in multidimensional computerized adaptive testing. *Educational Psychology Bulletin. 38*(2), 193–210.

Conger, A. J., & Lipshitz, R. (1973). Measures of reliability for profiles and test batteries. *Psychometrika 38*(3), 411–427.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Davison, M. (1996). *Multidimensional scaling interest and aptitude profiles: Idiographic dimensions, nomothetic factors*. Presidential address to Division 5, American Psychological Association, Toronto.

Davison, M. L., Chang, Y., & Davenport, E. C., Jr. (2012). Assessing the Reliability and Convergent Validity of Individual Differences in Profile Patterns. *Unpublished manuscript.*

Davison, M. L., Kim, S-K., & Close, C. W. (2009). Factor analytic modeling of within person variation in score profiles. *Multivariate Behavioral Research*, *44*, 668–687.

de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. *Applied Psychological Measurement*, *32*, 355–370.

de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement, 33,* 465–485.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT approach. *Applied Psychological Measurement*, *34*, 267–285.

de la Torre, J., & Patz, R.J. (2005). Making the most of what we have: A practical application of multidimensional IRT in test scoring. *Journal of Educational and Behavioral Statistics*, *30*, 295–311.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement, 33*, 620–639.

de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement, 35*(4), 296–316.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43* (2), 145–168.

Dimitrov, D. M. (2003). Marginal true-score measures and reliability for binary items as a function of their IRT parameters. *Applied Psychological Measurement, 27*, 440–458.

Douglas, J., Roussos, L. A., & Stout, W. F. (1996). Item bundle DIF hypothesis testing: Identifying suspect bundles and assessing their DIF. *Journal of Educational Measurement, 33*, 465–485.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement,* 7, 189–199.

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006). *A comparison of subscale score augmentation methods using empirical data*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Edwards, M. C. & Vevea, J. L. (2006). An empirical Bayes approach to subscore augmentation: How much strength can we borrow? *Journal of Educational and Behavioral Statistics, 31*(3), 241–259.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (third ed., pp. 105–146). New York: Macmillan.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis-based models. *Applied Psychological Measurement, 34*(1), 10–26.

Fraser, C. (1988). *NOHARM II: A Fortran program for fitting unidimensional and multidimensional normal ogive models of latent trait theory* [Software]. Armidale, New South Wales: University of New England, Centre for Behavioral Studies.

Gibbons, R. D., & Hedeker, D. R. (1992). Full information bifactor analysis. *Psychometrika, 57,* 423–436.

Goodman, D. & Hambleton, K. (2004). Student test score reports and interpretive guides: Review of current practices and suggestions for future research. *Applied Measurement in Education, 17,* 145–220.

Haberman, S.J. (2008).When can subscores have value? *Journal of Educational and Behavioral Statistics, 33,* 204–229.

Haberman, S.J., & Sinharay, S. (2010). Reporting subscores using multidimensional item response theory. *Psychometrika, 75*(2), 209–227.

Haberman, S. J., Sinharay, S., & Puhan, G. (2006). Subscores for institutions (ETS Research Rep. No. RR-06-13). Princeton, NJ: ETS.

Haberman, S., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology, 62*(1), 79–95.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation & The Health Professions, 27*(4), 349–368.

Hambleton, R.K. (2000). *Introduction to Item Response Theory*. Breakout session presented at the Sylvan Prometric Results Conference, Tucson AZ.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Thousand Oaks: Sage Publications. Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, *9*(2), 139–164.

Han, K. T. (2012). Fixing the c parameter in the three-parameter logistic model. *Practical Assessment, Research & Evaluation, 17*(1), 1–24. Available from http://pareonline.net/pdf/v17n1.pdf

Hartig, J., & Hohler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35,* 57–63.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: applications to true-score prediction from a possible nonparallel test. *Psychometrika, 68*(1), 123–149.

Hu, L.T., & Bentler, P. M. (1999). Cut off criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Kelly, T. L. (1927). *Interpretation of educational measurements*. New York, NY: World Book Company.

Kelly, T. L. (1947). *Fundamentals of statistics*. Cambridge: Harvard University Press.

Kim, S., Feldt, L. S. (2010). The estimation of the IRT reliability coefficient and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review 11,* 179–188.

Kline, R. B. (2005). *Principle and Practice of Structural Equation Modelling*. The Guilford Press, New York.

Kolen, M. J. & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.

Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(8), 151–160.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*(2), 187–212.

Masters, G.N. 1982. A Rasch model for partial credit scoring. *Psychometrika, 47,* 149–74.

May, K., & Nicewander, W. A. (1994). Reliability and information functions for percentile ranks. *Journal of Educational Measurement, 31,* 313–325.

McDonald, R. P. (1985). Unidimensional and multidimensional models for item response theory. In D. J. Weiss (Ed.), *Proceedings of the 1982 Computerized Adaptive Testing Conference* (pp. 127–148). Minneapolis: University of Minnesota, Department of Psychology, Psychometrics Methods Program.

McDonald, R. P. (1997). *Normal-ogive multidimensional model*. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 257-269). New York: Springer.

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

Md Desa, Z. N. D. (2012). Bi-factor multidimensional item response theory modeling for subscore estimation, reliability, and classification. (Doctoral dissertation). Retrieved from http://kuscholarworks.ku.edu/dspace/handle/1808/10126.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29,* 133–161.

Mislevy, R. J., & Bock, R. D. (1997). BILOG version 3.11. [Software]. Chicago, IL: Scientific Software International.

Monaghan, W. (2006). The facts about subscores. *R&D Connections*. Princeton, NJ: Educational Testing Service. Available from http://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

Mulaik, S. A. (1972). *A mathematical investigation of some multidimensional Rasch models for psychological tests*. Paper presented at the annual meeting of the Psychometric Society.

Muraki, E. (1999). *POLYFACT Version 2* [Software]. Princeton, NJ: Educational Testing Service.

Muthén, L. K., & Muthén, B. O. (1998-2011). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén & Muthén.

Nandakumar, R. (1994). Assessing dimensionality of a set of item responses: Comparison of different approaches. *Journal of Educational Measurement, 31*(1), 17–35.

Osteeen, P. (2010). An introduction to using multidimensional item response theory to assess latent factor structures. *Journal of the Society for Social Work and Research, 1*(2), 66-82.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2001). *Practical considerations in computer testing*. New York: Springer-Verlag.

R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedogogische Institute.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230.

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

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement, 21*, 25–36.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York: Springer.

Reckase, M. & Hirsh, T. (1991). *Interpretation of number correct scores when the true number of dimensions assessed by a test is greater than two*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Reckase, M. D., & McKinley, R. L. (1982). *The feasibility of a multidimensional latent trait model*. Paper presented at the annual meeting of the American Psychological Association, Washington.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*, 229–244.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika, 70*, 533–555.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, 61, 331–354.

Sheng, Y. (2005). Bayesian analysis of hierarchical IRT models: Comparing and combining the unidimensional & multi-unidimensional IRT models (Doctoral dissertation). Retrieved from https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/4153/research.pdf?sequence=3

Sheng, Y., & Wikle, C. K. (2007). Comparing Multiunidimensional and unidimensional item response theory models. *Educational and Psychological Measurement, 67(*6) 899–919.

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement, 68,* 413–430.

Shojima, K., & Toyoda, H. (2002). Estimation of Cronbach's alpha coefficient in the context of item response theory. *The Japanese Journal of Psychology, 73,* 227–233.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*(2), 150–174.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21–28

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29–40.

Skorupski, W. P. (2008). *A review and empirical comparison of approaches for improving the reliability of objective level scores*. Paper presented at the annual meeting of A Study of the Council of Chief State School Officers.

Skorupski, W. P. & Carvajal, J. (2009). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement, 70*(3), 357–375.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*(5), 893-898.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23,* 63–86.

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

Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (Fifth ed.). Boston: Allyn and Bacon.

Tate, R. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education, 17*(2), 89–112.

Thissen, D. & Edwards, M. C. (2005). *Diagnostic scores augmented using multidimensional item response theory: Preliminary investigation of MCMC strategies*. Paper presented at the annual the National Council on Educational Measurement, Montreal, Canada.

Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 141-186). Mahwah, NJ: Erlbaum.

Van der Linden, W., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 law school admissions test as an example. *Applied Measurement in Education, 8*(2), 157–186.

Wainer, H. & Dorans, N. J. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum.

Wainer, H. & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics, 12*, 339–368.

Wainer, H., Vevea, J. L., Camacho, F., Reeve III, B. B., Rosa, K., Nelson, L., Swygert, K. A., & Thissen, D. (2001). Augmented scores—"Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–388). Mahwah, NJ: Erlbaum.

Wainer, H. & Wang, C. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement, 37*, 203–220.

Walker, C. M., & Beretvas, S. N. (2000). *Using multidimensional versus unidimensional ability estimates to determine student proficiency in mathematics*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement, 38,* 147–163.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40,* 255–275.

Waller, N. (2002). MicroFACT: *A microcomputer factor analysis program for ordered polytomous data and mainframe size problems* [Software]. St. Paul, MN: Assessment Systems Corporation.

Wang, W., Chen, P., & Cheng, Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods,* 9, 116–136.

Watkins, M., Glutting, J., & Youngstrom, E. (2005). Issues in subtest profile analysis. In D.P. Flanagan & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 251-268). New York: Guilford Press.

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimation. *Applied Psychological Measurement, 12,* 239–252.

Whitely, S. E. (1991). Measuring aptitude processes with multicomponent latent trait models. Technical report, Lawrence: University of Kansas.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika*, *45*, 497-494.

Wilson, D. T., Wood, R., & Gibbons, R. (1998). *TESTFACT: Test scoring, item statistics, and item factor analysis* [Software]. Chicago: Scientific Software International.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *ACER ConQuest: Generalized item response modeling software version 2.0* [Software]. Melbourne, Australia: Australian Council for Educational Research.

Yao, L. (2003). *BMIRT: Bayesian multivariate item response theory* [Software]. Monterey, CA: Defense Manpower Data Center. Available from http://www.bmirt.com.

Yao, L. (2003). *SimuMIRT* [Software]. Monterey, CA: Defense Manpower Data Center. Available from http://www.bmirt.com.

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement, 47*(3), 339–360.

Yao, L. (2011). Multidimensional linking for domain scores and overall scores for nonequivalent groups. *Applied Psychological Measurement, 35*, 48–66.

Yao, L. (2013). Multidimensional Item Response Theory for Score Reporting. In Chang & Cheng (Ed), *Advances in Modern, International Testing: Transition from Summative to Formative Assessment*. Charlotte, NC: Information Age Publishing.

Yao, L., & Boughton, K. A. (2007). A Multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement*. *31*, 83–105.

Yao, L., & Boughton, K. A. (2009). Multidimensional linking for tests containing polytomous items. *Journal of Educational Measurement, 46,* 177—197.

Yao, L., & Schwarz, R. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement, 30*, 469–492.

Yen, W. M. (1987). *A Bayesian/IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada.

Yon, H. (2006). Multidimensional Item Response Theory (MIRT) approaches to vertical scaling (Doctoral Dissertation). Michigan State University, Lansing, MI.

Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement, 36*(5), 375–398.

Appendix A1

SimuMIRT Control and Batch Files

```
REM SimuMIRT Batch File
REM Simulating 10 datasets based on MIRT compensatory model
for %%f in (1  2  3  4  5  6  7  8  9  10) do call SimulateRwo
Parameter/sctl1_%%f.par Datasets/sctl1_%%f
pause;
```

sctl1_1 through sctl1_10 are the parameter files. An example item parameter file is shown below. The first line shows the total number of items, sample size, number of dimensions, means for each dimension, variance-covariance matrix of dimensions, maximum response level, random seed for ability generation, random seed for data generation. After the first line, item parameters based on a simple structure is presented.

```
30 1500 3 0 0 0 1 0.3 0.3 0.3 1 0.3 0.3 0.3 1 5 9001 867989
1    1    2.25      0    0    1.8186    0.0747
2    1    2.4133    0    0    -1.0097   0.2175
3    1    2.0419    0    0    -0.0084   0.2335
4    1    0.969     0    0    0.4482    0.1840
5    1    0.9449    0    0    -0.9046   0.0514
6    1    2.3515    0    0    -1.1728   0.1696
7    1    2.227     0    0    -0.1464   0.0967
8    1    1.5589    0    0    0.1868    0.2220
9    1    1.6131    0    0    -0.5554   0.0129
10   1    1.8727    0    0    0.3856    0.1970
11   1    0    2.4685    0    1.0619    0.0683
12   1    0    1.5533    0    -1.1986   0.0348
13   1    0    1.024     0    -0.1942   0.1002
14   1    0    0.914     0    -0.947    0.1084
15   1    0    1.1497    0    1.5473    0.0879
16   1    0    1.8895    0    1.6996    0.2309
17   1    0    2.2246    0    0.5514    0.1291
18   1    0    1.2891    0    0.8959    0.0816
19   1    0    1.4597    0    0.0865    0.0720
20   1    0    0.8712    0    1.4818    0.2419
21   1    0    0    1.2896    1.4225    0.2405
22   1    0    0    2.4976    -0.2124   0.1590
23   1    0    0    1.1607    -1.5689   0.1371
24   1    0    0    1.9731    0.5155    0.0786
25   1    0    0    1.7548    -0.6218   0.0081
26   1    0    0    2.4172    1.2701    0.1891
27   1    0    0    1.5476    -0.1565   0.1465
28   1    0    0    0.9018    -0.2182   0.0297
29   1    0    0    1.2977    -1.6267   0.2445
30   1    0    0    1.5863    -0.3844   0.0181
```

Appendix A2

BMIRT Control and Batch Files

```
REM BMIRT Control File
REM Estimating subscores based on compensatory MIRT model
1500 30 3 100 99631 0.01 1
11111111111111111111111111111
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
1 0.3 0.3 0.3 1 0.3 0.3 0.3 1 0 0 0
```

First line: Sample size, number of items, number of dimensions, number of iterations, random seed, delta, estimation method (1: MAP, 0: MLE).

Second line: 1 is indicating multiple-choice format for all items

Third line: Item numbers

Fourth line: Variance-covariance matrix of dimensions and means of dimensions. These are strong priors.

To run MAP estimation for 10 replications, the following batch file should be used:

```
REM BMIRT Batch File
for %%f in (1 2 3 4 5 6 7 8 9 10) do call BayesianModeAbility control_%%f.ctl
sctl1_%%f.rwo sctl1_%%f.par out/tctl1_%%f
pause;
```

This batch file uses the same item parameter files used for SimuMIRT. It reads control files and item parameter files for each replication, estimates subscores, and creates an output file in “out” folder.

Appendix A3

R Codes for Computing Reliability Coefficients

```
profile.r <- function(form1, form2) {

#Form1 and Form2 are data frames that include the subscores
n <- ncol(form1)
k <- nrow(form1)

#Average score for each person (level)
f1 <- as.matrix(rowMeans(form1),ncol=1,nrow=k)
f2 <- as.matrix(rowMeans(form2),ncol=1,nrow=k)

pattern1 <- matrix(,ncol=n, nrow=k)
pattern2 <- matrix(,ncol=n, nrow=k)

#Creating pattern scores
for (i in 1:n) {
        pattern1[,i] <- form1[,i]-f1
        pattern2[,i] <- form2[,i]-f2
}

#Overall profile reliability
covar1 <- matrix(,ncol=n, nrow=1)

for (i in 1:n) {covar1[,i]=cov(form1[,i],form2[,i])}
num1=rowSums(covar1)

variance.form1 <- sum(apply(form1,2,var))
variance.form2 <- sum(apply(form2,2,var))

denum1 <- sqrt(variance.form1*variance.form2)

overall <- num1/denum1

#Level reliability
num2 <- n*(cov(f1,f2))
denum2 <- sqrt((n*var(f1))*(n*var(f2)))
level <- num2/denum2

#Pattern reliability
covar2 <- matrix(,ncol=n, nrow=1)

for (i in 1:n) {covar2[,i]=cov(pattern1[,i],pattern2[,i])}
num3=rowSums(covar2)

var.form1 <- sum(apply(pattern1,2,var))
var.form2 <- sum(apply(pattern2,2,var))
```

```
denum3 <- sqrt(var.form1*var.form2)

pattern <- num3/denum3

  #Function profile.r returns the estimated values of level, pattern, and overall reliabilities.
  result <- cbind(level,pattern,overall)
  return(result)
}
```

Appendix B1

Correlations of Multidimensional Subscore Estimates

Table B1.1. Correlations of the Multidimensional Subscore Estimates from Three Subtests in Form 1

Subtest Length	ρ		S1	S2	S3
10		S1	1		
	.3	S2	.36	1	
		S3	.36	.36	1
		S1	1		
	.5	S2	.59	1	
		S3	.59	.59	1
		S1	1		
	.8	S2	.89	1	
		S3	.89	.89	1
20		S1	1		
	.3	S2	.34	1	
		S3	.34	.33	1
		S1	1		
	.5	S2	.56	1	
		S3	.56	.56	1
		S1	1		
	.8	S2	.87	1	
		S3	.87	.87	1
40		S1	1		
	.3	S2	.32	1	
		S3	.32	.32	1
		S1	1		
	.5	S2	.53	1	
		S3	.54	.53	1
		S1	1		
	.8	S2	.85	1	
		S3	.85	.85	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3. ρ: True correlation between subscores used in the data generation.

Table B1.2. Correlations of the Multidimensional Subscore Estimates from Three Subtests in Form 2

Subtest Length	ρ		S1	S2	S3
10		S1	1		
	.3	S2	.36	1	
		S3	.36	.36	1
		S1	1		
	.5	S2	.59	1	
		S3	.59	.59	1
		S1	1		
	.8	S2	.89	1	
		S3	.89	.89	1
20		S1	1		
	.3	S2	.34	1	
		S3	.34	.33	1
		S1	1		
	.5	S2	.56	1	
		S3	.56	.56	1
		S1	1		
	.8	S2	.87	1	
		S3	.87	.87	1
40		S1	1		
	.3	S2	.32	1	
		S3	.32	.32	1
		S1	1		
	.5	S2	.54	1	
		S3	.54	.53	1
		S1	1		
	.8	S2	.85	1	
		S3	.85	.85	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3. ρ: True correlation between subscores used in the data generation.

Table B1.3. Correlations of the Multidimensional Subscore Estimates from Five Subtests in Form 1

Subtest Length	ρ		S1	S2	S3	S4	S5
10	.3	S1	1				
		S2	.35	1			
		S3	.35	.35	1		
		S4	.35	.35	.35	1	
		S5	.35	.35	.35	.35	1
	.5	S1	1				
		S2	.59	1			
		S3	.59	.59	1		
		S4	.59	.59	.59	1	
		S5	.59	.59	.59	.59	1
	.8	S1	1				
		S2	.89	1			
		S3	.89	.89	1		
		S4	.89	.89	.89	1	
		S5	.89	.89	.89	.89	1
20	.3	S1	1				
		S2	.33	1			
		S3	.33	.33	1		
		S4	.33	.34	.33	1	
		S5	.33	.34	.33	.33	1
	.5	S1	1				
		S2	.56	1			
		S3	.56	.56	1		
		S4	.56	.56	.56	1	
		S5	.56	.56	.56	.56	1
	.8	S1	1				
		S2	.87	1			
		S3	.87	.87	1		
		S4	.87	.87	.87	1	
		S5	.87	.87	.87	.87	1
40	.3	S1	1				
		S2	.32	1			
		S3	.32	.33	1		
		S4	.33	.32	.32	1	
		S5	.32	.32	.32	.32	1
	.5	S1	1				
		S2	.53	1			
		S3	.53	.53	1		
		S4	.54	.53	.53	1	
		S5	.53	.53	.53	.53	1
	.8	S1	1				
		S2	.84	1			
		S3	.84	.84	1		
		S4	.84	.84	.85	1	
		S5	.85	.84	.84	.84	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5. ρ: True correlation between subscores used in the data generation.

Table B1.4. Correlations of the Multidimensional Subscore Estimates from Five Subtests in Form 2

Subtest Length	ρ		S1	S2	S3	S4	S5
10		S1	1				
		S2	.35	1			
	.3	S3	.35	.35	1		
		S4	.35	.35	.35	1	
		S5	.35	.35	.35	.35	1
		S1	1				
		S2	.59	1			
	.5	S3	.59	.59	1		
		S4	.59	.59	.59	1	
		S5	.59	.59	.59	.59	1
		S1	1				
		S2	.89	1			
	.8	S3	.89	.89	1		
		S4	.89	.89	.89	1	
		S5	.89	.89	.89	.89	1
20		S1	1				
		S2	.33	1			
	.3	S3	.33	.33	1		
		S4	.33	.34	.33	1	
		S5	.33	.34	.33	.33	1
		S1	1				
	.5	S2	.56	1			
		S3	.56	.56	1		
		S4	.56	.56	.56	1	
		S5	.56	.56	.56	.56	1
		S1	1				
		S2	.87	1			
	.8	S3	.87	.87	1		
		S4	.87	.87	.87	1	
		S5	.87	.87	.87	.87	1
40		S1	1				
		S2	.32	1			
	.3	S3	.32	.33	1		
		S4	.33	.32	.32	1	
		S5	.32	.32	.32	.32	1
		S1	1				
		S2	.53	1			
	.5	S3	.53	.53	1		
		S4	.54	.53	.53	1	
		S5	.53	.53	.53	.53	1
		S1	1				
	.8	S2	.84	1			
		S3	.84	.84	1		
		S4	.84	.84	.85	1	
		S5	.85	.84	.84	.84	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5. ρ: True correlation between subscores used in the data generation.

Table B1.5. Correlations of the Multidimensional Subscore Estimates from Seven Subtests in Form 1

Subtest Length	ρ		S1	S2	S3	S4	S5	S6	S7
10		S1	1						
		S2	.36	1					
		S3	.36	.35	1				
	.3	S4	.36	.36	.36	1			
		S5	.36	.35	.36	.36	1		
		S6	.36	.36	.36	.36	.36	1	
		S7	.36	.36	.36	.36	.36	.3	1
		S1	1						
		S2	.59	1					
		S3	.58	.59	1				
	.5	S4	.59	.59	.59	1			
		S5	.59	.58	.59	.59	1		
		S6	.59	.59	.59	.59	.58	1	
		S7	.59	.59	.59	.59	.58	.59	1
		S1	1						
		S2	.89	1					
		S3	.89	.88	1				
	.8	S4	.89	.89	.89	1			
		S5	.89	.89	.89	.89	1		
		S6	.89	.89	.88	.89	.89	1	
		S7	.89	.89	.89	.89	.88	.89	1
20		S1	1						
		S2	.33	1					
		S3	.33	.33	1				
	.3	S4	.33	.33	.33	1			
		S5	.33	.33	.33	.33	1		
		S6	.33	.33	.33	.33	.33	1	
		S7	.33	.33	.33	.33	.33	.33	1
		S1	1	.					
		S2	.56	1					
		S3	.56	.56	1				
	.5	S4	.56	.56	.56	1			
		S5	.57	.56	.56	.56	1		
		S6	.56	.56	.56	.56	.56	1	
		S7	.56	.56	.57	.56	.56	.56	1
		S1	1	.					
		S2	.87	1					
		S3	.86	.87	1				
	.8	S4	.87	.87	.87	1			
		S5	.87	.86	.86	.87	1		
		S6	.86	.87	.87	.87	.87	1	
		S7	.87	.87	.87	.87	.87	.87	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore7. ρ: True correlation between subscores used in the data generation.

Table B1.5. Correlations of the Multidimensional Subscore Estimates from Seven Subtests in Form 1 (Cont.)

Subtest Length	ρ		S1	S2	S3	S4	S5	S6	S7
40		S1	1						
		S2	.33	1					
		S3	.32	.32	1				
	.3	S4	.33	.32	.32	1			
		S5	.33	.32	.32	.32	1		
		S6	.32	.32	.32	.33	.32	1	
		S7	.32	.32	.32	.32	.32	.32	1
		S1	1						
		S2	.53	1					
		S3	.54	.53	1				
	.5	S4	.53	.53	.53	1			
		S5	.53	.53	.54	.53	1		
		S6	.54	.53	.53	.54	.53	1	
		S7	.54	.53	.54	.53	.54	.53	1
		S1	1						
		S2	.84	1					
		S3	.84	.85	1				
	.8	S4	.85	.84	.84	1			
		S5	.84	.84	.84	.84	1		
		S6	.84	.84	.84	.84	.84	1	
		S7	.84	.84	.84	.84	.85	.84	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore7. ρ: True correlation between subscores used in the data generation.

Table B1.6. Correlations of the Multidimensional Subscore Estimates from Seven Subtests in Form 2

Subtest Length	ρ		S1	S2	S3	S4	S5	S6	S7
10		S1	1						
		S2	.36	1					
		S3	.35	.35	1				
	.3	S4	.36	.36	.36	1			
		S5	.36	.35	.36	.36	1		
		S6	.36	.36	.36	.36	.36	1	
		S7	.36	.36	.36	.36	.36	.3	1
		S1	1						
		S2	.58	1					
		S3	.58	.59	1				
	.5	S4	.59	.59	.59	1			
		S5	.59	.58	.59	.59	1		
		S6	.59	.59	.59	.59	.58	1	
		S7	.59	.59	.59	.59	.58	.59	1
		S1	1						
		S2	.89	1					
		S3	.89	.88	1				
	.8	S4	.89	.89	.89	1			
		S5	.89	.89	.89	.89	1		
		S6	.89	.89	.88	.89	.89	1	
		S7	.89	.89	.89	.89	.88	.89	1
20		S1	1						
		S2	.33	1					
		S3	.33	.33	1				
	.3	S4	.33	.33	.33	1			
		S5	.33	.33	.33	.33	1		
		S6	.33	.33	.33	.33	.33	1	
		S7	.33	.33	.33	.33	.33	.33	1
		S1	1	.					
		S2	.56	1					
		S3	.56	.56	1				
	.5	S4	.56	.56	.56	1			
		S5	.57	.56	.56	.56	1		
		S6	.56	.56	.56	.56	.56	1	
		S7	.56	.56	.57	.56	.56	.56	1
		S1	1	.					
		S2	.86	1					
		S3	.86	.87	1				
	.8	S4	.87	.87	.87	1			
		S5	.87	.86	.86	.87	1		
		S6	.86	.87	.87	.87	.87	1	
		S7	.87	.87	.87	.87	.87	.87	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore7. ρ: True correlation between subscores used in the data generation.

Table B1.6. Correlations of the Multidimensional Subscore Estimates from Seven Subtests in Form 2 (Cont.)

Subtest Length	ρ		S1	S2	S3	S4	S5	S6	S7
40		S1	1						
		S2	.32	1					
		S3	.32	.32	1				
	.3	S4	.33	.32	.32	1			
		S5	.33	.32	.32	.32	1		
		S6	.32	.32	.32	.33	.32	1	
		S7	.32	.32	.32	.32	.32	.32	1
		S1	1						
		S2	.53	1					
		S3	.53	.53	1				
	.5	S4	.53	.53	.53	1			
		S5	.53	.53	.54	.53	1		
		S6	.54	.53	.53	.54	.53	1	
		S7	.54	.53	.54	.53	.54	.53	1
		S1	1						
		S2	.84	1					
		S3	.84	.85	1				
	.8	S4	.85	.84	.84	1			
		S5	.84	.84	.84	.84	1		
		S6	.84	.84	.84	.84	.84	1	
		S7	.84	.84	.84	.84	.85	.84	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore7. ρ: True correlation between subscores used in the data generation.

Appendix B2

Correlations of Unidimensional Subscore Estimates

Table B2.1. Correlations of the Unidimensional Subscore Estimates from Three Subtests in Form 1

Subtest Length	ρ		S1	S2	S3
10		S1	1		
	.3	S2	.22	1	
		S3	.22	.21	1
		S1	1		
	.5	S2	.36	1	
		S3	.36	.36	1
		S1	1		
	.8	S2	.59	1	
		S3	.58	.58	1
20		S1	1		
	.3	S2	.25	1	
		S3	.25	.25	1
		S1	1		
	.5	S2	.42	1	
		S3	.42	.42	1
		S1	1		
	.8	S2	.67	1	
		S3	.67	.67	1
40		S1	1		
	.3	S2	.27	1	
		S3	.27	.27	1
		S1	1		
	.5	S2	.45	1	
		S3	.45	.45	1
		S1	1		
	.8	S2	.73	1	
		S3	.73	.73	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3. ρ: True correlation between subscores used in the data generation.

Table B2.2. Correlations of the Unidimensional Subscore Estimates from Three Subtests in Form 2

Subtest Length	ρ		S1	S2	S3
10		S1	1		
	.3	S2	.22	1	
		S3	.21	.21	1
		S1	1		
	.5	S2	.36	1	
		S3	.36	.36	1
		S1	1		
	.8	S2	.58	1	
		S3	.58	.58	1
20		S1	1		
	.3	S2	.25	1	
		S3	.25	.25	1
		S1	1		
	.5	S2	.42	1	
		S3	.42	.42	1
		S1	1		
	.8	S2	.67	1	
		S3	.67	.67	1
40		S1	1		
	.3	S2	.27	1	
		S3	.27	.27	1
		S1	1		
	.5	S2	.45	1	
		S3	.45	.45	1
		S1	1		
	.8	S2	.73	1	
		S3	.73	.73	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3. ρ: True correlation between subscores used in the data generation.

Table B2.3. Correlations of the Unidimensional Subscore Estimates from Five Subtests in Form 1

Subtest Length	ρ		S1	S2	S3	S4	S5
10		S1	1				
		S2	.21	1			
	.3	S3	.20	.20	1		
		S4	.22	.22	.21	1	
		S5	.21	.23	.21	.21	1
		S1	1				
		S2	.36	1			
	.5	S3	.34	.34	1		
		S4	.36	.36	.36	1	
		S5	.35	.37	.36	.36	1
		S1	1				
		S2	.58	1			
	.8	S3	.57	.57	1		
		S4	.58	.58	.58	1	
		S5	.57	.58	.57	.58	1
20		S1	1				
		S2	.24	1			
	.3	S3	.24	.23	1		
		S4	.26	.26	.24	1	
		S5	.25	.26	.25	.26	1
		S1	1				
	.5	S2	.41	1			
		S3	.41	.41	1		
		S4	.42	.42	.41	1	
		S5	.41	.43	.41	.42	1
		S1	1				
		S2	.67	1			
	.8	S3	.67	.67	1		
		S4	.67	.68	.67	1	
		S5	.67	.68	.67	.68	1
40		S1	1				
		S2	.26	1			
	.3	S3	.26	.25	1		
		S4	.28	.28	.27	1	
		S5	.26	.28	.27	.27	1
		S1	1		.		
		S2	.45	1			
	.5	S3	.44	.44	1		
		S4	.45	.46	.45	1	
		S5	.45	.46	.45	.45	1
		S1	1				
	.8	S2	.72	1			
		S3	.73	.72	1		
		S4	.73	.73	.73	1	
		S5	.72	.73	.73	.73	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5. ρ: True correlation between subscores used in the data generation.

Table B2.4. Correlations of the Unidimensional Subscore Estimates from Five Subtests in Form 2

Subtest Length	ρ		S1	S2	S3	S4	S5
10		S1	1				
		S2	.21	1			
	.3	S3	.20	.20	1		
		S4	.21	.22	.21	1	
		S5	.21	.22	.21	.21	1
		S1	1				
		S2	.35	1			
	.5	S3	.35	.35	1		
		S4	.35	.36	.36	1	
		S5	.35	.36	.36	.35	1
		S1	1				
		S2	.57	1			
	.8	S3	.57	.57	1		
		S4	.58	.58	.57	1	
		S5	.58	.58	.58	.58	1
20		S1	1				
		S2	.24	1			
	.3	S3	.23	.23	1		
		S4	.25	.25	.24	1	
		S5	.25	.26	.25	.24	1
		S1	1				
	.5	S2	.41	1			
		S3	.41	.41	1		
		S4	.42	.42	.41	1	
		S5	.41	.43	.41	.42	1
		S1	1				
		S2	.67	1			
	.8	S3	.67	.67	1		
		S4	.67	.68	.67	1	
		S5	.67	.68	.67	.68	1
40		S1	1				
		S2	.27	1			
	.3	S3	.26	.26	1		
		S4	.28	.28	.27	1	
		S5	.26	.28	.27	.27	1
		S1	1		.		
		S2	.45	1			
	.5	S3	.44	.44	1		
		S4	.45	.46	.45	1	
		S5	.45	.46	.45	.45	1
		S1	1				
	.8	S2	.72	1			
		S3	.73	.72	1		
		S4	.73	.73	.72	1	
		S5	.72	.73	.73	.72	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5. ρ: True correlation between subscores used in the data generation.

Table B2.5. Correlations of the Unidimensional Subscore Estimates from Seven Subtests in Form 1

Subtest Length	ρ		S1	S2	S3	S4	S5	S6	S7
10		S1	1						
		S2	21	1					
		S3	21	.22	1				
	.3	S4	21	21	21	1			
		S5	21	.22	21	21	1		
		S6	22	21	21	21	21	1	
		S7	21	21	21	21	21	21	1
		S1	1						
		S2	.35	1					
		S3	.36	.35	1				
	.5	S4	.35	.35	.35	1			
		S5	.35	.36	.35	.35	1		
		S6	.35	.35	.35	.35	.36	1	
		S7	.35	.35	.35	.35	.36	.35	1
		S1	1						
		S2	.57	1					
		S3	.57	.58	1				
	.8	S4	.57	.57	.57	1			
		S5	.57	.57	.57	.57	1		
		S6	.58	.57	.58	.57	.57	1	
		S7	.57	.57	.57	.57	.58	.57	1
20		S1	1						
		S2	.24	1					
		S3	.24	.24	1				
	.3	S4	.25	.25	.24	1			
		S5	.24	.24	.25	.24	1		
		S6	.24	.24	.24	.24	.24	1	
		S7	.25	.24	.24	.25	.24	.24	1
		S1	1						
		S2	.41	1					
		S3	.41	.41	1				
	.5	S4	.41	.44	.41	1			
		S5	.42	.41	.42	.41	1		
		S6	.41	.41	.41	.41	.41	1	
		S7	.41	.41	.42	.41	.43	.41	1
		S1	1	.					
		S2	.67	1					
		S3	.68	.67	1				
	.8	S4	.67	.67	.67	1			
		S5	.67	.68	.68	.67	1		
		S6	.68	.67	.67	.67	.67	1	
		S7	.67	.67	.67	.67	.67	.67	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore7. ρ: True correlation between subscores used in the data generation.

Table B2.5. Correlations of the Unidimensional Subscore Estimates from Seven Subtests in Form 1 (Cont.)

Subtest Length	ρ		S1	S2	S3	S4	S5	S6	S7
40		S1	1						
		S2	27	1					
		S3	.26	.26	1				
	.3	S4	27	.26	.26	1			
		S5	27	.26	.26	.26	1		
		S6	.26	.26	.26	27	.26	1	
		S7	.26	.26	.26	.26	.26	.26	1
		S1	1						
		S2	.45	1					
		S3	.44	.45	1				
	.5	S4	.45	.45	.45	1			
		S5	.45	.45	.44	.45	1		
		S6	.44	.45	.45	.44	.45	1	
		S7	.44	.45	.44	.45	.44	.45	1
		S1	1						
		S2	.72	1					
		S3	.72	.73	1				
	.8	S4	.73	.72	.72	1			
		S5	.72	.72	.72	.72	1		
		S6	.72	.72	.72	.72	.72	1	
		S7	.72	.72	.72	.72	.73	.72	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore7. ρ: True correlation between subscores used in the data generation.

Table B2.6. Correlations of the Unidimensional Subscore Estimates from Seven Subtests in Form 2

Subtest Length	ρ		S1	S2	S3	S4	S5	S6	S7
10		S1	1						
		S2	22	1					
		S3	21	.22	1				
	.3	S4	22	21	21	1			
		S5	21	.22	21	21	1		
		S6	22	21	21	22	21	1	
		S7	21	21	21	21	21	21	1
		S1	1						
		S2	.35	1					
		S3	.35	.35	1				
	.5	S4	.36	.35	.35	1			
		S5	.36	.36	.35	.35	1		
		S6	.35	.35	.37	.35	.36	1	
		S7	.35	.35	.35	.35	.36	.35	1
		S1	1						
		S2	.57	1					
		S3	.57	.57	1				
	.8	S4	.56	.57	.57	1			
		S5	.57	.57	.57	.57	1		
		S6	.58	.58	.58	.57	.57	1	
		S7	.57	.57	.57	.57	.57	.57	1
20		S1	1						
		S2	.24	1					
		S3	.24	.24	1				
	.3	S4	.24	.24	.24	1			
		S5	.24	.24	.25	.24	1		
		S6	.24	.24	.24	.24	.24	1	
		S7	.25	.24	.24	.24	.24	.24	1
		S1	1						
		S2	.41	1					
		S3	.42	.41	1				
	.5	S4	.41	.44	.41	1			
		S5	.41	.41	.42	.42	1		
		S6	.41	.41	.41	.41	.41	1	
		S7	.41	.41	.42	.41	.43	.41	1
		S1	1	.					
		S2	.67	1					
		S3	.68	.67	1				
	.8	S4	.67	.67	.67	1			
		S5	.67	.68	.68	.67	1		
		S6	.68	.67	.67	.67	.67	1	
		S7	.67	.67	.67	.67	.67	.67	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore7. ρ: True correlation between subscores used in the data generation.

Table B2.6. Correlations of the Unidimensional Subscore Estimates from Seven Subtests in Form 2 (Cont.)

Subtest Length	ρ		S1	S2	S3	S4	S5	S6	S7
40		S1	1						
		S2	27	1					
		S3	.26	.26	1				
	.3	S4	27	.27	.26	1			
		S5	27	.27	.26	.26	1		
		S6	.26	.26	.26	27	.26	1	
		S7	.26	.26	.26	.26	.26	.26	1
		S1	1						
		S2	.44	1					
		S3	.44	.45	1				
	.5	S4	.45	.45	.45	1			
		S5	.45	.45	.45	.45	1		
		S6	.45	.45	.45	.43	.45	1	
		S7	.44	.45	.44	.45	.44	.45	1
		S1	1						
		S2	.72	1					
		S3	.71	.73	1				
	.8	S4	.73	.71	.72	1			
		S5	.72	.72	.72	.72	1		
		S6	.72	.72	.72	.72	.73	1	
		S7	.72	.72	.72	.72	.73	.72	1

Note: S1: Subscore 1; S2: Subscore 2; S3: Subscore 3; S4: Subscore 4; S5: Subscore 5; S6: Subscore 6; S7: Subscore7. ρ: True correlation between subscores used in the data generation.

Appendix C

Sampling Distributions of Between-person, Within-person, and Total Profile Reliability

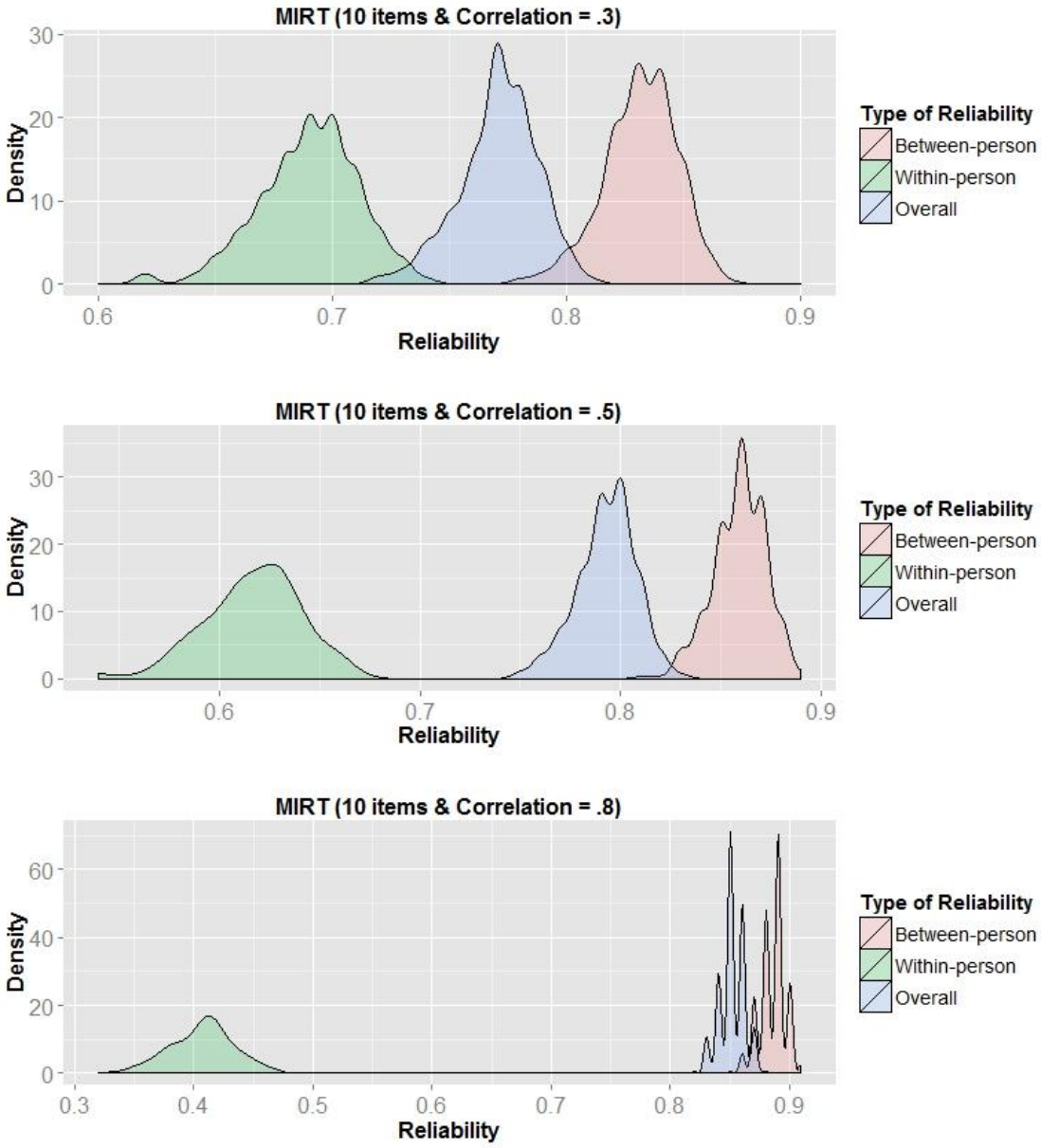

Figure C1.1. Sampling distributions of reliability estimates from 3-dimensional MIRT model with 10 items.

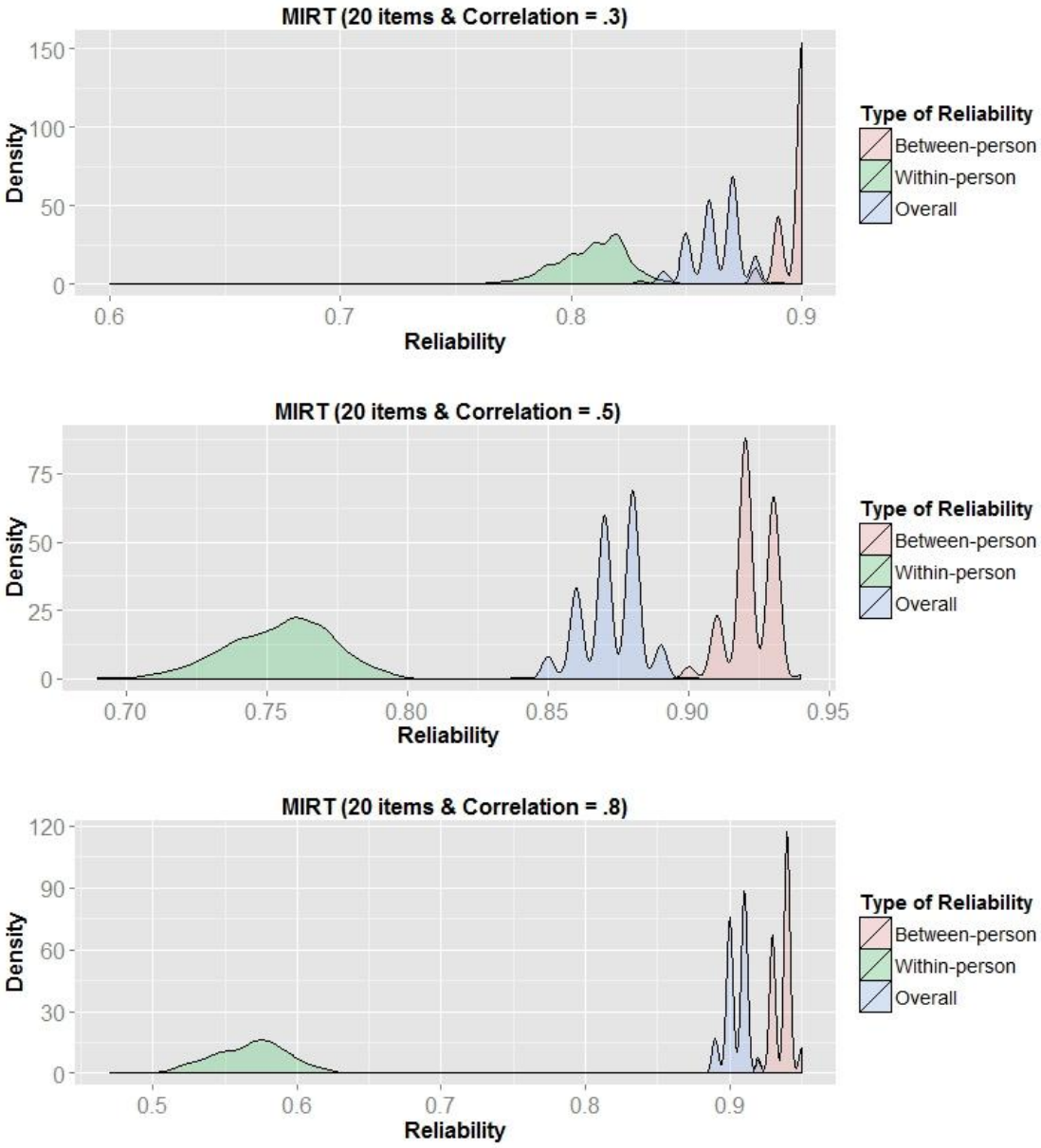

Figure C1.2. Sampling distributions of reliability estimates from 3-dimensional MIRT model with 20 items.

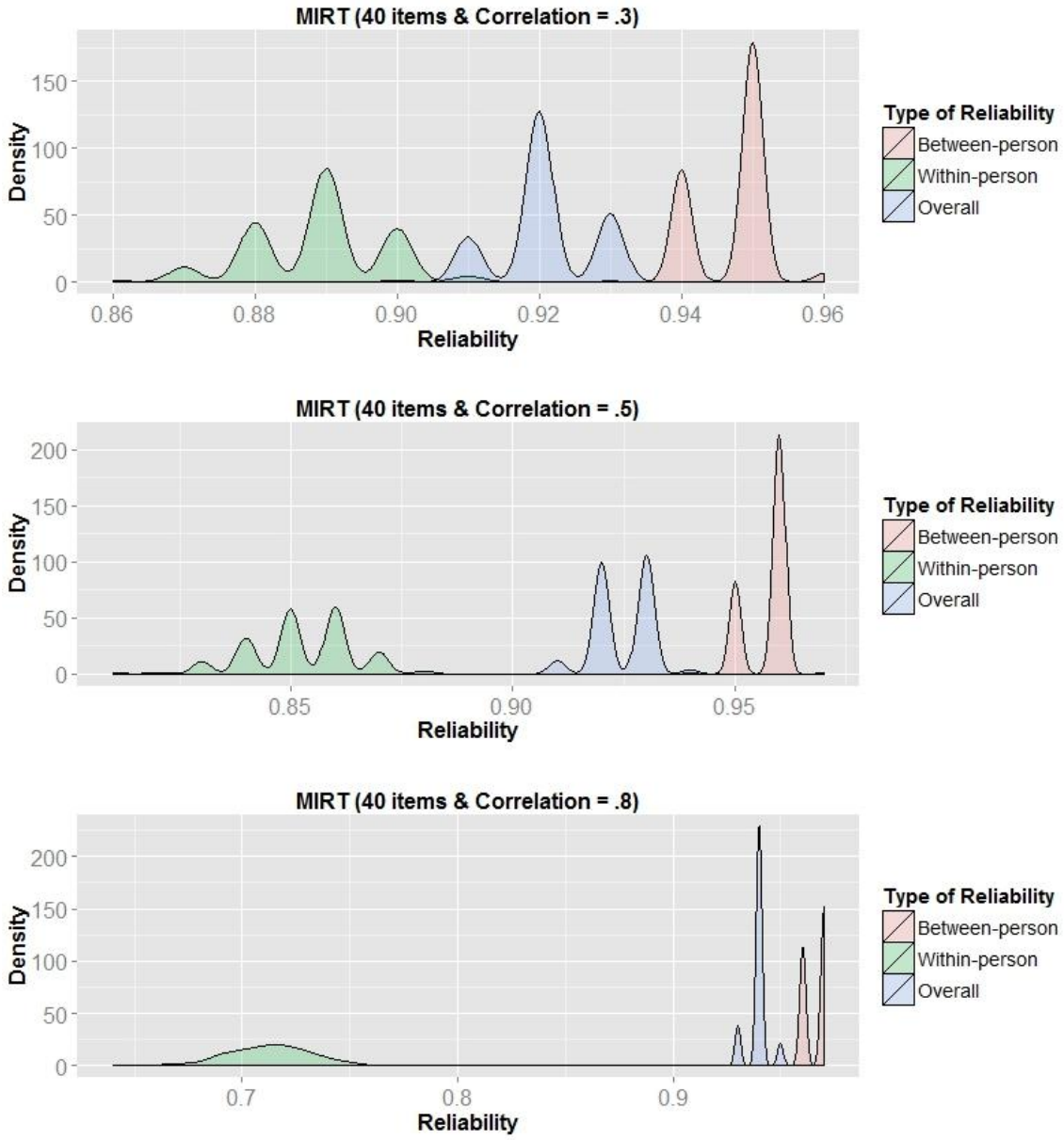

Figure C1.3. Sampling distributions of reliability estimates from 3-dimensional MIRT model with 40 items.

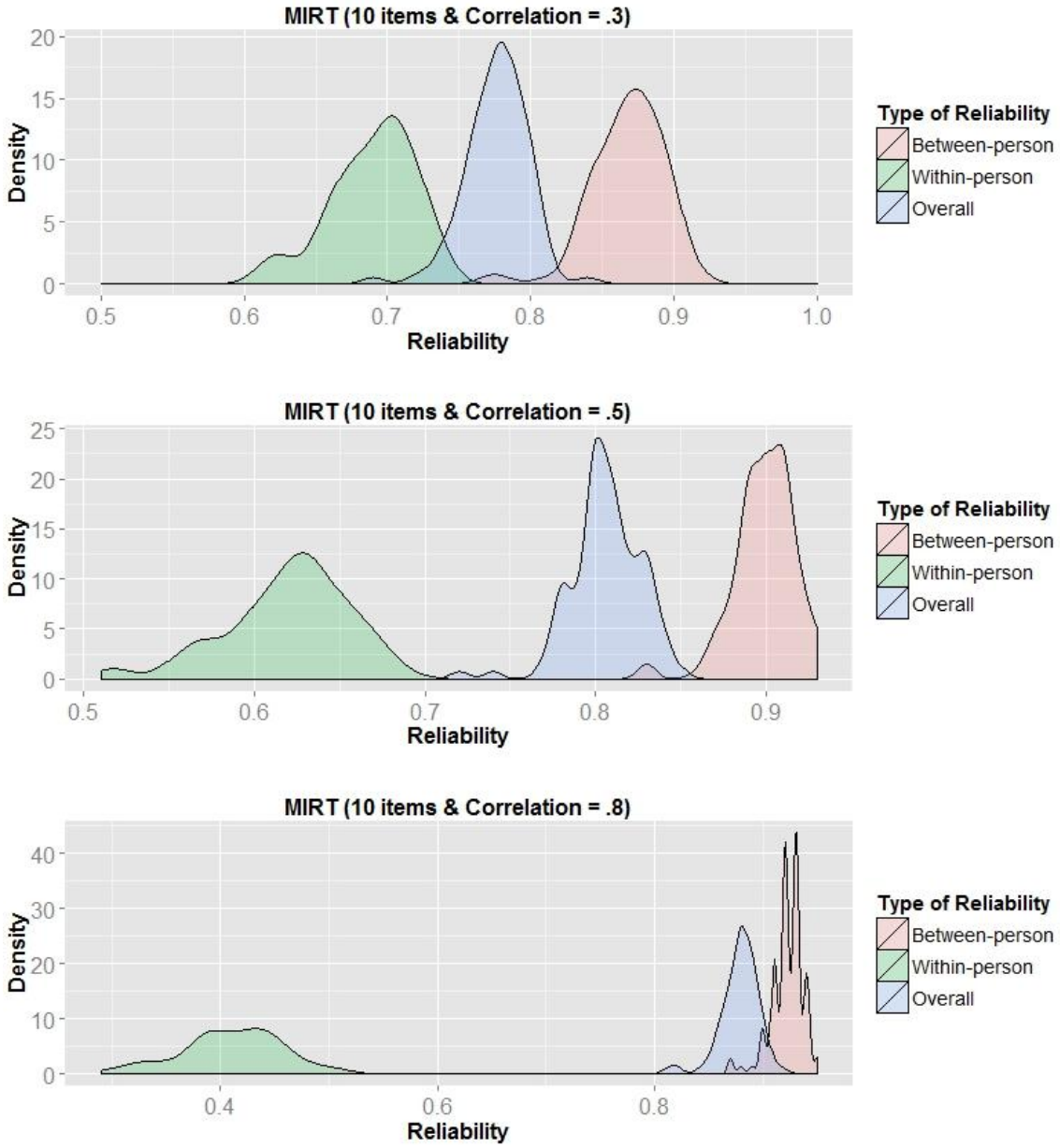

Figure C1.4. Sampling distributions of reliability estimates from 5-dimensional MIRT model with 10 items.

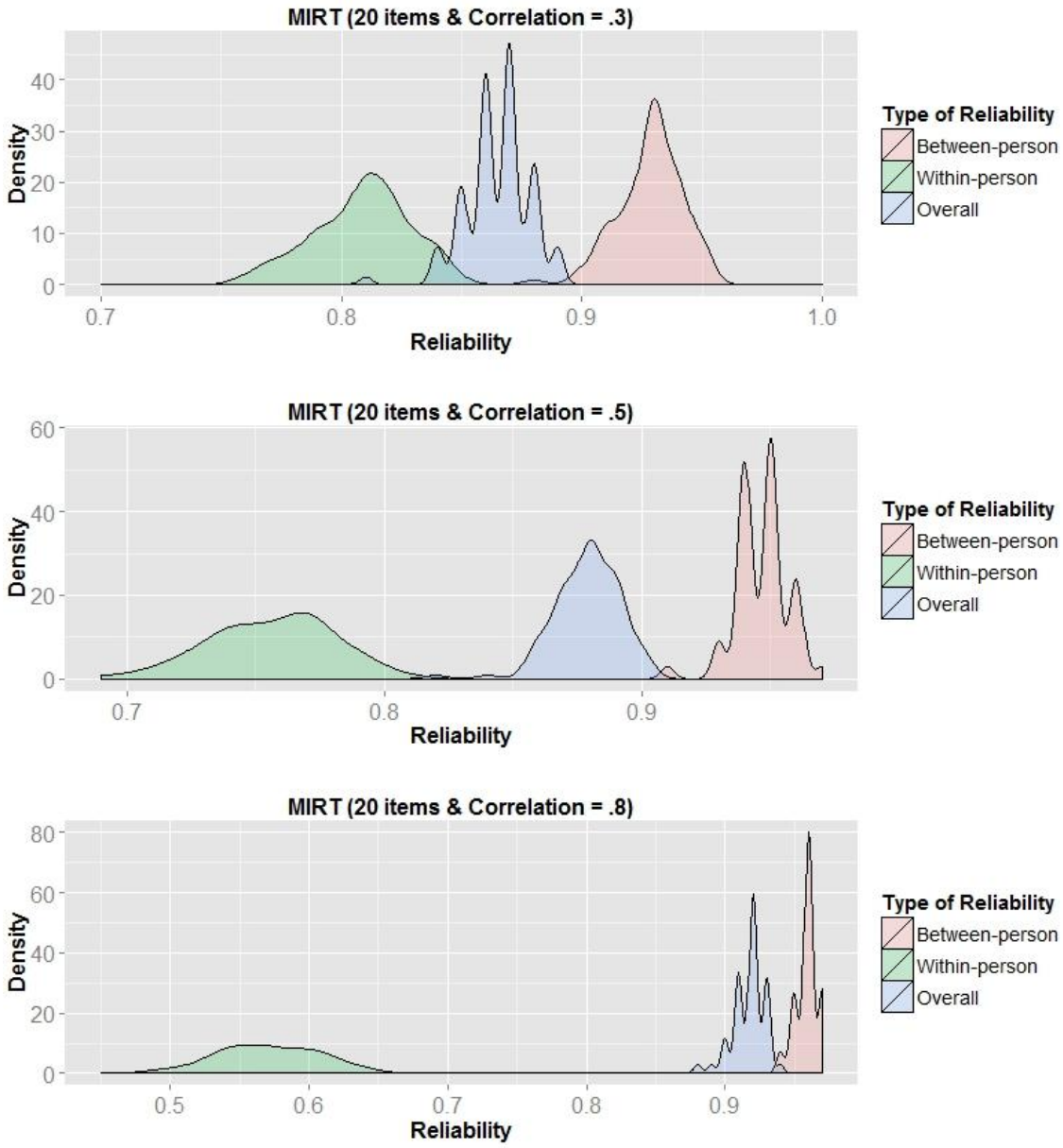

Figure C1.5. Sampling distributions of reliability estimates from 5-dimensional MIRT model with 20 items.

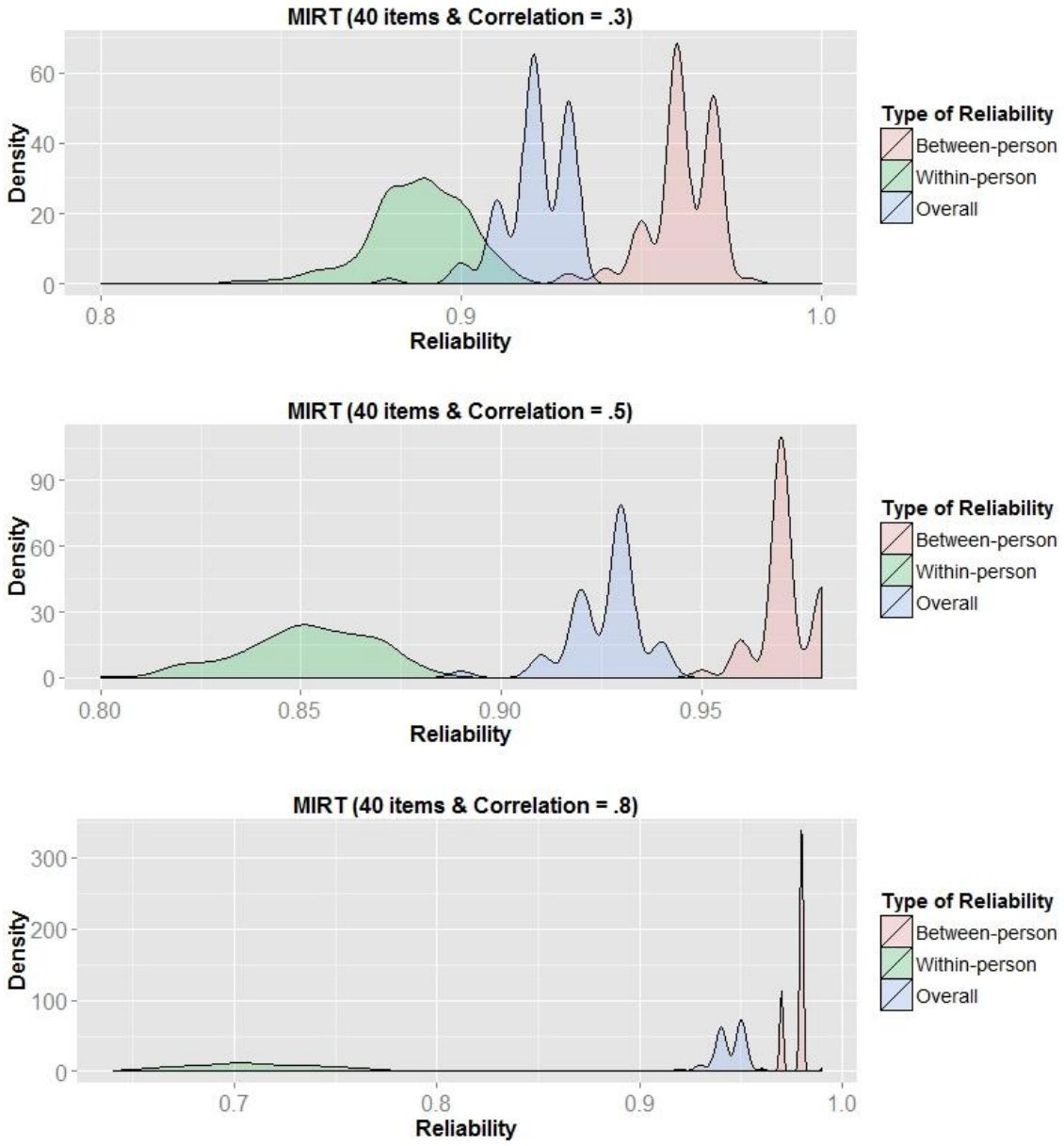

Figure C1.6. Sampling distributions of reliability estimates from 5-dimensional MIRT model with 40 items.

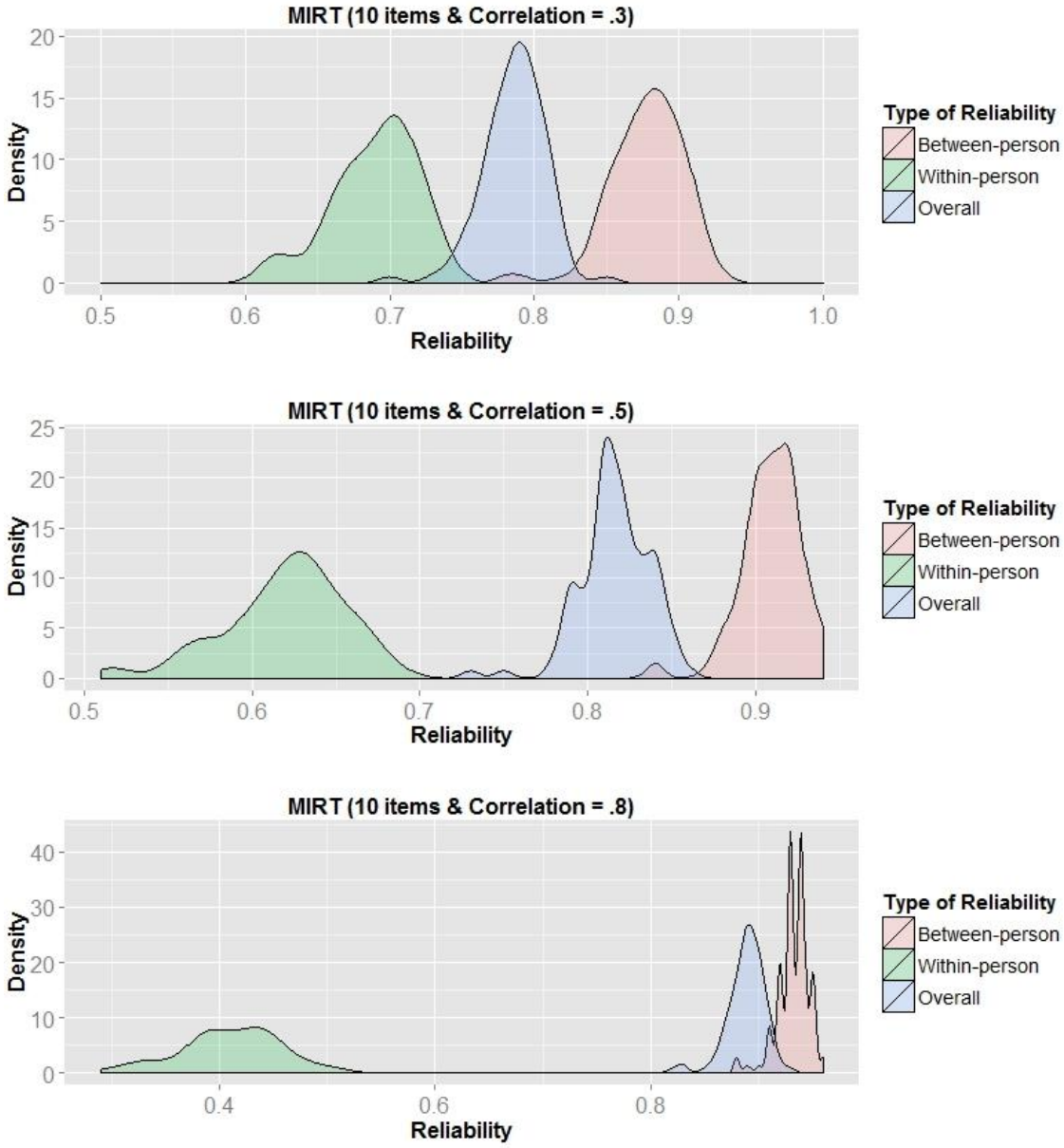

Figure C1.7. Sampling distributions of reliability estimates from 7-dimensional MIRT model with 10 items.

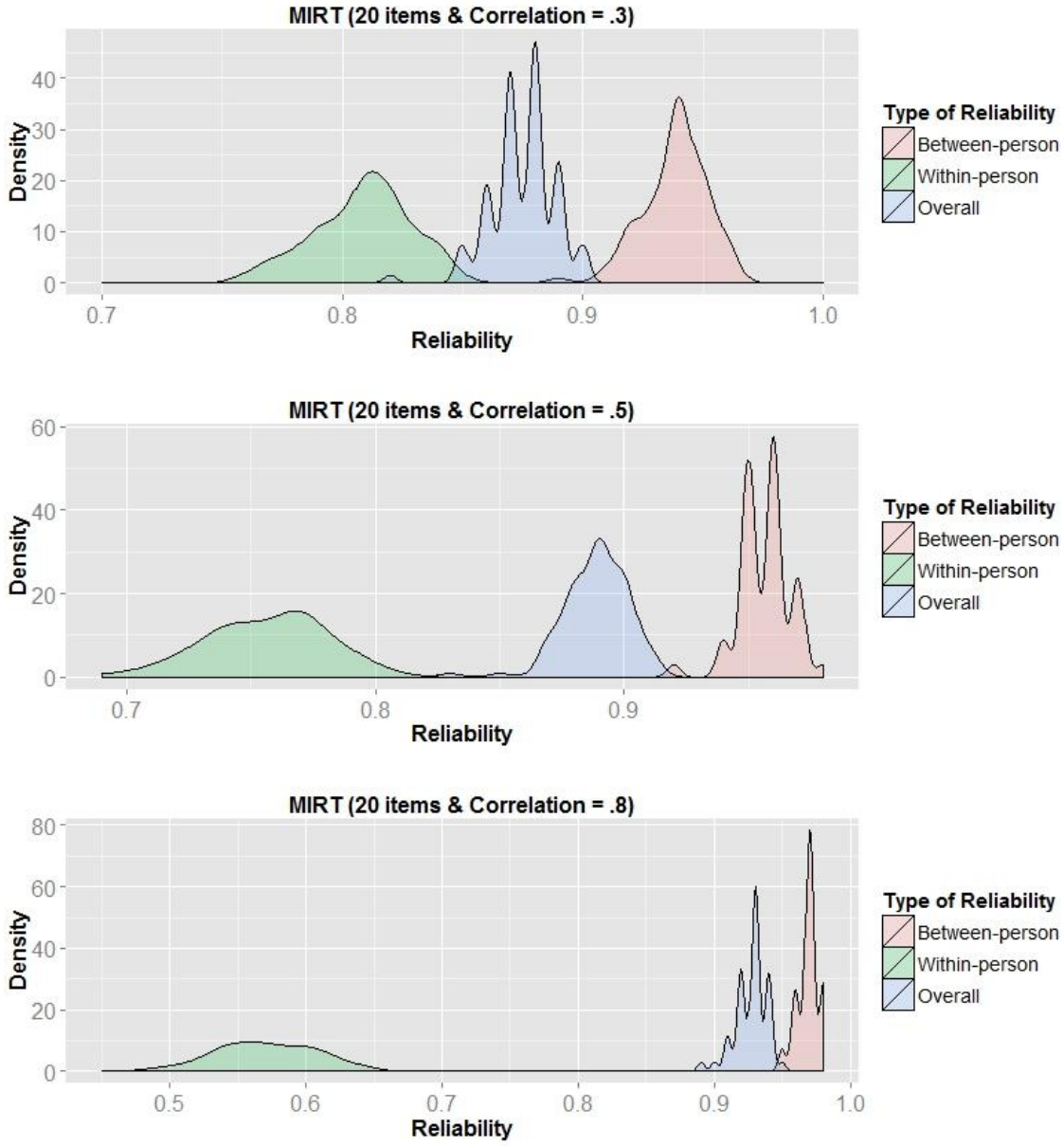

Figure C1.8. Sampling distributions of reliability estimates from 7-dimensional MIRT model with 20 items.

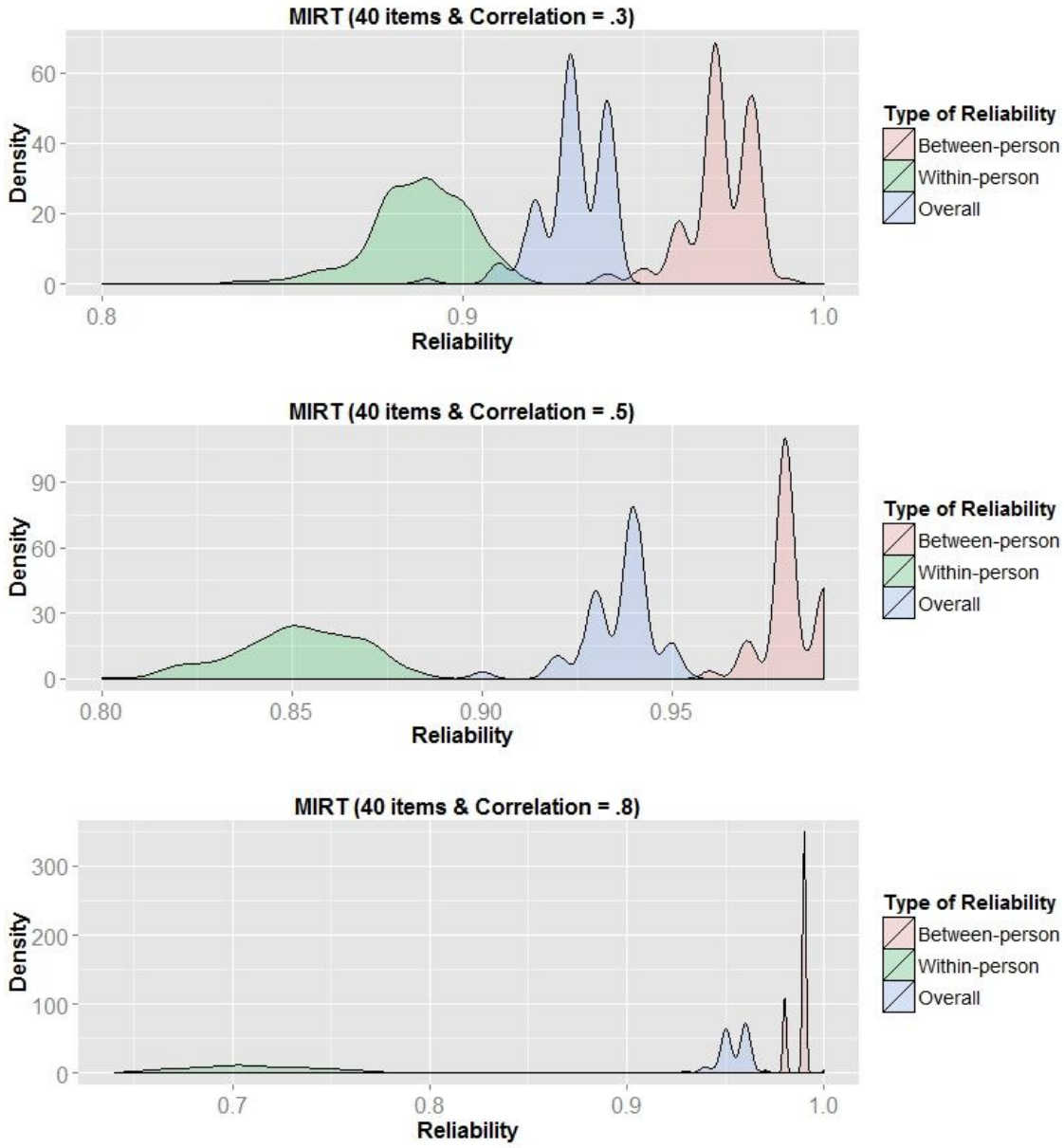

Figure C1.9. Sampling distributions of reliability estimates from 7-dimensional MIRT model with 40 items.

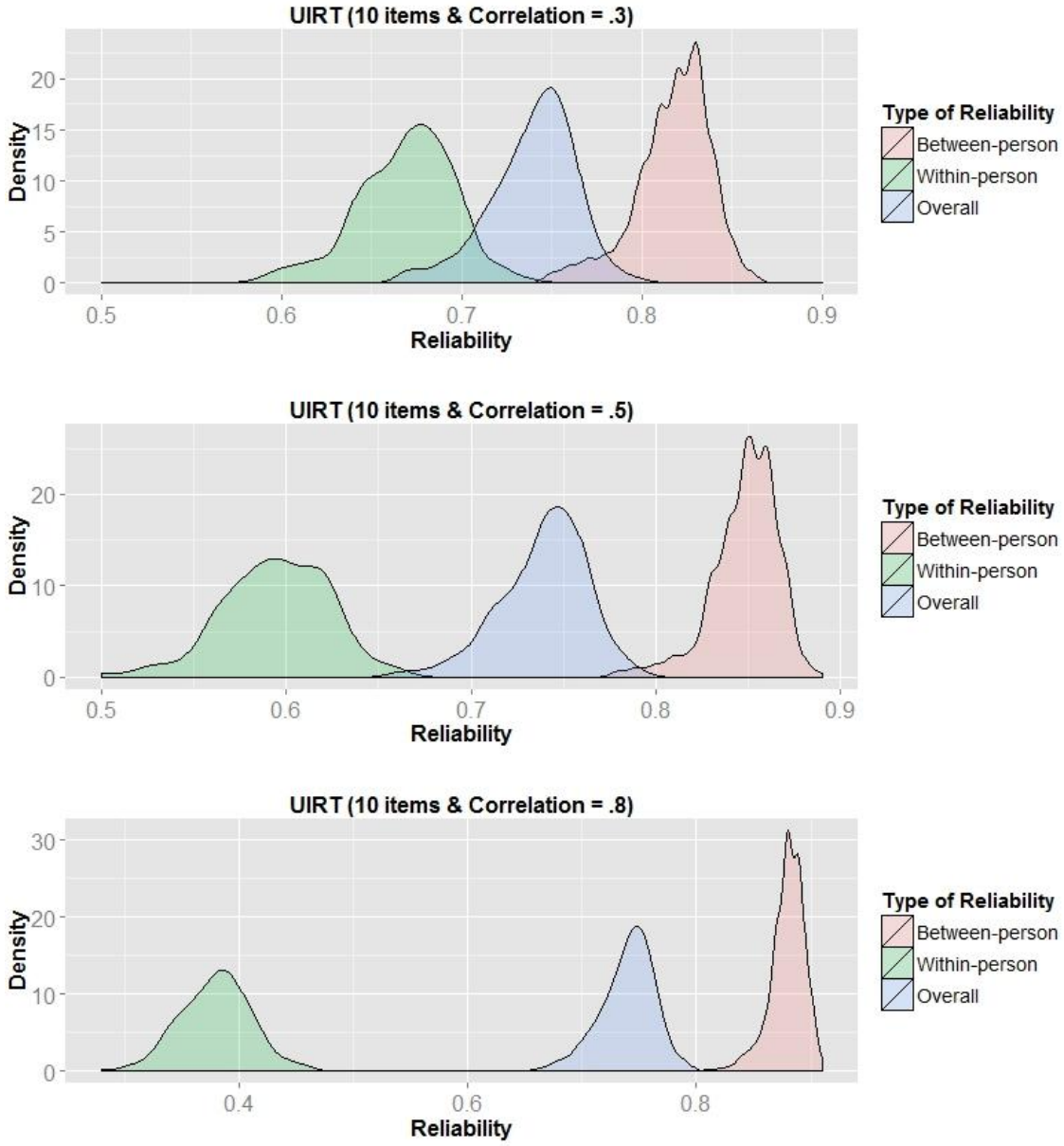

Figure C1.10. Sampling distributions of reliability estimates from 3-dimensional UIRT model with 10 items.

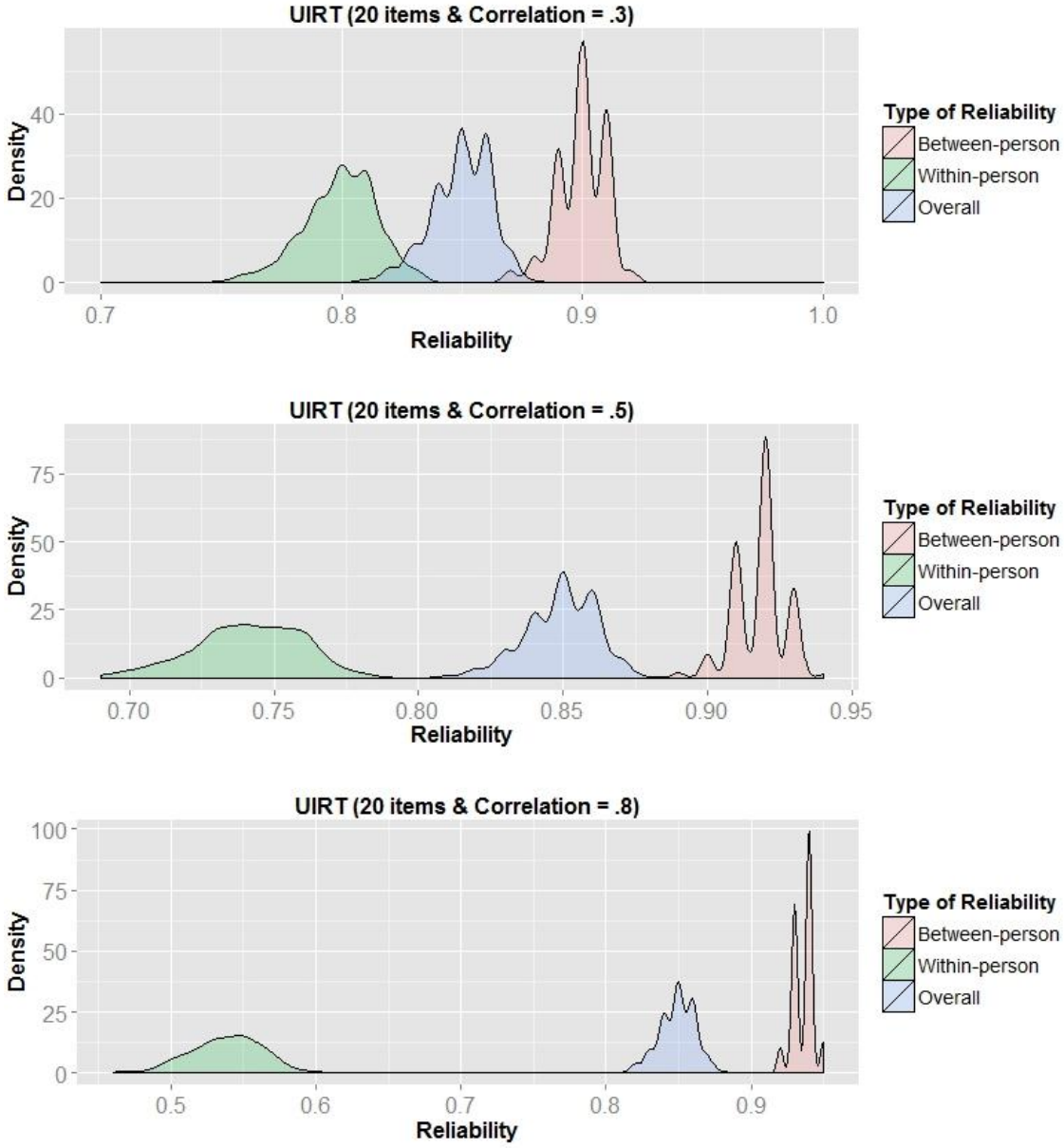

Figure C1.11. Sampling distributions of reliability estimates from 3-dimensional UIRT model with 20 items.

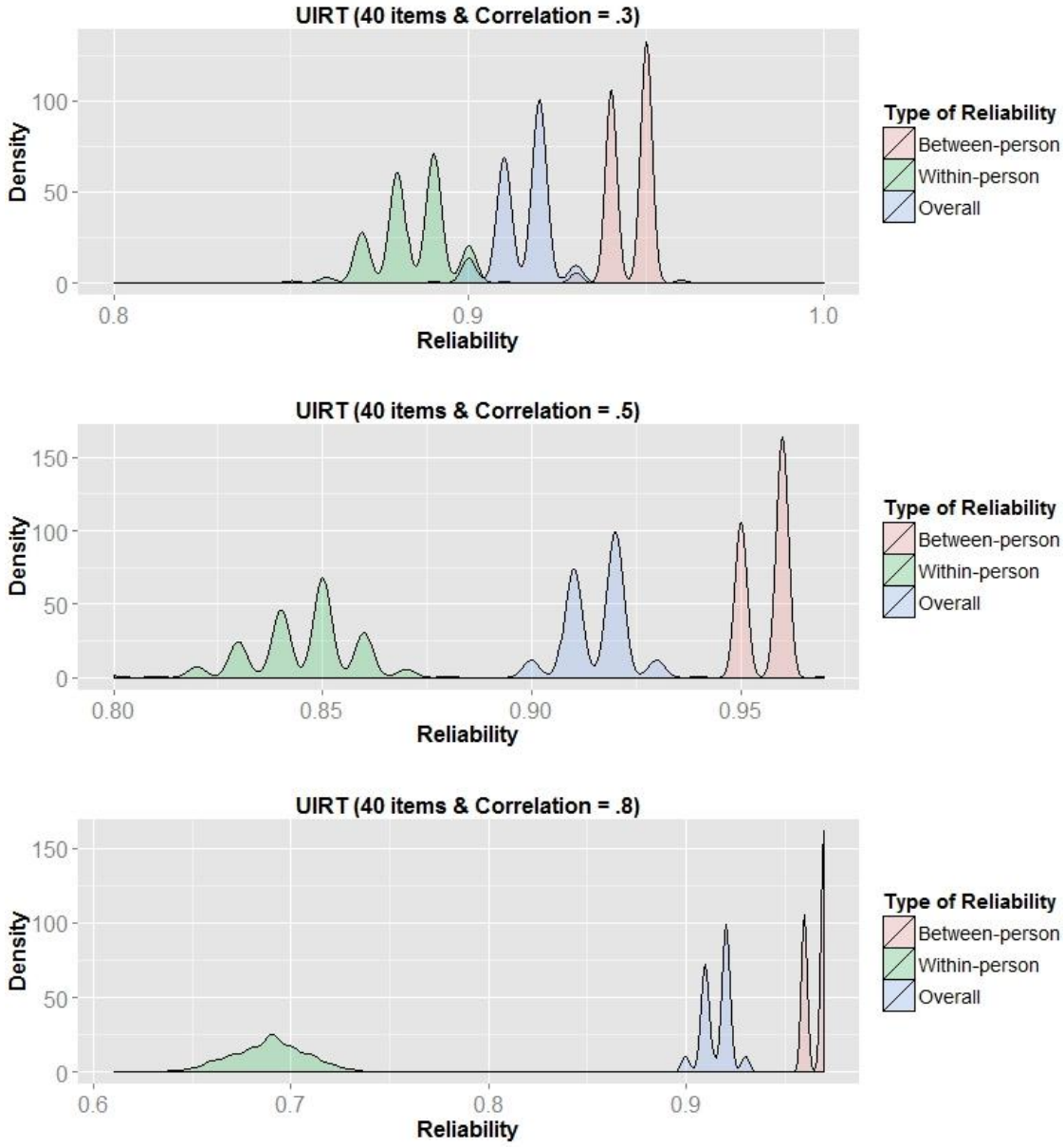

Figure C1.12. Sampling distributions of reliability estimates from 3-dimensional UIRT model with 40 items.

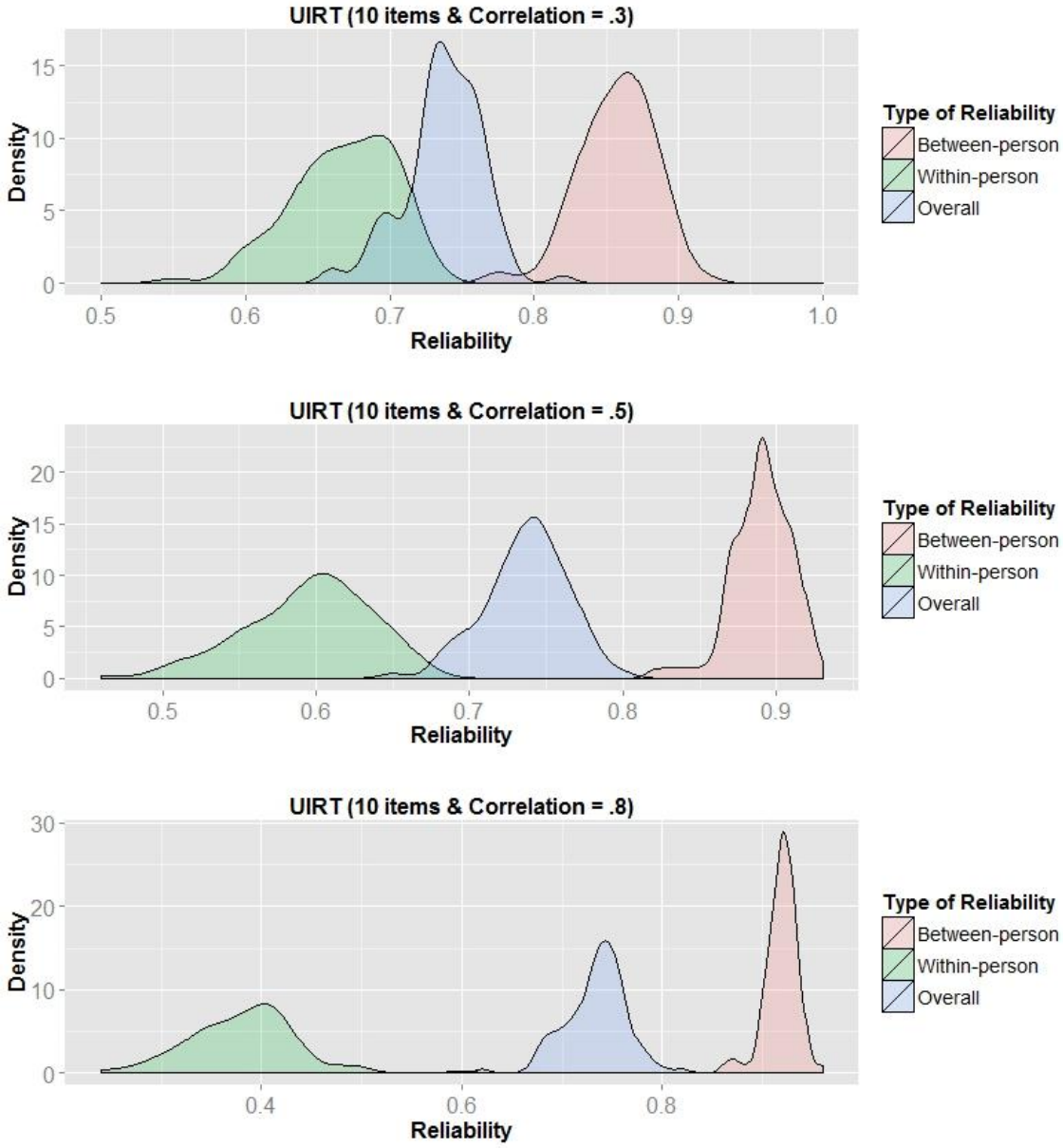

Figure C1.13. Sampling distributions of reliability estimates from 5-dimensional UIRT model with 10 items.

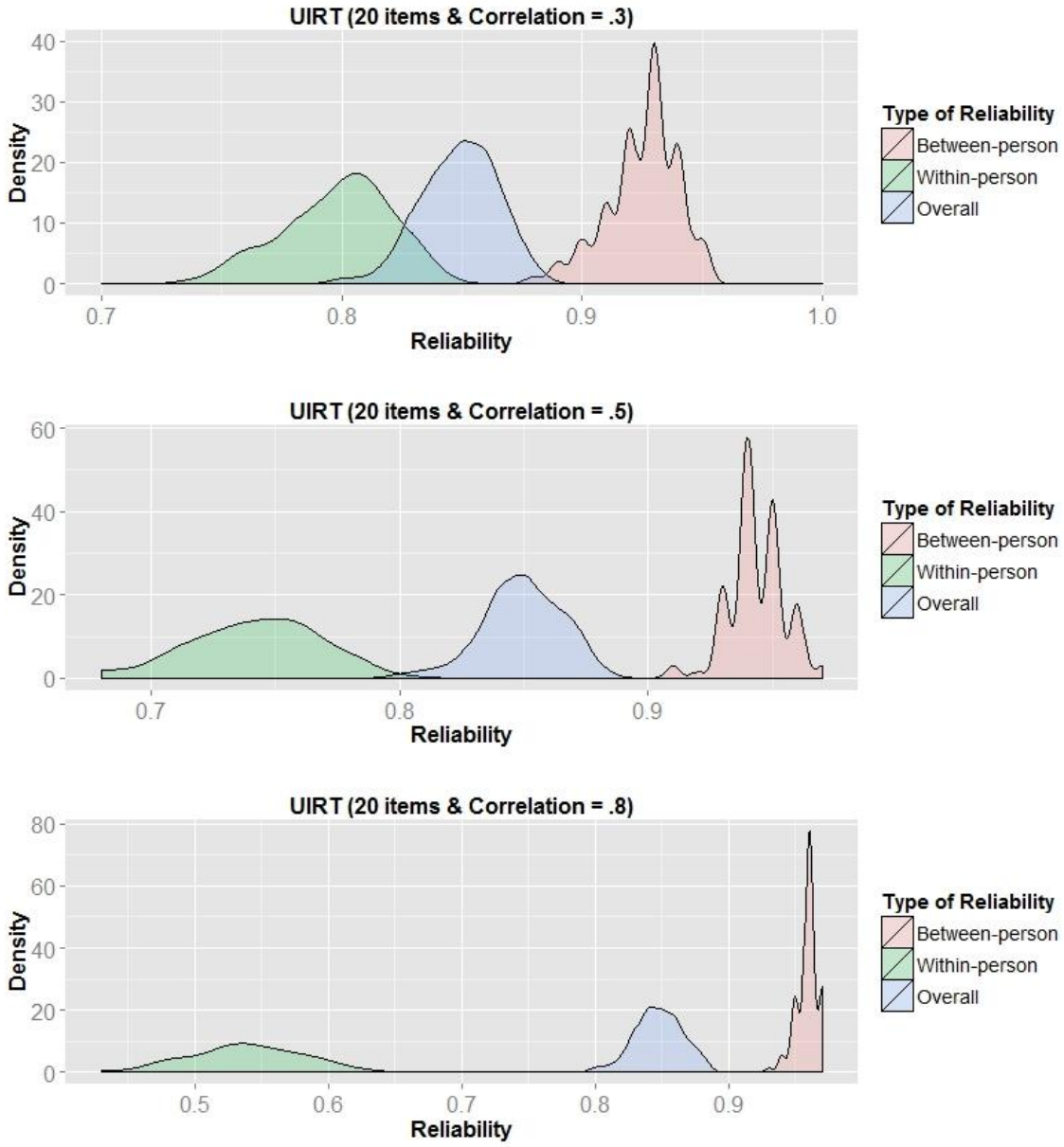

Figure C1.14. Sampling distributions of reliability estimates from 5-dimensional UIRT model with 20 items.

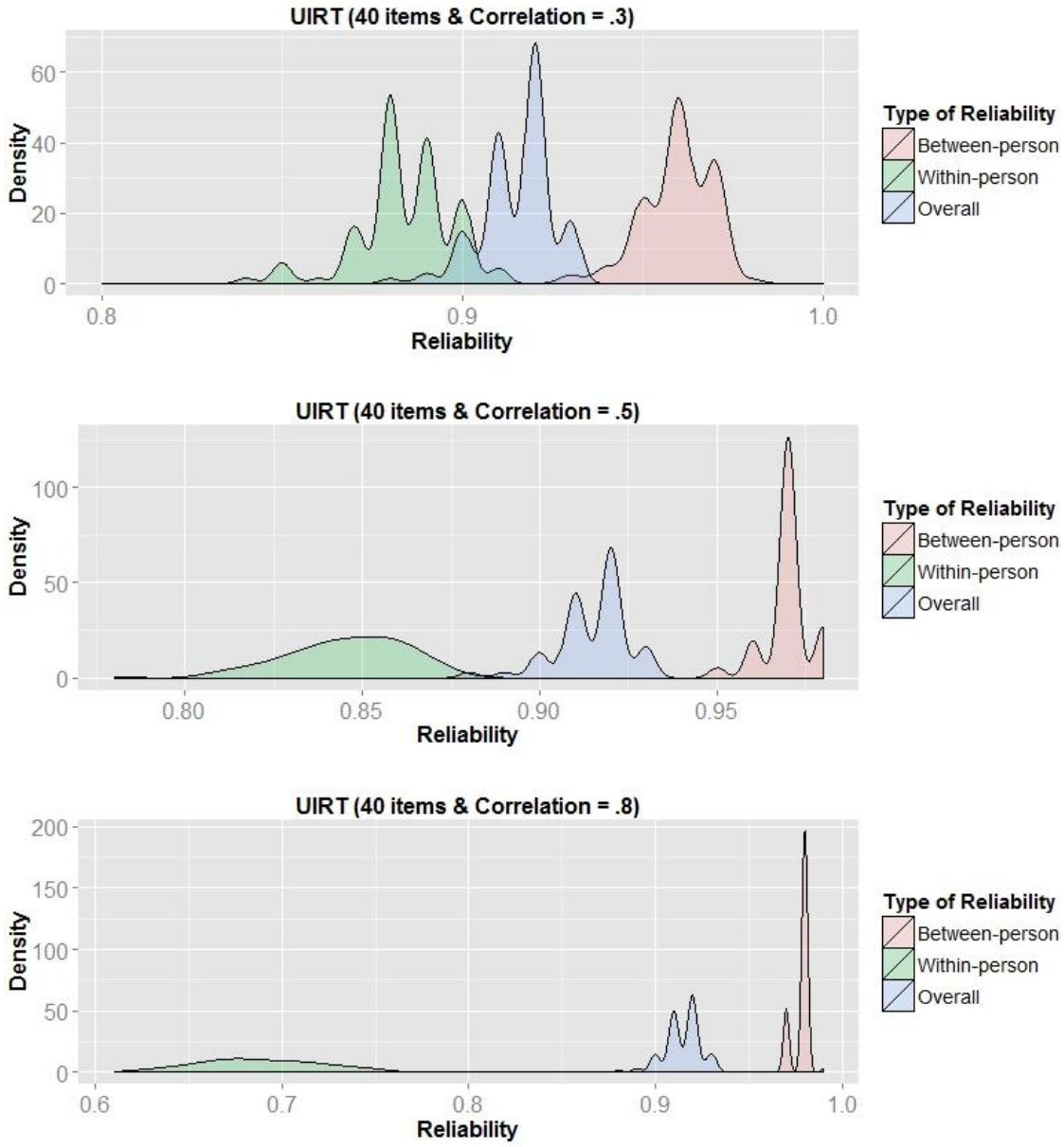

Figure C1.15. Sampling distributions of reliability estimates from 5-dimensional UIRT model with 40 items.

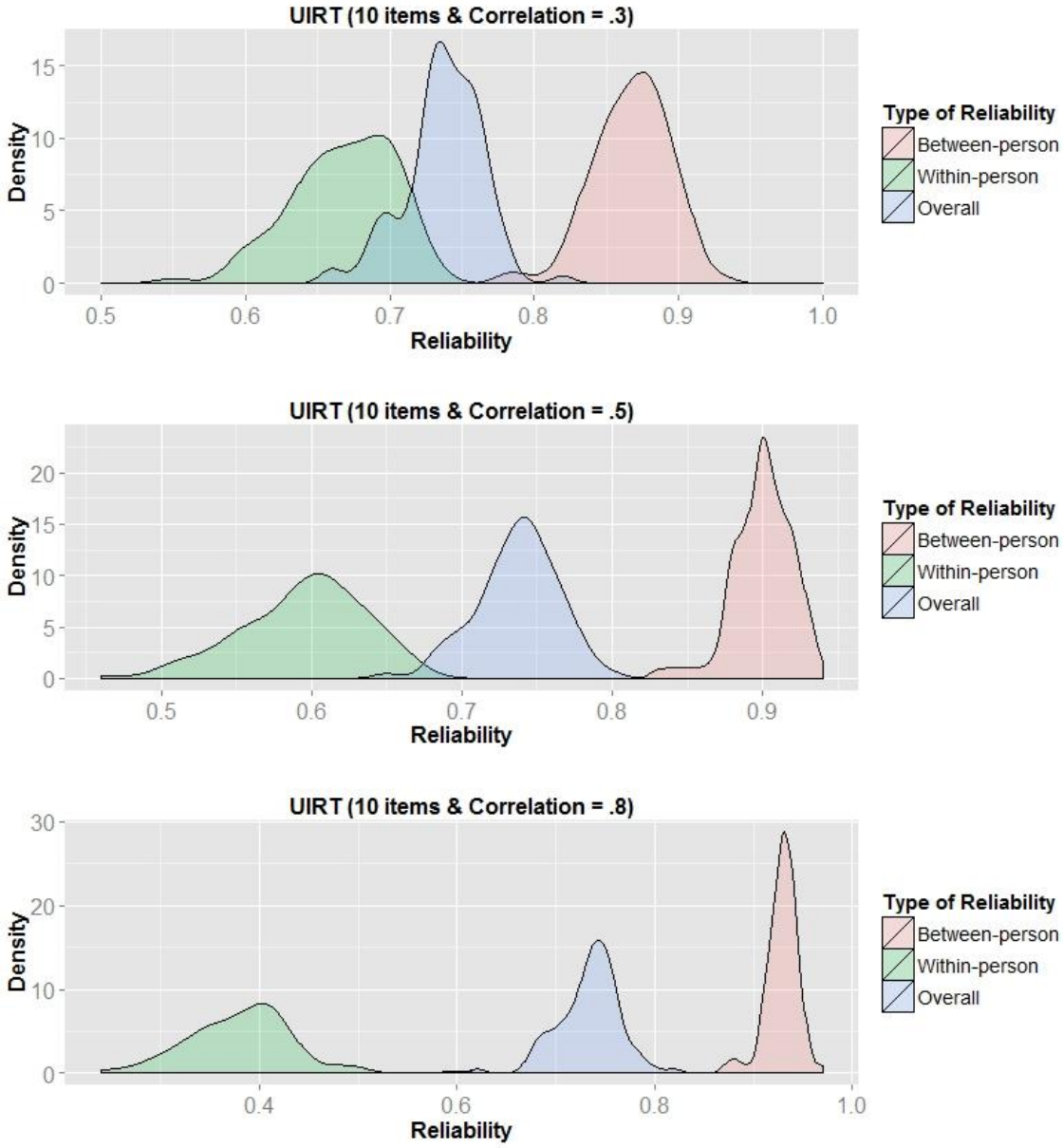

Figure C1.16. Sampling distributions of reliability estimates from 7-dimensional UIRT model with 10 items.

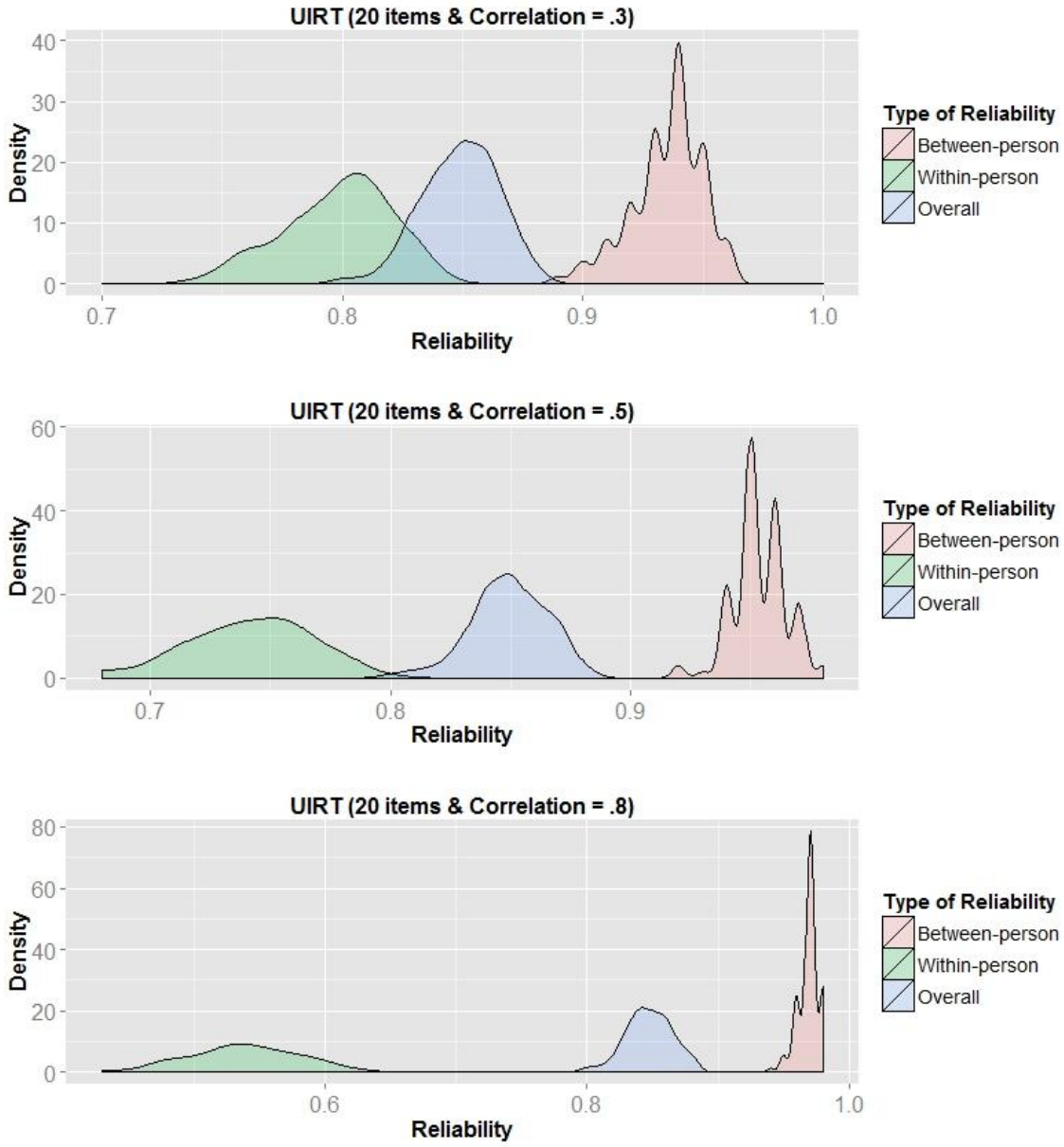

Figure C1.17. Sampling distributions of reliability estimates from 7-dimensional UIRT model with 20 items.

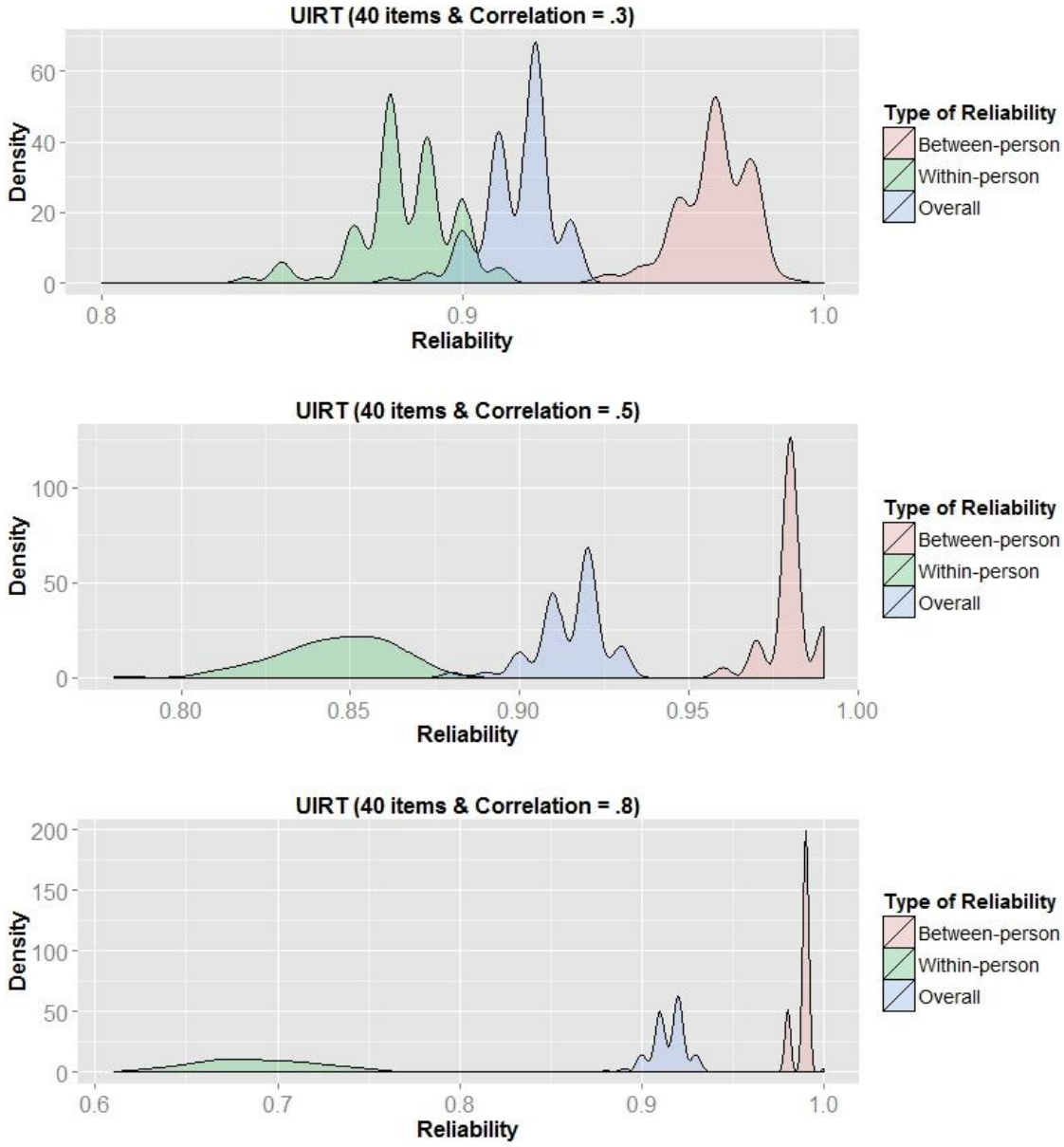

Figure C1.18. Sampling distributions of reliability estimates from 7-dimensional UIRT model with 40 items.