IDENTIFYING DATA CONDITIONS TO ENHANCE SUBSCALE SCORE ACCURACY BASED ON VARIOUS PSYCHOMETRIC MODELS

A Dissertation
Presented to
The Academic Faculty

by

HeaWon Jun

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Psychology

Georgia Institute of Technology
August 2016

IDENTIFYING DATA CONDITIONS TO ENHANCE SUBSCALE SCORE ACCURACY BASED ON VARIOUS PSYCHOMETRIC MODELS

Approved by:

Dr. Susan Embretson, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Charles Parsons
Scheller College of Business
Georgia Institute of Technology

Dr. Richard Catrambone
School of Psychology
Georgia Institute of Technology

Dr. Jonathan Templin
Department of Educational psychology
University of Kansas

Dr. Rick Thomas
School of Psychology
Georgia Institute of Technology

Date Approved: April 25, 2016

ACKNOWLEDGEMENTS

I wish to express my deepest thanks to my advisor, Dr. Susan Embretson, for her constant and insightful guidance, expertise, and encouragement, without whose support this work would not have been possible. I am also grateful to all other committee members, Dr. Richard Catrambone, Dr. Rich Thomas, Dr. Charles Parsons, and Dr. Jonathan Templin for their assistance and suggestions. Most of all, I would like to thank my parents, my husband, and my son for their love and supports.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

As a result of the requirements in the NCLB Act of 2001, subscale score reporting has drawn much attention from educational researchers and practitioners. Subscale score reporting has an important diagnostic value because it can give information about respondents' cognitive strengths and weaknesses in specific content domains. Although several testing programs have reported their results in subscales, there have been many concerns about the reported subscale scores due to their lack of appropriate psychometric quality, especially in reliability. Various subscale scoring methods have been proposed to overcome the lack of reliability (Monaghan, 2006; Haberman, 2008). However, their efficiency in subscale scoring seems to fluctuate under different data conditions. The current study seeks the optimal data conditions for maximizing reliability or accuracy of subscale scores using CTT- and IRT-based methods. Both real-world data and simulation data are used to compute subscale scores, and their accuracies of these estimations (i.e., reliability) are compared. For a real-world data study, response data of a math achievement test from 5,000 eighth grade students in a Midwestern state are used. For the simulation study, response data are generated varying the subscale length, between-subscales correlations, within-subscale correlations, and level of item difficulty. Each data condition has 100 replications.

CHAPTER 1

INTRODUCTION

Most educational tests are designed to rank examinees along a single continuum with regard to the measured construct for the purpose of assessing students' educational progress and deciding whether they qualify for advancement to the next grade or graduation. However, current educational policies require student assessments to provide more information than just a single score, including diagnostic information about students' specific cognitive strengths and weaknesses regarding knowledge or skills. In particular, the No Child Left Behind Act (NCLB) of 2001 demanded that states measure student achievement relative to state standards and report the results to students, parents, educators, and other educational stakeholders, so that the information from the results may be used to plan instruction or learning as well as modify educational programs. The goal was to enable that all U.S. students would ultimately reach the state-mandated achievement goals. Accordingly, testing institutes have devoted themselves to designing appropriate tests that offer diagnostic information or help in finding or developing psychometric models for diagnostic results. Practitioners' interest in and need for diagnostic information naturally attracted researchers' attention, spurring on the research and development of methodologies for diagnostic measurement. From this effort, one of representative research outcomes is the development of cognitive diagnostic assessment models (CDA).

CDA models, combining cognitive psychology with measurement theory, are relatively new psychometric models for measuring respondent's proficiency levels, with regard to skills or knowledge consisting of items. CDA models may provide diagnostic information for a respondent's strength or weakness in content, skill, or knowledge areas. The development of

these CDA models was initially instigated by numerous researchers who have emphasized the role of cognitive psychology in measured constructs (Embretson, 1983; Messick, 1989; Nichols, Chipman, & Brennan, 1995; Mislevy, 1993; Snow & Lohman, 1989, 1993), and spurred by the NCLB Act, resulted in the development of many different CDA psychometric models. Despite the effort to develop CDA models and the advantage (i.e., diagnostic information) that they provide, practitioners have been reluctant to use CDA models for reporting diagnostic results. This is due to a number of factors, including their computational inefficiency relative to parameter estimation (i.e., large number of parameters), their insufficient evidence regarding the psychometric quality of resulting scores (e.g., accuracy of estimation, model-fit, etc.), and the substantially large number of items required for measuring each skill construct.

Other researchers have considered using subscale scores (e.g., number-correct scores, percent-correct scores, IRT estimated domain scores, etc.) as diagnostic scores, which are available from traditional psychometric frameworks, such as classical test theory (CTT) and item response theory (IRT). They believe that these methods can provide relatively easy and simple methods for computing subscale scores. In practice, testing programs such as ETS, ACT, and LSAT reported these types of subscale scores, in order to provide their test users with diagnostic information (Sinhary, Puhan, & Haberman, 2011). However, there has been much disagreement among researchers with the appropriateness of subscale score reporting, their major argument being the lack of reliability or accuracy in subscale scores. Many researchers criticize the fact that the reported subscale scores for the most part lack reliability or precision, and should not be reported.

Not surprisingly, numerous studies have focused on the factors allowing subscale scores to yield better reliability or accuracy. In particular, these studies include the examination of the

impacts of the different lengths of a subscale or the different correlations between subscale scores or between subscale score and total score for reliability. Other studies were on the development of psychometric models, which can yield subscale scores without reducing reliability or estimation precision (Monaghan, 2006). Haberman (2008) believes that adding information from a total score to an observed subscale score can improve the accuracy of subscale score estimation, ensuring high reliability, and suggests a method of weighted averages, combining both an observed score from a subscale and a total score with different weights. Introducing the Objective Performance Index (OPI), Yen (1987) intends to yield more accurate subscale scores by setting each respondent's global trait score θ as the prior distribution in subscale score estimation. Many other subscale scoring methods have been introduced; some of them based on the CTT framework, but others based on IRT. Subscale score estimates from these models show different levels of estimation accuracy, depending on various data structures (e.g., different correlation structure, internal consistency, the number of items, etc).

In the present study, I intend to review various psychometric models that have been suggested for subscale scoring, and compare which models are able to provide better subscale score estimates. In addition, I will examine specific data conditions known to have an impact on the precision of estimation and examine which conditions will enable improvement, based on the addition of other conditions to those already specified. Throughout the paper, the term "subscale score" will be used as a generic term for the diagnostic proficiency score, the domain score, the dimension score, and so on.

What Are Subscale scores?

Subscale scores refer to a test-taker's performance levels on multiple subject areas or on the subscales making up a test. A test may be clearly partitioned into a few subscales, or may be

divided in a way that the test developer believes is appropriate (Sinharay, Puhan, & Haberman, 2011). In the former case, where there are clearly defined subscale sections, it is relatively easy to decide how many subscale scores need to be reported. For example, two subscale scores should be reported in a general ability test that includes two subscales, such as reasoning and working memory scales, one subscale score for reasoning scale and the other for working memory scale. However, in most cases where subscale scores are being considered for reporting, a test is not clearly divided into subsections. For instance, large-scale assessments such as state assessments usually measure broad content areas, skills, or attributes in a subject, but the criterion for dividing an entire test into subscales is rather ambiguous, thus requiring an indicator to determine how many subscale scores should be offered. In these cases, a test blueprint may be used as an indicator for deciding the number of subscales (Haberman, Sinharay, & Puhan, 2006). Because the test blueprint specifies the skills, attributes, or knowledge structure that each item should measure, it is possible to categorize items that measure constructs that are more or less similar into the same or a different group.

Not having clearly defined subscales may indicate that a test may have multiple sets of subscales, based on different standards. For example, educational test blueprints display different hierarchical structures, in which multiple contents levels (e.g., superior levels vs. subordinate levels) may be categorized, permitting different sets of subscale scores. According to a test blueprint from a Midwestern State Department of Education, a math test involves three levels of hierarchy in its structure—Standards, Benchmarks, and Indicators—in which Standards involves Numbers/Computation, Algebra, Geometry, and Probability, these then include two or three benchmarks each. In turn, these Benchmarks include six or seven indicators (i.e., indicator skills) within each. Given these types of test blueprints, a set of subscale scores may be considered for

one selection among Standards, Benchmarks, or Indicators. Besides this test blueprint example, other sets of subscale scores are possible or likely. Specifically, Embretson (2006) identified cognitive components influencing math item-solving rather than content variables as in the test blueprint in the comparison study of alternative models, based on these different subscale structures. Similarly, Jun, Lutz, Morrison, & Embretson (2013) created two different sets of subscales based on cognitive complexity variables and standards-based variables, and compared the fit of models based on both subscale structures. In these studies, subscale scores could be provided based on cognitive components defined by researchers, and not based on content variables.

Multiple ways to constitute subscales seem to be plausible. A test may be partitioned into subscales by the test developer from the beginning. Alternatively, it may be divided based on either test blueprint or as defined by test developers or analysts. Once alternative subscale scores are available, the kind of subscale information given to test users must prioritize the purpose of testing or the test user's interests (DiBello, Roussous, and Stout, 2007).

Importance of Subscale Score Reporting

There is socially increasing demand for subscale scores. U.S. Educational policy (e.g., the NCLB) requires states to implement standards-based assessments and provide descriptive score reports, including diagnostic information aligned with state academic achievement standards. In addition, according to a national survey by Goodman & Huff (2006), most teachers (i.e., 93%) participating in the survey responded that large-scale assessment results should provide diagnostic information, but most of these assessment are not including sufficiently detailed information on specific content or skill domains.

Why are subscale scores important? First, subscale scores provide diagnostic information about the strengths and weaknesses of test takers' performance in specific knowledge, content, or skill domains. The knowledge of students' cognitive strengths and weaknesses in the domains allow teachers to plan the future instruction or adjust their current lessons, so that they effectively intervene and properly address student's academic needs. Next, diagnostic information can help students plan their own learning objectives. Based on the results, students will be more likely to direct their efforts towards their weak subject areas, and dedicate less or similar levels of effort to their strong subject areas. Diagnostic results may also be used when states and educational institutions appraise the effectiveness of their existing curriculum or need to propose modification. Furthermore, subscale scores may be considered a source of supplementary information in school admissions, personal selection, and placement. Monaghan (2006) indicates that subscale scores may be a valuable resource for the admission or the selection purposes to differentiate between candidates with identical total scores, thus when an additional criterion for selection of appropriate applicants is required. Besides, Haladyna and Kramer (2004) mentioned that subscale scores may also be useful as a tool for evaluating their training programs by comparing students' performance before and after the training program.

Although subscale scores are much needed and important for the advantages that they provide, it may not be easy to yield reliable and accurate subscale scores. The following section delineates several psychometric requirements that must be present in order for subscale scores to be reported.

Psychometric Requirements of Subscale Scores

Although subscale scores serve multiple purposes, not all are not permitted to be reported. Certain psychometrical criteria must be met, in order for subscores to be reported. The

most important criteria are their reliability and validity. Standard 5.12 of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014) mentions the features of reliability and validity that assessment results must maintain, stating that "Scores should not be reported for individuals unless the validity, the comparability, and the reliability of such scores have been established". This requirement is applied to scores from subscales, as well as to the total score. Specifically, Standard 2.1 states that the reliability of subscale scores have to be given with that of the total score, highlighting the reliability that subscale scores must meet. Also, Standard 1.12 also states that subscale scores from different domains should be interpreted with relevant evidence and the rationale to support the interpretation, underlining that scores from different subscales must provide valid evidence towards the constructs being measured. According to the *Standards*, reliability refers to "the consistency of measurement when the testing procedure is repeated on a population of individuals or groups". Subscale scores are defined as accurate and reliable when scores from multiple administrations of a subscale are consistent. Validity here refers to "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests". After ensuring that a subscale score measures the things that it intends to measure, the score would lead to meaningful interpretations. In other words, information from subscale scores would allow one to accurately assess a test taker's ability on attributes, once sufficient evidence of validity has been established.

Haberman (2008) claims that subscale scores must provide an added-value over the total test score. It mentions that subscale scores must provide additional or distinct information (i.e., added-value) over the total test score. Otherwise, if subscale scores do not provide information that is distinct from the total test score, there would be no reasons to report subscale scores,

along with the total test score. The distinctiveness of subscale scores seems to be related with the concept of validity in certain contexts. Suppose that we have a subscale measuring a set of constructs that are different from the other subscales. Subscales measuring different sets of constructs would lead to irrelevant subscale item responses, unless constructs are too closely correlated or too similar. The irrelevant item responses in different subscales would result in their decreased relevancy within the total score, probably increasing the added-value by subscale score over the total score. In other words, when subscales are assumed to measure different constructs, scores from a subscale will be probably irrelevant to those from the other subscales, having low or moderate correlations between any two subscales scores. On the other hand, if scores in any two different subscales are too similar, probably because they measure constructs that are too similar or equivalent, the correlations between scores of two different subscales will be quite high, which will, in turn, not provide any additional or distinct information over other subscale scores or the total test score.

Factors Impacting the Psychometric Quality of Various Subscale Scores

Various subscale scoring methods that provide subscale scores have been proposed. The simplest subscale scoring method is summed scores (i.e., number-correct scores) that are obtained by summing the number of correct responses for all items in a subscale. Although it is the easiest way to obtain subscale scores, the resulting subscale scores are reported to have greatly reduced reliability compared to the total test scores, mostly due to the short test length in a subscale. However, there are many other subscale scoring methods that yield more reliable and precise subscale scores in the shorten length. For example, Kelly's (1947) regressed subscale score, Holland and Hoskens' (2003) regressed subscale score, and Haberman's (2008) weighted average methods, which belong to the CTT-based models, are known as methods that increase

the reliability of subscale scores by approximating the true subscale scores, using different types of observed scores. Yen's (1986) objective performance index (OPI), Wainer et al.'s (2001) augmented scoring, and multidimensional item response theory (MIRT) models, which belong to the IRT-based models, are alternative subscale scoring methods that estimate subscale scores, without sacrificing the accuracy of subscale estimation under the condition which raw subscale scores are unreliable. Moreover, all types of CDA models can provide subscale scores.

There have been several studies comparing subscale scoring methods with regard to subscale score reliability or accuracy. Different methods improved subscale score reliability or accuracy in different degrees. Dwyer, Boughton, Yao, Steffen, and Lewis (2006) compared raw subscale scores, Yen's OPI subscale scores, and Wainer et al.'s augmented subscale scores in terms of their accuracy, and found that Wainer et al.'s augmented subscale scores generally provided the most reliable subscale score estimates, which were comparable to the MIRT model. Haberman and Sinharay (2010) examined the added-value of subscale scores from the MIRT model and several CTT-based models, and argued that subscale scores based on the MIRT model were more accurate than those from the CTT-based model, although their accuracy did not greatly differ. At most, the degree to which reliability or accuracy is enhanced differed under different data structures. Major factors known as influencing reliability or accuracy include subscale length, sample size, and between-subscales correlations.

First, subscale scores have different levels of reliability depending on the subscale length. Boughton, Yao, and Lewis (2006) compared the impact of the subscale length, varying the number of items contributing to each subscale from three to eighteen, and identified that as the number of subscale items increases, the accuracy of subscale scores improves. Sinharay et al. (2011) illustrated some research examples, supporting that subscales consisting of sufficient

number of items can provide reliable or accurate subscale scores. It makes perfect sense that one may obtain more reliable estimates from a greater number of items, because one may obtain more information from a larger number of item responses than from a small number of item responses. Nevertheless, considering that testing times are limited, it is practically impossible to achieve maximum reliability by calibrating the test length. Note that different subscale scoring methods may be more or less sensitive or adjustable with respect to the length of a subscale.

Second, subscale scores are accurate in their estimation to different degrees, depending on the sample size on which subscale score estimation was based. Yao and Boughton (2007) compared three sample groups of 1,000, 3,000, and 5,000 to examine how changes in sample sizes impact the estimation accuracy of subscale scores. They found that the increase in accuracy of subscale score parameter estimation is much greater when the sample size increases from 1,000 to 3,000 rather than from 3,000 to 5,000, and concluded that a sample size of approximately 3,000 was large enough to obtain accurate subscale scores.

Third, the different size of correlations between subscale scores could cause differences in subscale score reliability or accuracy. Yao and Boughton (2007) compared the accuracy of Yen's OPI values and MIRT dimension scores under four different between-subscales correlation conditions of r = 0.0, 0.3, 0.5, 0.7, and 0.9. The results indicated that subscale scores from the OPI method were as accurate as MIRT dimension scores when the between-subscales correlations are as high as 0.9. It seems to be reasonable, if once considers the point that the OPI method borrows information from the total score. On the other hand, where between-subscales correlations were between 0.0 and 0.5, OPI produced less accurate estimates with more errors than the MIRT models. De la Torre, Song, and Hong (2011) compared four IRT subscale scoring methods, MIRT, augmented scoring, higher order IRT, and OPI, in different test length, different

number of subscales, and between-subscales correlation conditions. The results indicated that the MIRT, augmented scoring, and HO-IRT methods yielded similar results, performing better than OPI. The more accurate estimates were obtained, as there are a greater number of subscales in the test, and scores from the subscale are highly correlated.

Different subscale scoring methods use somewhat distinct information to estimate more accurate or reliable scores, which may be from other subscale scores or the total test score. If a subscale scoring method borrows information from the total test or other subscales, the method will be able to improve subscale score reliability more effectively when between-subscales correlations are high. Skorupski and Carvajal (2010) argued that all augmentation approaches to subscale scoring, borrowing some information to increase subscale score accuracy, lead to the improvement of subscale score reliability, and the amount of increased reliability was greater, especially when the between-subscales correlations are high. However, Sinharay (2010) claims that in order for subscale scores to contain information that is distinct from other subscale scores or from the total test score, correlations between subscale scores should be less than a specified level (i.e., $r = 0.85$). The results show that correlations among subscale scores that are too high may cause validity issues regarding the measured constructs, because highly correlated subscale scores may be interpreted as evidence that the subscales are measuring the same constructs. That is, information from one subscale may be the duplicate of information found in the total test or other subscales.

Other possible factors affecting subscale score reliability may exist, though these have not been discovered in the previous studies. The purpose of this study will involve identifying new factors influencing subscale score accuracy.

The Purpose of This Study

The purpose of this study is to comprehensively identify data circumstances under which the various methods of scoring subscales will have the most accurate estimates of true subscale scores. Alternative subscale scoring methods are compared under all data conditions being considered. The various data conditions include: a) the levels of within-subscale correlations (i.e., internal consistency of subscale scores), b) item difficulty, c) subscale length, and d) between-subscales correlation. In turn, a total of seven alternative subscale scoring methods are employed to compare results: four different CTT-based subscale scoring methods including a) raw subscale scores, b) Kelly's regressed subscale scores, c) Holland and Hoskens' regressed subscale scores, and d) Haberman's weighted average method, and three different IRT-based subscale scoring methods including a) unidimensional 2PL model, b) OPI, and c) a multidimensional 2PL model. Subscale scores from these methods are discussed relative to their accuracy of subscore estimation. Research hypotheses follow.

Research Hypotheses

1) The four conditions are expected to impact subscale score reliability or accuracy, as computed from observed subscale scores, as follows:
    a. The number of items in a subscale will influence the degree of subscale score reliability. Specifically, as the number of items in each subscale increases, subscale score reliability is expected to increase.
    b. The size of correlation between subscales will impact the reliability of subscale scores differently. High correlations between subscales are expected to result in lower reliability in raw subscale scores than in the total score. However, moderate or low correlations between subscales seem to lead higher reliability of raw subscale scores than of the total score. Using CTT-based and IRT-based subscale

scoring methods, the amount of improved reliability will differ based on the methods.

c. The degree of within-subscale correlations (i.e., Subscale Consistency) will impact the reliability of subscale scores differently. As the within-subscale correlation will be high, the resulting subscale score reliability is expected to increase.

d. Subscale score reliability may differ in different test types: ability vs. achievement tests, in which subscales consist of different levels of difficulty.

2) The four conditions are expected to impact the relationship of observed subscale scores to true subscale scores as follows:

a. The various subscale scoring methods are expected to interact with the accuracy of predicting true subscale scores, depending on the specified data conditions. For example, the use of total score to approximate true score (i.e., Holland & Hoskens) should be effective only under conditions in which the subscale score correlations are high and the subscale observed score is based on few items with low internal consistency.

The data conditions described above may interactively influence the subscale score reliability, depending on alternative subscale scoring methods.

Chapter Overview

Large-scale assessments that are widely used for admission, selection, and evaluation often measure broad areas of content or skill domains, placing people on a single ability continuum relative to the measured construct. Recently, these types of assessments are considered useful in providing additional information regarding examinees' cognitive strengths

and weaknesses in specific subdomains. The current study seeks data conditions under which a test that is not initially designed for diagnosis may provide diagnostic results with appropriate reliability or the added-value. The criteria of whether these can provide valuable information will be mainly based on the reliability of subscale score estimates.

The current paper consists of four chapters. The following chapter (Chapter 2) introduces various subscale scoring methods based on the different measurement scaling models (CTT, IRT, and CDA models), and delineates reliability and validity measures as criteria for evaluating the psychometric quality of subscale scores. Chapter 3 includes the research methods and results from the real-world data study. The real-world study compares the psychometric quality of subscale scores from real data. Chapter 4 includes research designs (or methods) describing simulation procedures for the simulation data study and discusses the results of subscale scores obtained under various data conditions. Lastly, Chapter 4 includes a discussion in which results from the real-world and simulation data will be summarized and the significance and implications of this study will be discussed.

CHAPTER 2

LITERATURE REVIEW

This chapter provides a broad overview of the several subscale scoring methods that are provided under CTT, IRT, and CDA frameworks. The basic and important concepts under these frameworks are presented. Then, how the methods influence reliability and score accuracy can be achieved is explained in detail.

It should be noted that the current studies examine a subset of these methods: 1) Raw subscale scoring, 2) Kelley's method, 3) Holland-Hoskens' method, 4) Haberman's method, 5) Unidimensional IRT scoring, 6) Objective Performance Index (OPI) scoring, and 7) two-parameter logistic Multidimensional IRT (MIRT-2PL) scoring.

CTT-based Subscale Scores

In CTT, the most intuitive method of obtaining subscale scores is to compute the summation of item scores from a subset of items (i.e., raw subscale scoring). Although this method provides quite clear and simple rationale for subscale scoring, the resulting subscale scores suffer from the lack of reliability in the short length of subscale. Several CTT-based subscale scoring methods have been proposed to solve the lack of reliability in raw subscale scores and improve the accuracy of subscale scores. These methods employ the linear regression model to approximate the true subscale score on one of the following types of predictors: a) the observed subscale score, b) the observed total score, and c) the weighted combination of the observed subscale score and the observed total score. In the methods, the true subscale scores are not directly observable, and thus, must be inferred through the relationship with pertinent observable subscale scores and subscale score errors.

The current section overviews some basic CTT concepts that are prerequisites for understanding CTT subscale scoring methods, and describes three different types of CTT-based subscale scoring methods. This section is followed by a detailed description of mean square error (MSE) and proportional reduction in mean square error (PRMSE), which are proposed as criterion of measuring the reliability and the added-value of subscale scores in CTT.

Basic Concepts in CTT

CTT specifies the relationship among variables (i.e., observable variables, unobservable variables, and error) under specific assumptions. There are five main assumptions that CTT adopts. The first assumption is below:

$$X = T + E, \tag{2.1}$$

where X, T, and E are, respectively, the observed, the true, and the error scores. In the assumption, the observed score, X, is assumed to be the sum of the true score, T, and the error score, E. X is a score from each testing when the same test is repeatedly given to an examinee, and the true score is a fixed score that does not change over repeated testings. E represents the difference between the observed score and the true score.

The second assumption is as follows:

$$E(X) = T. \tag{2.2}$$

In CTT, the true score, T, is the theoretical mean of each person's scores based on multiple independent testings on the same test. That is, the true score, T, can be achieved by the expected value of the observed scores over repeated testings, $E(X)$.

The third assumption is as follows:

$$\sigma_{ET} = 0. \qquad (2.3)$$

That is, the error scores, Es, and the true scores, Ts, from all examinees in a population are assumed to be uncorrelated.

The forth assumption is as follows:

$$\rho_{E_1E_2} = 0. \qquad (2.4)$$

Supposing that E_1 and E_2 are the error scores for two different tests from all examinees in a population, the error scores on the tests are assumed to be uncorrelated.

The fifth assumption is as follows:

$$\rho_{E_1T_2} = 0. \qquad (2.5)$$

when E_1 and T_2 are, respectively, the error score for Test 1 and the true score for Test 2 from all examinees in a population, the error scores on Test 1, E_1, are assumed to be uncorrelated with true scores on Test 2, T_2. Note that equations (2.1) and (2.2) are based on repeated testings of an examinee, but the equations (2.3), (2.4), and (2.5) are based on all examinees in a population.

The major assumptions above postulate relationships among the observed, the true, and the error scores. In the assumptions, only Xs are observable. Because true scores, Ts, and error scores, Es, are unobservable and theoretical variables, they should be indirectly inferred. Also, when the assumptions hold well enough, several other inferences among variables are derived. The following section includes major five equations that can be driven when assumptions are reasonably correct.

First, the expected value of the observed scores, X, and the expected value of the true scores, T, are the same, which is shown below:

$$E(X) = E(T). \quad (2.6)$$

According to the equation (2.1), $E(X) = E(T + E) = E(T) + E(E)$. Because $E(E) = 0$, $E(X) = E(T)$.

Second, the expected value of the products of the error scores and the true scores from all examinees in a population is zero, as shown below:

$$E(ET) = 0. \quad (2.7)$$

Because $\sigma_{AB} = E(AB) - E(A)E(B)$ and $E(E) = 0$, $E(ET) = \sigma_{ET} + E(E)E(T) = \sigma_{ET}$. From the assumption equation (2.3), $\sigma_{ET} = 0$. Thus, $E(ET) = 0$.

Third, the variance of the observed scores, Xs, is equal to the sum of variance of the true scores, Ts, and the variance of the error scores, Es, as below:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2. \quad (2.8)$$

From the assumption equation (2.1), $\sigma_X^2 = \sigma_{T+E}^2 = \sigma_T^2 + \sigma_E^2 + 2\sigma_{TE}$. Because $\sigma_{TE} = 0$ from the equation (2.3), $\sigma_X^2 = \sigma_T^2 + \sigma_E^2$.

Fourth, the squared correlation of the observed scores and the true scores across examinees in a population is the ratio of true score variance to observed score variance, as shown below:

$$\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}. \quad (2.9)$$

This formula can be driven as follows:

$$
\begin{aligned}
\rho_{XT}^2 &= \left[\frac{\sigma_{XT}}{\sigma_X \sigma_T}\right]^2 \\
&= \frac{[E(XT) - E(X)E(T)]^2}{\sigma_X^2 \sigma_T^2} \\
&= \frac{\left[E\left(T(T+E)\right) - E(X)E(T)\right]^2}{\sigma_X^2 \sigma_T^2} \\
&= \frac{[E(T^2) - E(T)^2]^2}{\sigma_X^2 \sigma_T^2} \\
&= \frac{\left[E(T^2) + E(TE) - E(X)E(T)\right]^2}{\sigma_X^2 \sigma_T^2} \\
&= \frac{\left[E(T^2) - E(T)E(T)\right]^2}{\sigma_X^2 \sigma_T^2} \\
&= \frac{[\sigma_T^2]^2}{\sigma_X^2 \sigma_T^2} \\
&= \frac{\sigma_T^2}{\sigma_X^2}.
\end{aligned}
\tag{2.10}
$$

The squared correlation among variables from the linear relationship indicates the maximum proportion of variance of the dependent variable that is predictable from the independent variable. When it comes to CTT, in which the linear relationship between the observed scores, Xs and the true scores, Ts, is assumed, the squared correlation among Xs and Ts is equal to the ratio of true score variance to observed score variance, which corresponds to the definition of reliability in CTT.

Fifth, the correlation between observed scores from two parallel tests is equal to the squared correlation between observed scores and true scores, which is also identical to the ratio of true score variance to observed score variance, as shown below:

$$
\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT}^2. \tag{2.11}
$$

In this equation,

$$\rho_{XX'} = \frac{\sigma_{XX'}}{\sigma_X \sigma_{X'}}$$

$$= \frac{\sigma_{(T+E)(T'+E')}}{\sigma_X^2}$$

$$= \frac{\sigma_{TT'} + \sigma_{TE'} + \sigma_{T'E} + \sigma_{EE'}}{\sigma_X^2} \qquad (2.12)$$

$$= \frac{\sigma_{TT'}}{\sigma_X^2}$$

$$= \frac{\sigma_T^2}{\sigma_X^2}$$

That is, the correlation between scores from two parallel tests is equal to the reliability of scores.

The basic notions from the CTT framework, described above, are often useful in computing subscale scores, with additional assumptions that each subscale scoring method takes. However, note that these notions are suitably used only when the CTT assumptions reasonably hold. For further details on CTT, see Lord and Novick (1968) and Allen and Yen (2002).

CTT-based Subscale Scoring Methods

This section includes the rationale of how three CTT subscale scoring methods, Kelley's, Holland-Hoskens', and Haberman's methods, estimate subscale scores and the computation procedure of how they obtain subscale scores with details.

Kelley's Regressed Subscale Scores

Kelley (1927, 1947) suggested the linear regression model for approximating the true subscale score by the observed subscale score. The general linear regression equation of the predicted variable *Y* on the predictor variable *X* is expressed as follows:

$$\hat{Y}_i = B_{Y.X}(X_i - \bar{X}) + \bar{Y} \qquad (2.13)$$

where X_i is the observed score of examinee i, $B_{Y.X}$ is the regression coefficients, and $\bar{X}$ and $\bar{Y}$ are, respectively, the means of predicting and predicted scores. $B_{Y.X}$ is determined based on a standard criterion of finding the best prediction line that minimizes the sum of the squared prediction errors. Based on the standard criterion, the regression coefficient, $B_{Y.X}$, can be estimated through the following equation:

$$\widehat{B_{Y.X}} = r_{XY}\left(\frac{S_Y}{S_X}\right), \qquad (2.14)$$

where r_{XY} is the correlation coefficient between X and Y, and S_Y and S_X are the standard deviations of X and Y, respectively.

When it comes to the prediction of subscale scores, in which the observed subscale score, S_x, is regressed to predict the true subscale score, S_t, the regression equation can be written as follows:

$$S_t = E(S_t) + r_{S_x S_t}\left(\frac{s_{S_t}}{s_{S_x}}\right)[S_x - E(S_x)], \qquad (2.15)$$

where $E(S_t)$ = the expected value of the true subscale scores across examinees,

$E(S_x)$ = the expected value of the observed subscale scores across examinees,

s_{S_t} = the standard deviations of the true subscale scores,

s_{S_x} = the standard deviations of the observed subscale scores, and

$\rho_{S_x S_t}$ = the correlation between the true subscale scores and the observed subscale scores.

Although true subscale scores in the equation are unknown, the terms pertinent to the true subscale scores are attainable, considering specific relationships between the observed and the true scores from CTT assumptions. In the equation (2.15), S_x is simply an examinee's observed

score in a subscale, and $E(S_x)$ is obtainable from the sample score mean in the subscale. Also, similar to the equation (2.6), $E(S_x)$ would be equal to $E(S_t)$. Thus, the expected value of observed subscale scores across examinees can be substituted for that of true subscale scores. Then, correlations between the true and the observed subscale scores in a sample can be rewritten as below:

$$r_{S_t S_x} = \frac{s_{S_t}}{s_{S_x}}, \qquad (2.16)$$

because $\rho_{XT}^2 = \sigma_T^2/\sigma_X^2$ from the equation (2.9). Note that the terms σ and s indicate the standard deviations, and the terms ρ and r indicate correlation coefficients. However, σ and ρ are terms defined in the population, and the others terms are defined in a sample. Thus, the regression coefficient is $r_{S_x S_t}(s_{S_t}/s_{S_x}) = s_{S_t}^2/s_{S_x}^2$, which corresponds to the reliability of subscale scores.

Based on the relationships among the observed and the true subscale scores, the equation (2.15) can be simply rewritten as follows:

$$S_t = E(S_x) + \frac{s_{S_t}^2}{s_{S_x}^2}[S_x - E(S_x)]. \qquad (2.17)$$

As mentioned earlier, $E(S_x)$ is obtainable through the sample mean of the corresponding subscale, and the regression coefficient, $s_{S_t}^2/s_{S_x}^2$, is available through KR-20 or Cronbach α from items in the subscale.

Holland and Hoskens' Regressed Subscale Scores

Holland and Hoskens (2003) suggested a subscale scoring method of approximating the true subscale score by the total test score. In some situations, the linear prediction based on the observed total score leads to better prediction based on the observed raw subscale score. For

example, the true subscale score based on the observed total score would have more accuracy with smaller prediction error than those based on the observed subscale score when the true subscale score and the true total score are highly correlated. However, this method requires caution when it is considered to use with data whose true subscale score and true total score are scarcely correlated.

The linear regression equation, in which the observed total score, Y_x, is regressed to predict the true subscale score, S_t, can be written as follows:

$$S_t = E(S_t) + r_{Y_x S_t}(\frac{s_{S_t}}{s_{Y_x}})[Y_x - E(Y_x)], \qquad (2.18)$$

where $E(S_t)$ = the expected value of the true subscale scores across examinees,

$E(Y_x)$ = the expected value of the observed total scores across examinees,

s_{S_t}= the standard deviation of the true subscale scores,

s_{Y_x}= the standard deviation of the observed total scores, and

$r_{Y_x S_t}$= the correlation between the observed total scores and the true subscale scores.

Y_x is simply an examinee's observed total score in a whole test, and $E(Y_x)$ is obtainable from the sample score mean in the whole test. Similar to the equation (2.6), $E(S_x)$ would be equal to $E(S_t)$. Thus, the expected value of observed subscale scores across examinees is substituted for that of true subscale scores. Also, correlations between the true subscores and the observed total scores in a sample can be expressed as $r_{S_t Y_x} = s_{S_t}/s_{Y_x}$. Thus, the linear regression equation (2.18) can be rewritten as below:

$$S_t = E(S_x) + \frac{s^2_{S_t}}{s^2_{Y_x}}[Y_x - E(Y_x)]. \qquad (2.19)$$

$E(S_x)$ and $E(Y_x)$ are available by the mean of the observed raw subscale scores and the total test scores across examinees. According to Lord & Novick (1968), $r_{Y_xS_t}$ (i.e., $s^2_{S_t}/s^2_{Y_x}$) is defined as the product of a) the correlation between the observed and true total test scores, $r_{Y_xY_t}$, and b) the correlation between the true subscale scores and the true total test scores, $r_{S_tY_t}$. That is, $r_{Y_xS_t} = r_{Y_xY_t} \cdot r_{S_tY_t}$. In the equation, $r_{Y_xY_t}$ is the square root of reliability of the total test score because $r_{Y_xY_t} = s_{Y_t}/s_{Y_x}$. In turn, $r_{S_tY_t}$ can be defined as follows (see Lord & Novick, 1968):

$$r_{S_tY_t} = \frac{r_{S_xY_x}}{r_{S_xS_t} \cdot r_{Y_xY_t}} - \frac{s^2_{s_e}}{s_{S_t} s_{Y_x}}. \qquad (2.20)$$

Together, the linear regression equation (2.19) can be rewritten as below:

$$S_t = E(S_x) + [r_{Y_xY_t} \cdot \left(\frac{r_{S_xY_x}}{r_{S_xS_t} \cdot r_{Y_xY_t}} - \frac{s^2_{s_e}}{s_{S_t} s_{Y_x}}\right)]^2 [Y_x - E(Y_x)]. \qquad (2.21)$$

For estimating subscale scores, $E(S_x)$, $E(Y_x)$, $r_{Y_xY_t}$, $r_{S_xY_x}$, $r_{S_xS_t}$, $s^2_{s_e}$, s_{S_t} and s_{Y_x} need to be known. $E(S_x)$ and $E(Y_x)$ are, respectively, obtainable by computing the sample score mean in the subscale and the total test. Because $r_{Y_xY_t}$ and $r_{S_xS_t}$, respectively, equal the square root of score reliability in the total test and the subscale, $r_{Y_xY_t}$ and $r_{S_xS_t}$ are available from KR-20 or Cronbach α. Then, $r_{S_xY_x}$ is simply correlation between the observed subscale scores and the observed total scores which is available from Pearson correlation coefficient, and s_{S_t} and s_{Y_x} are, respectively, the standard deviations of the true subscale scores and the observed total test scores in a sample. Although s_{S_t} is not directly computable from subscale scores, it is available from $\sqrt{s^2_{S_x} - s^2_{s_e}}$ (see the equation (2.8)). $s^2_{s_e}$ represents the standard error of measurement of the observed subscale score, which is defined as $s_x\sqrt{1 - r^2_{xt}}$.

Haberman's Regressed Subscale Scores: Weighted Average Method

Many educational tests have subscales that have moderately high correlations. Thus, using some information from the other subscale scores as well as the corresponding subscale for estimating the true subscale score may improve the accuracy of the true subscale score estimation. Haberman (2008) suggested to use the jointed information of the observed total score and the corresponding observed subscale score for improving the accuracy of subscale prediction, in its estimation. Specifically, Haberman's (2008) weighted average method approximates the true subscale score by the weighted combination of the corresponding observed subscale score and the observed total score.

The multiple linear regression equation of the true subscale score, S_t, on the observed subscale score, S_x, and the observed total score, Y_x, is shown below:

$$S_t = E(S_t) + \beta_{Y_x \cdot S_x}[Y_x - E(Y_x)] + \beta_{S_x \cdot Y_x}[S_x - E(S_x)], \qquad (2.22)$$

where $\beta_{Y_x \cdot S_x}$ = the partial regression coefficient for the observed total score Y_x,

$\beta_{S_x \cdot Y_x}$ = the partial regression coefficient for the observed subscale score S_x ,

$E(S_t)$ = the expected value of the true subscale scores across examinees,

$E(Y_x)$ = the expected value of the observed total scores across examinees, and

$E(S_x)$ = the expected value of the observed subscale scores across examinees.

$\beta_{Y_x \cdot S_x}$ refers to the changes in the true subscale scores associated with a one-unit change in observed total scores, holding the observed subscale scores constant, and $\beta_{S_x \cdot Y_x}$ refers to the changes in the true subscale scores associated with a one-unit change in observed subscale scores, holding the observed total scores constant. These partial regression coefficients, $\beta_{Y_x \cdot S_x}$

and $\beta_{S_x \cdot Y_x}$, are determined based on the standard criterion of finding the best prediction line, in which the sum of squared errors in prediction line is minimized (i.e., Ordinary least-squares).

The partial regression coefficient of $\beta_{S_x \cdot Y_x}$ satisfying the standard criterion is obtained by the following equation:

$$\beta_{S_x \cdot Y_x} = \frac{s_{S_t}[r_{S_x S_t} - r_{S_t Y_x} r_{S_x Y_x}]}{s_{S_x}\left[1 - r^2_{S_x Y_x}\right]}, \qquad (2.23)$$

where s_{S_t} = the standard deviation of the true subscale scores,

s_{S_x} = the standard deviation of the observed subscale scores,

$r_{S_x S_t}$ = the correlation between the observed subscale scores and the true subscale scores,

$r_{S_t Y_x}$ = the correlation between the true subscale scores and the observed total scores, and

$r_{S_x Y_x}$ = the correlation between the observed subscale scores and the observed total scores.

Subsequently, the partial regression coefficient of $\beta_{Y_x \cdot S_x}$ satisfying the standard criterion is obtained by the following equation:

$$\beta_{Y_x \cdot S_x} = \frac{s_{S_t}[r_{S_t Y_x} - r_{S_x S_t} r_{S_x Y_x}]}{s_{Y_x}[1 - r^2_{S_x Y_x}]}, \qquad (2.24)$$

where s_{Y_x} is the standard deviation of the observed total scores, and other terms are the same as in the equation (2.23). For estimating the true subscale scores, $E(Y_x)$, $E(S_t)$, $E(S_x)$, $\beta_{S_x \cdot Y_x}$, and $\beta_{Y_x \cdot S_x}$terms should be known. $E(Y_x)$ and $E(S_x)$ are available from the sample means of the observed total scores and the observed subscale scores, respectively. Because $E(S_x)$ would be equal to $E(S_t)$, $E(S_t)$ is substituted with $E(S_x)$. For the computations of $\beta_{S_x \cdot Y_x}$, and $\beta_{Y_x \cdot S_x}$, three terms, $r_{S_x S_t}$, $r_{S_x Y_x}$, and $r_{S_t Y_x}$have to be computed. $r_{S_x S_t}$ is the square root of score reliability in the subscale, achieved by KR-20 and Cronbach α, and $r_{S_x Y_x}$is simply the correlation between the

observed subscale scores and total scores. Refer to the equation (2.20) for the computation of $r_{S_t Y_x}$.

Mean Squared Error (MSE)

One statistic used to evaluate the CTT-based subscale scores is the mean squared error (MSE), based on observed scores. When a statistical model is considered, errors occur because observed scores and their model-predicted scores differ. For example, let an observed subscale score and an estimated subscale score based on a model, X_s and T_s, respectively. In the context of CTT, the error of measurement is obtained by subtracting the estimated score from the observed score (i.e., $E_s = X_s - T_s$). That is, T_s is based on the obtained estimate from the subscale score model rather than an actual true score as based on a simulation. Here, the expected value of squared errors is the mean squared error (MSE). In other words, MSE can be expressed by $E((X-T)^2)$ or $E(e)^2$. Because $Var(X) = E(X^2) - (E(X))^2$, the following equation can be derived:

$$E((X-T)^2) = \sigma^2(X-T) + (E(X-T))^2. \qquad (2.25)$$

Because $X - T = e$, this equation can be rewritten as $E(e^2) = \sigma^2(e) + (E(e))^2$. Namely, the MSE, $E(e^2)$, is the sum of the error variance and the square of error score mean. Because the expected value of errors, $E(e)$, is zero, the MSE is simply abbreviated by $\sigma^2(X-T)$ or $\sigma^2(e)$. The MSE in a squared unit is often transformed into the same scale as scores by taking its squared root value and it will be described as the model-based root mean squared error (RMSE-MB), $\sqrt{E((X-T)^2)}$ or $\sqrt{E(e^2)}$ throughout the manuscript.

The estimates of RMSE-MB vary somewhat across models. Specifically, the standard error of measurement based on the linear regression line from Kelley's method is defined as

$\sigma_{S_t}\sqrt{1-\rho^2_{S_xS_t}}$, and the squared standard error of measurement, $\sigma^2_{S_t}(1-\rho^2_{S_xS_t})$. The standard error of measurement based on the linear regression from Holland-Hoskins' method is $\sigma_{S_t}\sqrt{1-\rho^2_{S_xS_t}}$, and the MSE is $\sigma^2_{S_t}(1-\rho^2_{Y_xS_t})$. Subsequently, the standard error of measurement based on the linear regression equation from Haberman's method is $\sigma_{S_t}\sqrt{(1-\rho^2_{S_xS_t})[1-\rho^2_{S_tY_x\cdot S_x}]}$ or $\sigma_{S_t}\sqrt{(1-\rho^2_{Y_xS_x})[1-\rho^2_{S_tS_x\cdot Y_x}]}$, and the MSE is $\sigma^2_{S_t}(1-\rho^2_{S_xS_t})[1-\rho^2_{S_tY_x\cdot S_x}]$ or $\sigma^2_{S_t}(1-\rho^2_{Y_xS_x})[1-\rho^2_{S_tS_x\cdot Y_x}]$. Large MSEs indicate high amount of prediction error, and vice versa, small MSEs indicate low prediction error.

Proportional Reduction in Mean Squared Error (PRMSE-MB)

The proportional reduction in mean squared error (PRMSE-MB), which measures the added-value of subscale scores over a total score, is the ratio of MSE reduced for a standard value. Specifically, the PRMSE-MB is computed by the ratio of MSE from subscale scores estimated based on a model and that from a constant predictor (i.e., standard or criterion value), in which the standard value is the resulting MSE value when the constant predictor $E(S)$ approximates the true subscale score. This can be written by $1-\frac{MSE_{predictor}}{MSE_{E(S)}}$. By computing the MSEs, the PRMSE-MB is easily computed. First, the MSE of the constant predictor $E(S)$ is

$$E((E(S)-S_t)^2) = s^2(E(S)-S_t) + [E(E(S)-S_t)]^2 \qquad (2.26)$$

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

Note that $\sigma^2_{X+c} = \sigma^2_X$, where *X* is a variable and *c* is a constant. That is, inserting a constant into a variable does not influence the computation of variance.

The PRMSE-MB of subscale scores from Kelley's method, which approximates of the true subscale score, S_t, by an observed subscale score, S_x, is computed as follows:

$$PRMSE - MB_{Kelley} = 1 - \frac{\sigma_{S_t}^2(1-\rho_{S_x S_t}^2)}{\sigma_{S_t}^2} = \rho_{S_x S_t}^2 \qquad (2.27)$$

Given the linear relationships between the true subscale scores and the observed subscale scores, the MSE from the prediction of the true subscale score by the observed subscale score is $\sigma_{S_t}^2(1 - \rho_{S_x S_t}^2)$, as shown earlier, and the resulting $PRMSE - MB_{Kelley}$ is the squared correlation between the observed and the true subscale scores, which is the reliability of raw subscale scores.

Second, the PRMSE-MB of subscale scores from Holland-Hosken's method, which approximates of the true subscale scores, S_t, by the observed total scores, Y_x, is computed as follows:

$$PRMSE - MB_{HH} = 1 - \frac{\sigma_{S_t}^2\left(1-\rho_{Y_x S_t}^2\right)}{\sigma_{S_t}^2} = \rho_{Y_x S_t}^2 \qquad (2.28)$$

The ratio of reduced MSE (i.e., PRMSE-MB) in subscale scores based on the Holland-Hosken's method can be computed by the squared correlations between the observed total scores and the true subscale scores. For the computation of $\rho_{Y_x S_t}^2$, refer to the equation (5). Based on the equation (5), the observed total score can approximate the true subscale score better than the observed subscale score if the product of reliability coefficient of a total test and the squared correlation of the true total score and the true subscale score is higher than the reliability of the observed subscale score. However, because any types of reliability cannot exceed 1.0, the $PRMSE - MB_{HH}$ cannot be higher than that obtained from a whole test. Generally, if the PRMSE-MB from the linear regression of the true subscale score on the total score is quite

small, this method would not be considered as an appropriate method for obtaining subscale scores. The use of total score for approximating the true subscale score is favored in the following situations: a) high reliability of the total score is high enough, b) low correlation between the true subscale score and the true total score, and c) low reliability of the observed subscale scores.

Third, the PRMSE-MB based on the linear regression of the true subscale score on the observed total score and the observed subscale score can be obtained from either the first or the second equation below:

$$PRMSE - MB_{Haberman} = 1 - \left[1 - \rho_{s_t}^2\right]\left[1 - \rho_{s_t Y_x \cdot s_x}^2\right], \text{ or} \qquad (2.29)$$

$$= 1 - [1 - \rho_{Y_x s_t}^2][1 - \rho_{s_t s_x \cdot Y_x}^2].$$

According to Lord & Novick (1968), $\rho_{S_t Y_x \cdot S_x}$ and $\rho_{s_t s_x \cdot Y_x}^2$, respectively, are defined as follows:

$$\rho_{S_t Y_x \cdot S_x} = \frac{\rho_{S_t Y_x} - \rho_{S_t S_x}\rho_{S_x Y_x}}{[1-\rho_{S_t S_x}^2]^{1/2}[1-\rho_{S_x Y_x}^2]^{1/2}}, \text{ and} \qquad (2.30)$$

$$\rho_{S_t S_x \cdot Y_x} = \frac{\rho_{S_x S_t} - \rho_{S_t Y_x}\rho_{S_x Y_x}}{[1-\rho_{S_t Y_x}^2]^{1/2}[1-\rho_{S_x Y_x}^2]^{1/2}}. \qquad (2.31)$$

For the computation of these partial correlation coefficients, sample correlations among the true subscale scores, the observed total scores, and the observed subscale scores can be used. Table 2.1 summarizes the computations of MSEs and PRMSE-MBs of subscale scores from different subscale scoring methods.

Table 2. 1. MSE and PRMSE-MB from Different Predictors

Predictors	MSE	PRMSE-MB

Subscale score	$\sigma_{S_t}^2(1-\rho_{S_xS_t}^2)$	$\rho_{S_xS_t}^2$
Total score	$\sigma_{S_t}^2(1-\rho_{Y_xS_t}^2)$	$\rho_{Y_xS_t}^2$
Subscale score & Total score	$\sigma_{S_t}^2(1-\rho_{S_xS_t}^2)[1-\rho_{S_tY_x\cdot S_x}^2]$ or $\sigma_{S_t}^2(1-\rho_{Y_xS_x}^2)[1-\rho_{S_tS_x\cdot Y_x}^2]$	$1-[1-\rho_{S_t}^2][1-\rho_{S_tY_x\cdot S_x}^2]$ or $1-[1-\rho_{Y_xS_t}^2][1-\rho_{S_tS_x\cdot Y_x}^2]$

If the PRMSE-MB of subscale scores estimated based on the observed subscale scores and the total score is large enough compared to that of subscale scores based on only observed total score, the subscale scores are likely to be desirable for reporting. In the other way, if it is not sufficiently large, it may be inappropriate to report the resulting subscale scores, because it add only slight information over the total test score.

IRT-based Subscale scores

Several IRT-based subscale scoring methods are available. Similar to the raw subscale scores in CTT, unidimensional IRT models estimate an examinee's scale score, θ, based on a subset of items, which can be used as the subscale score. However, these types of subscale scores can have less accuracy than other methods, in that they use only information from items in the corresponding subscale, disregarding other available information from the other subscales in the test. As methods of improving the accuracy of subscale score estimation, Objective Performance Index (OPI; Yen, 1987), augmented subscale scoring (Wainer et al., 2001), and the multidimensional IRT models are available. This section reviews essential concepts in IRT models, including their model configurations, assumptions and estimation follow. Then, the descriptions of subscale scoring methods are followed.

Basic Concepts in IRT

Item response theory (IRT) models specify a relationship between the underlying traits and the probability of item response, which is nonlinear. Item Characteristic Curve (ICC)

represents such nonlinear relationship, in which the probability of item success is monotonically increasing as the trait level increases. In IRT, because the trait level is estimated based on both item properties (e.g., difficulty, discrimination, and guessing) and examinees' response patterns on items. ICC is practically drawn based on the relationship among the probability of item success, item properties, and examinees' item scores. Figure 2.1 presents examples of three ICCs from three dichotomous items. All three ICCs have *S*-shaped curves in which the probability of item success monotonically increases with escalations in the trait level. From the ICCs, small changes in the medium trait level appear to lead large changes in the probability of item success, whereas large changes in the extreme trait levels appear to lead relatively small changes in the probability of item success.

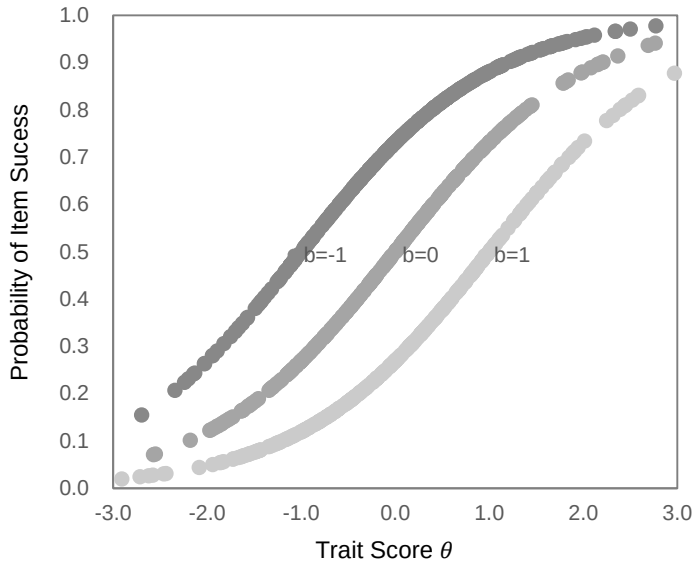

Figure 2.1. Item Characteristic Curves of Three Dichotomous Items

Three dichotomous items used in Figure 2.1 differ in item difficulty. At the same level of trait estimate, the probability of correct item is always the highest for item of b = -1 and the lowest for item of b = 1. Thus, item of b = -1 seems to be the hardest and b = 1 seems to be the

easiest. Although not illustrated in the figure above, ICCs may differ in item discrimination (i.e., slope) and item guessing (i.e., low asymptote), making the interpretation of ICC somewhat complex. For example, when items differ in their discrimination levels, the ICCs from the items may be different in their slopes. Some ICCs may sharply increase as the trait estimates change, while other ICCs may gradually increase as the trait scores increase. Specifically, items with high discrimination values will have large changes in their probabilities in a narrow range of trait estimates, whereas those with low discrimination values will have relatively small probability changes over a broad range of trait estimates. The low asymptote of an ICC corresponds to item guessing that represents the probability of correct item response at the extremely low trait level. Various IRT models, depending on whether item difficulty, item discrimination, or item guessing parameters is involved, can be specified. One Parameter Logistic Model (1-PLM) is the simplest unidimensional IRT model, in which only item difficulty parameters are involved. Two Parameter Logistic Model (2-PLM) involve item discrimination parameters as well as item difficulty. In turn, Three Parameter Logistic Model (3-PLM) includes item guessing parameters as well as item difficulty and discrimination parameters. These models are all unidimensional IRT models because only a single latent trait underlies item-solving. Alternatively, there are several multidimensional IRT models, in which multiple latent traits are assumed, and complex dependency among items are considered.

IRT Assumptions

A common assumption on all IRT models is local independence. Local independence refers to an assumption that an examinee's response to items are independent of each other, after controlling the level of underlying latent traits. Under the assumption, an examinee's performance on an item must not influence his or her performance on other items once his or her

trait level is fixed. Local independence is a key concept in IRT because all likelihood functions for estimating parameters are based on this idea. In IRT, the likelihood function indicates the probability that a person has specific item response patterns given θ. When local independence is met, the likelihood function of an examinee's response patterns on items consisting of a test is computable by the products of probabilities of score patterns on respective items.

This assumption should be required in both unidimensional and multidimensional models as well. The only difference is the number of traits that should be controlled when local independency holds (Embretson & Reise, 2000). Wainer and Wang (2000) argued that the violation of local independence can overestimate test reliability by underestimating the standard error of ability estimates.

Three Main IRT Models: 1-PLM, 2-PLM and 3-PLM

There are three popular IRT models based on which item parameters are involved: One-, Two-, and Three- Parameter Logistic Models. The 1-PLM is the simplest IRT model, in which only item difficulty parameter is involved. The 1-PLM assumes that items in a test have the same item discrimination and their lower asymptotes are low enough to be negligible. In the 1PLM, the probability that person *s* successfully performs item *i* is formulated as below:

$$P_i(\theta_s) = \frac{exp(1.7a(\theta_s - b_i))}{1+exp(1.7a(\theta_s - b_i))}, \qquad (2.32)$$

where a is a common discrimination parameter, b_i is the difficulty parameter for item i, and θ_s is the trait level for person *s*. The probability of solving an item correctly is determined by two factors: person θ and item difficulty. The constant of 1.7 is a scaling factor that transforms the logit scale into the probit scale. Item difficulty parameters are determined at the location on θ continuum at which the probability of a correct response equals 0.5.

The 2-PLM (Birnbaum, 1957, 1958) assumes variable item difficulty and discrimination parameters across items, still keeping the lower asymptotes trivially low. The probability that person *s* successfully performs item *i* in the 2-PL is formulated as follows:

$$P_i(\theta_s) = \frac{exp\ 1.7a_i(\theta_s - b_i)}{1+exp 1.7a_i(\theta_s - b_i)}, \qquad (2.33)$$

where a_i and b_i are the item difficulty and the item discrimination parameters, respectively, and θ_s is the trait level for person *s*. Items with different discrimination weights would differently influence item performance. For example, an item with a high discrimination value will discriminate respondents between low and high ability levels, and an item with a low discrimination value may not discriminate respondents, whose ability levels largely differ, making the ICC flat. Same as in the 1PLM, item difficulty values are the location of θ at which the probability of item equals 0.5 (i.e., p=0.5). Item discrimination values are defined as the slop at the level of *p*=0.5.

The 3-PLM is a model assuming that lower asymptote parameters are involved. In the 3-PLM, the probability that person *s* successfully performs item *i* is shown below:

$$P_i(\theta_s) = c_i + (1 - c_i)\frac{exp(1.7a_i(\theta_s - b_i))}{1+exp(1.7a_i(\theta_s - b_i))}, \qquad (2.34)$$

where c_i is the lower asymptote for item *i*, and the remaining terms are the same as in 2-PLM. The lower asymptote is interpreted as the probability that a person with very low ability answers an item correctly.

Estimation Methods for IRT Modeling

For estimating item parameters, joint maximum likelihood (JML), marginal maximum likelihood (MML), and conditional maximum likelihood (CML) are popular. These three types

of methods differ in how they handle unknown person estimates. JML estimates person parameters with item parameters fixed, and then item parameters are re-estimated with the acquired person estimates. These calibration processes between items and persons are iterated until convergence criterion is satisfied. JML is not frequently recommended because it often yields biased and inconsistent estimators (Embretson & Reise, 2000). Similarly, the CML method treats person parameters as known, and uses total score as known thetas to estimate person parameters. CML is limitedly used in the Rasch model, in which the total scores are sufficient statistics. Therefore, the following method, MML, is the most common.

The MML method estimates item parameters under the assumption that population θs are distributed with specific means and variances (e.g., normal distribution), although each θ is specifically unknown. MML includes the computation of integration when the likelihood of response patterns is weighted and added together across all rectangles that are created by setting several quadrature points in the prior distribution. When a large number of quadrature points are set, MML is reported to have some issues related to integration. The complexity of integration computation is also dramatically increased for multidimensional models. The more quadrature points selected, the more accurate estimates we can obtain. In addition, a number of examinees are required for more accurate estimation. Bock and Aitken (1981) employed the expectation-maximization (EM) implementation procedure of MML and resolved this problem to some degree. In the Expectation stage, the number of people at each quadrature and the number of persons answering an item successfully are predicted, and in the Maximization stage, item parameters are determined based on the criteria to maximize the likelihoods. These two stages are iterated until the changes in likelihood values are minimalized. Estimators are determined based on a Newton-Gauss procedure (See Hambleton & Swaminathan, 1985 for more details).

Three methods for estimating person scores are common: maximum likelihood (ML), maximum a posteriori (MAP), and expected a posteriori (EAP) estimations. The ML method finds θ that maximizes the likelihood function of a response pattern, with item parameters assumed to be known. Specifically, given θ, the likelihood of each item response pattern is computed and summed up over all items, and θ value that maximizes the likelihood value is determined as a person estimate. Although the ML method generally produces consistent estimators, the application of this method requires more caution because it may fail to locate appropriate estimates if all items or none are correctly answered.

The next methods, MAP and EAP methods, estimates person scores based on the Bayesian approach. The basic concept of Bayes' theorem is that the probability that an event will occur is conditioned on the event that was previously occurred. The Bayes' theorem assumes that the information from the previous events influence the probability that the subsequent events occur. This can be formulated as follows:

$$P(A|B) = (P(B|A)P(A))/P(B), \qquad (2.35)$$

where $P(A)$ is a prior distribution of latent variables, $P(B|A)$ is the likelihood of observed responses given the prior distribution, $P(B)$ is the marginal distribution of the observed response pattern, and $P(A|B)$ is the posterior distribution of latent variables given response data. If all parameters are assumed to have a prior distribution based on prior knowledge, estimation is fully Bayesian. This theorem is employed as a way of estimating continuous posterior estimates, θ estimators, in MAP and EAP methods. Specifically, the formula above can be rewritten as follows:

$$f(\theta|U) = f(U|\theta)f(\theta)/f(U), \qquad (2.36)$$

where $f(\theta|U)$ indicates the posterior density function of θs that we try to obtain, $f(U|\theta)$ is equal to the likelihood function, and $f(\theta)$ corresponds to the prior distribution. Because the probability of specific item response pattern is pre-specified for a given set of item responses, $f(U)$ is a constant. Thus, the posterior density function of θs is determined by the products of the likelihood function and prior distribution. Although both the MAP and EAP methods are based on the Bayes theorem, they differ in selecting the mode and the mean of the posterior distribution as ability θ, respectively. Note that the Bayesian approach may not be properly used when there is not reasonable information on prior distribution of θs.

IRT-based Subscale Scoring Methods

Objective Performance Index (OPI)

Yen (1987) proposed a subscale scoring method, OPI, that estimates a true subscale score based on the performance of items in a subscale. OPI combines information about an examinee's overall test performance into his or her subscale score. Yen's approach to subscale scores is analogous to the Haberman's weighted average method for subscale scores in that it uses collateral information from a total score so that it increases the accuracy of the true subscale score. However, unlike to Haberman's method, OPI basically uses the IRT scale as prior information in the Bayesian procedure. Specifically, it estimates a global trait score based on the entire test for each examinee, and uses this information to build a prior distribution for estimating the subscale score. Thus, Yen's method will provide more accurate estimation about one's subscale score when subscales are highly correlated. In the meantime, the prior distribution for more stable estimation is person-specific. Each examinee has his or her own individual prior

distribution for estimating the subscale score because each examinee has his or her own ability estimate. OPI method may increase the estimation accuracy in that it uses an informative prior distribution based on examinees' global ability estimates, rather than a random prior distribution (e.g., standard normal distribution).

Estimation of the Prior Distribution of T_s

The OPI procedure assumes that a test of *N*-items consists of *S* subscales with n_s items, in which each item is related to only a single dimension rather than multiple dimensions. In OPI, the true subscale score, T_s, is defined as the expected value of *P*-values for observed number-correct scores, $E(X_s/n_s)$, where X_s are obtained from an examinee's repeated administration of a subscale. Yen (1987) believed that if there is additional information for the true subscale score estimation, we would be able to obtain more accurate and stable subscale scores by combining such information (i.e., prior distribution) into true subscale score estimation.

In the OPI procedure, the prior distribution of T_s for an examinee follows a beta distribution, $beta(\alpha_s, \beta_s)$, and is expressed by the equation as follows:

$$g(T_s) = \frac{(\alpha_s+\beta_s-1)!T_s^{\alpha_s-1}(1-T_s)^{\beta_s-1}}{(\alpha_s-1)!(\beta_s-1)!}, \qquad (2.37)$$

where α and β are shape parameters as exponents of the random variables and have larger values than zero. For estimating the prior distribution of T_s in practice, several procedures are required. First, OPI estimates the trait level θ for each examinee and item parameters based on the whole test performance on a whole test using the 3PLM. Then, the mean and the variance of $\hat{T}_s$ is obtainable using the 3PLM parameter values as below:

$$\mu_{\hat{T}_s} = \frac{1}{n_s}\sum_{i=1}^{n_s} P_{js}(\theta_j), \text{ and} \qquad (2.38)$$

$$\sigma_{\hat{T}_s}^2 = I(T_s, \hat{T}_s)^{-1}, \tag{2.39}$$

where $\mu_{\hat{T}_s}$ approximately equal the mean and the variance of T_s, and $\sigma_{\hat{T}_s}^2$ equals the amount of information that $\hat{T}_s$ accounts for T_s. For estimating the variance of prior distribution,

$$I\left(T_s, \hat{T}_s\right)^{-1} = \frac{I(\theta, \hat{T}_s)}{\left[\frac{\partial T_s}{\partial \theta}\right]^2}, \tag{2.40}$$

where $\frac{\partial T_j}{\partial \theta} = \frac{\partial[\frac{1}{n_s}\sum_{i=1}^{n_s} P_{js}(\theta_j)]}{\partial \theta} = \frac{\frac{1}{n_s}\sum_{t=1}^{n_s} \partial P_{js}(\theta_j)}{\partial \theta} = \frac{1}{n_s}\sum_{t=1}^{n_s} P'_{js}(\theta_j)$ and $P'_{js}\left(\theta_j\right) = \frac{Da_{is}[1-P_{js}(\theta)][P_{js}(\theta)-c_{is}]}{1-c_{is}}$. From Lord (1980), $I\left(\theta, \hat{T}_s\right)$ can be approximately estimated by $I\left(\theta, \hat{\theta}_s\right)$, and $I\left(\theta, \hat{\theta}_s\right) = \sum_{s=1}^{S}\sum_{t=1}^{n_s} \frac{[P'_{js}(\theta_j)]^2}{P_{js}(\theta)[1-P_{js}(\theta)]}$. Thus, $\sigma_{\hat{T}_s}^2$ can be expressed as follows:

$$\sigma_{\hat{T}_s}^2 = \frac{\frac{1}{n_s}\sum_{t=1}^{n_s} P'_{js}(\theta_j)^2}{\sum_{s=1}^{S}\sum_{t=1}^{n_s} \frac{[P'_{js}(\theta_j)]^2}{P_{js}(\theta)[1-P_{js}(\theta)]}}. \tag{2.41}$$

Also, the prior distribution of T_s follows the $beta(\alpha_s, \beta_s)$ distribution, where the shape parameter of the beta distribution can be expressed in terms of the mean, $\mu_{\hat{T}_s}$, and the variance, $\sigma_{\hat{T}_s}^2$, explained above. The mean and the standard deviation of the $beta(\alpha_s, \beta_s)$ are, respectively, $\frac{\alpha_s}{\alpha_s+\beta_s}$ and $\frac{\alpha_s\beta_s}{(\alpha_s+\beta_s)^2(\alpha_s+\beta_s+1)}$. In the other way, the shape parameters, α_s and β_s, can be expressed by formulation of these mean and variance. Thus,

$$\alpha_s = \frac{\mu_{\hat{T}_s}{}^2(1-\mu_{\hat{T}_s})}{\sigma_{\hat{T}_s}^2} - \mu_{\hat{T}_s}, \text{ and} \tag{2.42}$$

$$\beta_s = \frac{\mu_{\hat{T}_s}(1-\mu_{\hat{T}_s})^2}{\sigma_{\hat{T}_s}^2} + \mu_{\hat{T}_s} - 1. \tag{2.43}$$

Given the prior distribution of T_s, X_s is assumed to follow a binomial distribution and specified as follows:

$$P(X_s = x_s|T_s) = \binom{n_s}{x_s} T_s^{x_s}(1 - T_s)^{n_s - x_s}, \qquad (2.44)$$

where x_s is the observed correct score for items in subscale *s*. When the posterior distribution of T_s is defined as $g(T_s|X_s = x_s) = beta(\gamma_s, \delta_s)$, the parameters of the posterior distribution can be expressed in terms of the parameters of the prior information as below:

$$\gamma_s = \alpha_s + x_s, \text{ and} \qquad (2.45)$$

$$\delta_s = \beta_s + n_s - x_s. \qquad (2.46)$$

OPI is estimated by the mean of the posterior distribution as the true subscale score, $\frac{\gamma_s}{\gamma_s+\delta_s}$. If the term n^* substitutes $\frac{\mu_{\hat{T}_s}(1-\mu_{\hat{T}_s})}{\sigma^2_{\hat{T}_s}} - 1$, the OPI value can be expressed as $\frac{\mu_{\hat{T}_s} n_s^* + x_s}{n_s^* + n_s}$. In turn, the variance of the posterior distribution is obtained by $\frac{\gamma_s \delta_s}{(\gamma_s+\delta_s)^2(\gamma_s+\delta_s+1)}$. From Lord (1980), the standard errors of estimates can be derived by the square root of the variance. Yen (1987) also suggested the computation of weighted OPI value. Namely, the mean of the posterior distribution is $w_s \hat{T}_s + (1 - w_s)\frac{x_s}{n_s}$, in which w_s is the relative weight of the prior estimate and the observed proportion-correct score, $\frac{x_s}{n_s}$, and is computed using $\frac{n_s^*}{n_s^* + n}$.

The OPI procedure requires to compute the statistic Q in order to check how accurate and stable estimates the prior information can lead. In specific, the Q statistic identifies unexpected item responses on the subscales in a test.

$$Q = \sum_{s=1}^{S} \frac{n_s(\frac{x_s}{n_s} - \hat{T}_s)^2}{\hat{T}_s(1-\hat{T}_s)}. \quad (2.47)$$

When $Q > x^2(J, .10)$, n_s^* is set to zero in the equations above.

Adjustment of OPI scores

Because the prior mean of T_s, $\hat{T}_s$, is calculated based on the performance of total items in a test, the prior mean is not independent of x_s. Although it seems to be reasonable to use only items that are not relevant to the x_s, it would make the computational procedure of OPI scores more complex. Thus, OPI values obtained by equations above can overestimate the amount of the prior information independent of x_s. Considering this situation, reducing the overlapping information may produce more accurate estimation. Yen (1987) suggested an adjusted OPI value by weighting the total test information by $\frac{n-n_s}{n}$. That is, the adjusted OPI value is obtained by multiplying the total test information by the ratio of the number of items that are not relevant to x_s to the total items.

Augmentation Method

Wainer et al. (2001) proposed the subscale score augmentation method, in which subscale scores are augmented by employing subsidiary information from the remaining subscale scores as well as a subscale score being considered. This method estimates true subscale scores through multiple stages. In the first stage, unidimensional IRT ability scores or the observed subscale scores are estimated based on the responses of items within each subscale. Here, unidimensional IRT ability scores can be estimated based on one of the ML, MAP, or EAP methods as described earlier. In the second stage, the estimated IRT ability scores are approximated by weighted subscale scores, in which the weights of subscale scores depend on the IRT-based reliability

estimates and all subscale scores are involved. The following three equations present the formula for the augmented subscale scores:

$$ML(\hat{\theta}) = \overline{ML(\theta)} + \rho\big(ML(\theta) - \overline{ML(\theta)}\big), \tag{2.48}$$

$$MAP(\hat{\theta}) = \overline{MAP(\theta)} + \rho\big(MAP(\theta) - \overline{MAP(\theta)}\big), \text{ and} \tag{2.49}$$

$$EAP(\hat{\theta}) = \overline{EAP(\theta)} + \rho\big(EAP(\theta) - \overline{EAP(\theta)}\big), \tag{2.50}$$

where the estimated IRT-based ability scores substitute the true subscale scores, and ρ indicates the reliability index of these estimates. Wainer et al. (2001) described two different ways of estimating the reliability: MAP based reliability and EAP based reliability. The reliability of $MAP(\theta)$ values is computed as below:

$$\hat{\rho} = \frac{VAR(\theta_{MAP})}{VAR(\theta_{MAP}) + E(SE^2(\theta_{MAP}))}, \tag{2.51}$$

in which $VAR(\theta_{MAP})$ is the variance of θs obtained based on the MAP method, and the $E(SE^2(\theta_{MAP}))$ is the expected value of the squared errors of estimates. Also, the reliability of $EAP(\theta)$ can be computed by the formula below:

$$\hat{\rho} = 1 - \bar{\sigma}_e^2, \tag{2.52}$$

where $\bar{\sigma}_e^2$ is the expected value of error variance of θs.

The subscale score augmentation method uses either the number-correct scores (i.e., raw subscale scores) or the IRT trait level estimates as the observed subscale score. The combination of scale scores is used for approximating a true subscale score. In IRT scaling, estimation errors vary among different scale scores. However, when the scaled scores are used to approximate the true subscale score, the equal levels of measurement errors are required. Thus, this method

ignores individual standard errors and substitutes them by a constant as the measurement error in CTT.

The augmentation of raw scores by using all the available observed subscale scores may produce more reliable estimates by using information from other subscale scores in the test. It is reasonable to utilize available collateral sources of information as well as the information from the corresponding subscale items for computing more accurate subscale scores with the little number of items. This method seems to be working better when correlations among subscale scores are fairly high. However, note that the subscale scores do not have values over a total score if correlations among subscale scores are too high, indicating that test items are unidimensional. This approach is not applicable when items only have impacts only on a subscale with simple structure.

Multidimensional Latent Trait Models

Multidimensional latent trait models yield a set of trait scores on multiple dimensions that may influence item performance by relating them to a set of item parameters. These types of models can be considered for uses in educational tests or personality tests that are designed for measuring broad areas and multiple subject domains. Multidimensional latent trait model can be categorized by compensatory and noncompensatory models. Compensatory models assume that high ability on a dimension can make up for low ability on other dimensions, whereas noncompensatory models assume that low ability on a dimension is not compensated by high ability on other dimensions. Compensatory models include Multidimensional Rasch Model (Adams, Wilson, and Wang, 1997), Multidimensional Two Parameter Logistic Model (Reckase & McKinley, 1991), and Multidimensional Three Parameter Logistic Model, and non-compensatory models include the Multicomponent Latent Trait Model (MLTM; Whitely, 1980),

the Generalized Latent Trait Model (GLTM; Embretson, 1984), and an extension of the GLTM, the Multicomponent Latent Trait Model for Diagnosis (MLTM-D; Embretson, & Yang, 2013)

Compensatory MIRT Models

The multidimensional Rasch model is one of multidimensional models where multiple dimension θs are involved to formulate the probability of a correct item response. Specifically, multidimensional Rasch model is formulated as follows:

$$P(\underline{\theta}_s) = \frac{\exp(\sum_{m=1}^{M} \theta_{sm} + \delta_i)}{1+\exp(\sum_{m=1}^{M} \theta_{sm} + \delta_i)}, \quad (2.53)$$

where θ_{sm} is the trait estimate from person *s* on dimension *m*, δ_i is the intercept for item *i*, and M is the number of dimensions. The probability of answering an item correctly is determined by the combination of trait estimates with equal weights and an item difficulty (i.e., intercept). The $\underline{\theta}_s$ vector presents a set of latent trait estimates on multiple dimensions.

The multidimensional two parameter logistic model considers item discriminations (i.e., dimension weights) as well as an item difficulty as item properties. Unlike to Rasch model, these item discriminations are unequal across dimensions. Specifically, the probability of a correct item response is written as follows:

$$P(\underline{\theta}_s) = \frac{\exp(\sum_{m=1}^{M} a_{im}\theta_{sm} + \delta_i)}{1+\exp(\sum_{m=1}^{M} a_{im}\theta_{sm} + \delta_i)}, \quad (2.54)$$

where a_{im} is the item discrimination for dimension *m* related to item *i*. The probability of answering an item correctly is determined by a weighted combination of the trait estimates and item difficulty values (i.e., intercept). If a dimension weight is high for a specific dimension, the impact of the corresponding trait score to the item probability gets high. In contrast, a dimension weight is low for the other dimension, the impact of the corresponding trait score is less

influential to item performance. The multidimensional three parameter logistic model is formulated as follows:

$$P(\underline{\theta}_s) = c_i + (1 - c_i)\ \frac{\exp(\sum_{m=1}^{M} a_{im}\theta_{sm} + \delta_i)}{1+\exp(\sum_{m=1}^{M} a_{im}\theta_{sm} + \delta_i)}, \qquad (2.55)$$

where c_i indicates the guessing parameter. The guessing parameter is interpreted in the same way as that of the unidimensional 3PL model.

These three types of models, as above, are compensatory in that weighted dimension scores are summed over dimensions related to an item to compute the probability of a correct item response. In compensatory models, a low weighted dimension seems to be compensated by a high weighted dimension to increase the probability of item success, implying that examinees do not require high dimension scores on all relevant dimensions. Marginal maximum likelihood estimators (MMLE) using expectation-maximization (EM) and Markov chain Monte Carlo (MCMC) algorithms have been developed for calibrating parameters.

Non-compensatory MIRT Models

According to the different interactions of dimensions underlying items, noncompensatory models include GLTM, MLTM, and MLTM-D, in which item probability is formulated with the product of relating parameters.

MLTM

The MLTM is one of the noncompensatory multidimensional models in which multiple processing or skill components (i.e., dimension) are involved in item solving. The MLTM assumes that each item consists of multiple subtasks measuring the processing or skill components, and the responses from all the subtasks are required. The MLTM models three situations depending on relationships among components: independent components,

sequentially-dependent components, and components of repeatable dependent-sequence. The independent component model of the MLTM is applied in the case in which components are independent and exhaustive, and the other two models are applied in cases in which components are assumed to be sequentially dependent.

The MTLM requires both component responses and item responses, and estimate component parameters (i.e., component θs and component difficulty values) by linking the component responses to the corresponding item responses. The following formulas are the mathematical equations of defining the MLTM for independent components, in which components are independent:

$$P\left(x_{ij}=1\middle|\theta_j,b_i\right)=a\prod_{k=1}^{K}P\left(x_{ijk}=1\middle|\theta_{jk}\right)+g(1-\prod_{k=1}^{K}P\left(x_{ijk}=1\middle|\theta_{jk}\right)),\text{ and} \quad (2.56)$$

$$P\left(x_{ijk}=1\middle|\theta_{jk}\right)=\frac{\exp(\theta_{jk}-b_{ik})}{1+\exp(\theta_{jk}-bi_k)}, \quad (2.57)$$

where the probability of item success, $P\left(x_{ij}=1\right)$, is termed by the products of component success probabilities, $P\left(x_{ijk}=1\right)$, related to the item performance. Here, the component probability is estimated by the Rasch model, in which component difficulty parameter, b_{ik}, and a component level theta, θ_{jk}, score are involved. In the Rasch model, item discrimination is fixed to one. The *a* and *g* parameters, respectively, represent component information (i.e., meta-component or executive functioning) and an alternative solving method of item (i.e., guessing or rote association to the stem). Specifically, *a* is the probability of item solving when all the required subtasks are responded, and *g* is the probability of item solving when at least one required subtask is not responded.

In the other two cases, where components are sequentially-dependent or with repeatable dependent-sequence, the correct response to a component requires information of prerequisite components. Although the response from the lowest sequence of subtask reflects component outcomes, the responses from the second or higher sequence of subtasks are the jointed outcomes, in which the preceding component outcomes will be influenced by the jointed outcomes in the second sequence. The major distinction between these models is from that the sequentially-dependent component model requires that all components are executed only once, whereas the repeatable dependent-sequence component model allows components to be executed over and over. More detailed information about these specific cases, see Whitely (1980) and Embretson (1984, 1985).

GLTM

The GLTM is an extended model of the MLTM, in which an item component difficulty is replaced by a linear combination of complexity factors for each component, as in the Log Linear Test Model (LLTM; Fischer, 1973). The GLTM equals the MLTM except that the probability of correct component is based on the LLTM instead of Rasch model. The GLTM generalizes both the MLTM and the LLTM. The formula of the GLTM is specified as below:

$$P(x_{ij}=1|\theta_j) = a\prod_{k=1}^{K} P(x_{ijk}=1|\theta_{jk}) + g(1-\prod_{k=1}^{K} P(x_{ijk}=1|\theta_{jk})), \text{ and} \quad (2.58)$$

$$P(x_{ijk}=1|\theta_{jk}) = \frac{\exp(\theta_{jk}-(\sum_m c_{imk}\eta_{mk}+d_k))}{1+\exp(\theta_{jk}-(\sum_m c_{imk}\eta_{mk}+d_k))}, \quad (2.59)$$

where c_{imk} is the complexity factor related to component *m* in item *i*, η_{mk} is the weight of the difficulty for complexity factor *k* related to component *m*, and θ_{jk} is the complexity level of

examinee j. The remaining terms are equal to those in the MLTM. If there is only one component (a subtask) within an item, then the GLTM equals the LLTM, and if there are no complexity factors underlying each component and multiple subtasks are involved in an item, then the model equals the MLTM.

MLTM-D

The MLTM-D is a noncompensatory latent trait model for diagnosis. The MLTM-D estimates dimension properties at two different levels with hierarchy: components and attributes. In the MLTM-D, the probability that examinee *j* solves item *i* is the products of the probabilities of the components that are relevant to the item, in which component probabilities are modeled with the weighted combination of the nested attribute variables and component θs.

$$P_{ij} = P(X_{ij} = 1) = \prod_{m=1}^{M} P_{ijm}^{cim}, \tag{2.60}$$

where c_{im} is a binary variable presenting the involvement if component *m* is required to solve an item *i*, and P_{ijm} represents the probability that person *j* performs component *m* in item *i* successfully. The probability of component success is similar to the LLTM as follows:

$$P_{ijm} = P\left(X_{ijm} = 1 \middle| \theta_{jm}, \underline{q}_{im}, \eta_m\right) = \frac{\exp(1.7\ (\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}))}{1 + \exp(1.7\ (\theta_{jm} - \sum_{k=1}^{K} \eta_{mk} q_{imk} + \eta_{m0}))}, \tag{2.61}$$

where θ_{jm} reprsents the examinee *j* on the component *m*, η_{mk} represents the weight of feature *k* on component *m*, and q_{imk} indicates the score of stimulus *k* on component *m* of item *i*.

The MLTM-D requires two Q-matrices, C_{bxM} and Q_{IxK_m}, specifying both component structure and attribute structure. Specifically, the C_{bxM} matrix specifies all possible patterns among components and items in a test, in which M components can yield $2^M - 1$ patterns. In practice, the matrix of items and components is usually one of the subsets of the possible

patterns. The Q_{IxK_m} matrix specifies the relationship between attributes and items within each component. In the MLTM-D, trait level estimates are obtained at the component levels. For example, if there are two components required to solve items, trait levels on only these two components are to be estimated for each person rather than attribute level estimates. However, component estimates, θ_{jm}, can be linked and directly comparable to the attribute weights in special cases (i.e., attributes are linearly ordered) because they can be located on the common scale. Each person's performance can be evaluated compared to the level of specific attributes as well as the level of components.

The MLTM-D is applicable in large-scale tests that are designed with hierarchical knowledge structures with broad skills and more specific skills. Especially when the number of attributes is large, the MLTM-D has an advantage in parameter estimation by decreasing the computational load.

CDA Model Based Subscale scores

CDA models embrace all psychometric models that were developed or utilized for the purpose of providing examinees for attribute mastery profiles on cognitive processes, skills, and knowledge structures underlying items. Although diagnostic models have been developed for the diagnostic use, they can be used for tests that are to be analyzed and reported for the purpose of providing diagnostic information. This section introduces several diagnostic models as measuring tools for diagnostic information. The models that are described here are limited to general classification models for diagnosis. For more details information about specific models, see Roussos, Templin, and Henson (2007) and Rupp and Templin (2008). A general model takes a form that can be expressed in various forms based on its parameterization. The log-linear cognitive diagnostic model (LCDM; Henson, Templin, & Willse, 2009), the general diagnostic

models (GDMs; von Davier & Yamamoto, 2004, 2007), generalized deterministic input, noisy "and" gate (G-DINA; de la Torre, 2011), and the MLTM-D can be categorized as general models.

Diagnostic assessment models can be divided into two groups according to the type of measurement scale for attribute proficiency levels: diagnostic latent trait model vs. diagnostic latent classification model. Specifically, diagnostic latent trait model provides continuous scores on multiple attribute/dimensions, whereas diagnostic latent trait model provides discrete level of scores on these attributes. The selection of appropriate diagnostic models depends on multiple factors: the number of skills/attributes, data type of resulting attribute scores (e.g., dichotomous/polytomous), the structure of data (e.g., attributes hierarchy or attributes dependency), the availability of computer software, and so on. The following section includes the description of diagnostic latent class models. For the information of diagnostic latent trait models, see the previous section of multidimensional latent trait models.

Diagnostic Latent Class Model

All diagnostic classification models (DCMs) provide item parameters and attribute mastery patterns as classes, which determine the probability of a correct response. DCMs commonly assume that examinees with the same attribute mastery profile have the same probabilities of item responses. Similar to the latent trait models for diagnosis (e.g., multidimensional latent trait models), diagnostic latent class models are also classified into compensatory and noncompensatory models based on the interaction of attributes required for successful task performance. The deterministic input, noisy "and" gate (DINA; Haertel, 1989; Junker & Sijtsma, 2001; de la Torre & Douglas, 2008), the reparameterized unified model (RUM; Hartz, 2002), and the unified model (DiBello, Stout, & Roussos, 1995) belong to

noncompensatory models, and the Compensatory RUM, the deterministic inputs, noisy, "or" gate (DINO; Templin & Henson, 2006), and the noisy inputs, deterministic "or" gate (NIDO; Templin, 2006) models belong to compensatory classification models. A full taxonomy of DCM is described with more details in Rupp, Templin, and Henson's (2010) book. Their taxonomy includes eighteen diagnostic classification models according to the type of response data and attribute proficiency classes (i.e., dichotomous vs. polytomous) and the interactive relationships among attributes (i.e., compensatory vs. noncompensatory)

GDM

A class of GDMs is the most general form among all kinds of developed diagnostic models in which both continuous and categorical latent variables are permitted. GDMs are also general in that they can handle both compensatory and noncompensatory skill interactions of item-solving. Based on the parameterization, GDMs can be specialized by various IRT models (e.g., the Rasch model, the two-parameter logistic IRT model, the generalized partial credit model) and diagnostic models (e.g., the latent class analysis; Maris, 1999; the fusion model; Hartz, Roussos, & Stout; 2002). In GDMs, the marginal probability of a vector of observed variables given attribute patterns can be expressed as follow:

$$P\left(x_{j1}, x_{j2}, \ldots, x_{jn}\right) = \sum_g p(g) \int p\left(x_{j1}, \ldots, x_{jn} \middle| a, g\right) da, , \qquad (2.62)$$

where $p(a|g)$ represents the probability of a vector of latent attribute variables given g distribution, and $p\left(x_{j1}, \ldots, x_{jn} \middle| a, g\right)$ represents the conditional probability that person j has specific item responses (i.e., a vector of responses) given the vector of latent attribute variables (i.e., **a** = $(a_{j1}, a_{j2}, \ldots, a_{jn})$) and distribution g. The class of general diagnostic models is formulated in a logistic form as the following equation:

$$P(x|\beta_i, q_i, \gamma_i, \boldsymbol{a}) = \frac{\exp(\beta_{xi}+\gamma_{xi}^T h(q_i,\boldsymbol{a}))}{1+\sum_{y=1}^{m_i}\exp(\beta_{yi}+\gamma_{yi}^T h(q_i,\boldsymbol{a}))}, \quad (2.63)$$

where β_{xi}, r_{xi}, and $h(q_i, a)$ indicates overall difficulty parameters, a *k*-dimensional slope parameter for each response category, and a linear combination of attribute level and Q-matrix, respectively. The slop parameter, r_{xi}, represents the weight of attribute variables to determine the probability of an item success. In term $h(q_i, a)$, the q_i is a term to relate item *i* to skill *k*, and **a** is a vector of an examinee attribute proficiency. The *h* is a term to specify how Q-matrix elements are related to the skill patterns, which determines the specific cases of the GDMs. For example, if item *i* involves skill *k* (i.e., $q_{ik} = 1$), then the term is replaced by $\gamma_{xi}a_k$.

While many other complex models depend on the MCMC estimation algorithm, an MML estimation using the EM-algorithm for the GDMs was developed, and the parameter estimation was successfully recovered with simulated data (von Davier, 2005).

LCDM

The LCDM is a log-linear model in which latent class variables are involved. The log-linear model was originally formulated to predict the frequency in cells in which observable discrete variables intersect, but could be easily extended to latent variables. For such a reason, the log-linear model could be applied to formulate several cognitive diagnosis models (von Davier, 2005; Fu, 2005). The LCDM is a general log-linear model in which dichotomous latent variables and dichotomous response data are involved. The LCDM provides the probability of a correct item response given a binary item response and attribute patterns. The probability of a correct response is formulated as below:

$$\mathrm{P}\left(X_{ij} = 1 \middle| \alpha_j\right) = \frac{\exp(\lambda_j^T h(\alpha_j, q_i) - \eta_i)}{1+\exp(\lambda_j^T h(\alpha_j, q_i) - \eta_i)}, \quad (2.64)$$

where the vector λ_j^T is the vector of weights for item *j*, $h(\alpha_j, q_i)$ represents linear combinations of the α_j(i.e., attribute variables involved in person *j*) and the Q-matrix values of attributes in item *i*, q_i, and η_i represents the probability of a correct response for examinees in class who have not mastered any attributes. Specifically, $\lambda_j^T h(\alpha_i, q_j)$ can be expressed as below:

$$\lambda_j^T h(\alpha_j, q_i) = \sum_{k=1}^{K} \lambda_{ik}(\alpha_k q_{jk}) + \sum_{k=1}^{K} \sum_{v>k} \lambda_{ikv}(\alpha_k \alpha_v q_{ik} q_{iv}) + \cdots, \qquad (2.65)$$

where λ_{ik} and λ_{ikv} are terms that are relevant to the main effect for attribute *k* involved in item *i* and the two-way interaction effect for attribute *k* and *v* involved in item *i*, respectively. The remaining parts sum up all possible multiple interaction effects including three-way interaction, four-way interaction and so on. The natural logarithm of the probability of a correct response corresponds to a logit function. A logit function is directly expressed by a linear combination of intercept, main effects, and interaction effects. For example, once two attributes are involved in item-solving, the logit of a correct response to item *j* by examinee *i* can be expressed by $\text{logit}(X_{ij} = 1 | \alpha_j) = \lambda_{i,0} + \lambda_{i,1,(1)}\alpha_{j1} + \lambda_{i,1,(2)}\alpha_{j2} + \lambda_{i,2,(1,2)}\alpha_{j1}\alpha_{j2}$. In this function, $\lambda_{i,0}$ is the logit for nonmatery groups of both attribute 1 and 2 involved in item *i*. The $\lambda_{i,1,(1)}$ and $\lambda_{i,1,(2)}$ terms that correspond to main effects, respectively, account for the increases in the logits when mastering attribute 1 and 2 involved in item *i*. In turn, $\lambda_{i,2,(1,2)}$ as an interaction term of attribute 1 and 2 represents the increase in the logit when mastering both attribute 1 and 2. The different combinations of attribute mastery patterns yield different size of logits combined with the effect of each attribute, and in turn producing the different probability of a correct response.

A multiple-way ANOVA model resembles the LCDM in that an ANOVA model has main factors and interaction effect among factors. For example, a two-way ANOVA model is represented by a linear combination of main effects from two factors X and Y and an interaction

effect between X and Y. The factors from the ANOVA correspond to the attributes in the LCDM. However, they differ in that factors in the ANOVA are continuous, but the attributes in the LCDM are binary. Particularly, the ANOVA model becomes a very similar form to the LCDM by dummy-coding two factor variables. Both models predict an item response using a linear combination of main effects and interactions.

As a general model, the LCDM embraces both noncompensatory models such as the DINA, the NIDA, and the reduced NC-RUM and compensatory models such as the C-RUM, the DINO, and the NIDO. For example, when main effect terms in the LCDM are disregarded and the term estimates are set to 0, the LCDM is same as the DINA. On the contrary, if no interaction effects among attributes are assumed, only terms relevant to main effects remain, resulting in compensatory RUM (Hartz, 2002).

G-DINA

The G-DINA model is one of general models based on the DINA model. The DINA model is the most simple and parsimonious CDA models, requiring only two parameters per each item, slip and guess parameters, regardless of the number of attributes involved in item-solving. The DINA model as a conjunctive model assumes that all required attributes are required to answer an item correctly. The lack in at least one attribute drastically decreases the probability to answer the item correctly, and produce the same results as the case in which all required attributes were not acquired. That is, the DINA model may classify all examinees who did not master all the required attributes for item-solving, not considering the degrees of deficiency regarding the required attributes and simply classifying all examinees by two groups. The G-DINA model addresses this assumption of the DINA model that all kinds of attribute mastery patterns except for the case in which all required attributes were mastered by an

examinee have the same probability of item success by relaxing the assumption. Specifically, the G-DINA model divides the latent attribute groups into 2^{K_i}, where $K_i^* = \sum_{k=1}^{K} q_{jk}$ is the number of attributes involved in item i. Then, the reduced attribute vector consisting of only K_i attributes can be expressed as $\alpha_i^*=(\alpha_{i1}, \ldots \alpha_{iK_i^*})'$ without considering a full attribute vector, $\alpha_i=(\alpha_{i1}, \ldots \alpha_{iK})'$. In the G-DINA model, the probability that an examinee will answer an item successfully are conditioned on the specific attribute vector, α_{ij}^* , represented by $P(X_i = 1|\alpha_{ij}^*)$, and thus $2^{K_i^*}$ parameters need to be estimated for item i.

Three different types of link functions for the probability of an item success given attribute structure, identity, logit, and log, have been proposed. These link functions can be transformed into the DINA model, DINO model, NIDA model, or reduced RUM as special cases. All these link functions largely can be divided by two terms regarding the main effects of specific attributes and their interactions. Specifically, the identity link function is formulated as follow:

$$P(X_i = 1|\alpha_{ij}^*) = \delta_{i0} + \sum_{k=1}^{K_i^*} \delta_{ik}\alpha_{ik} + \sum_{k'=k+1}^{K_j^*} \sum_{k=1}^{K_j^*-1} \delta_{ikk'}\alpha_{ik}\alpha_{ik'\ldots} + \delta_{i12 \ldots K_j^*} \prod_{k=1}^{K_j^*} \alpha_{ik}, \quad (2.66)$$

where δ_{i0} represents the intercept of item i, δ_{ik} is the weight of the main effect due to α_{ik}, and $\delta_{ikk'}$ is the weight of the interaction effect due to α_{ik} and $\alpha_{ik'}$. $\delta_{i12 \ldots K_j^*}$ is the weight of interaction effect due to $\alpha_{i1} * \alpha_{i2} * \ldots * \alpha_{iK_j^*}$. Specifically, the intercept δ_{i0} indicates the probability of a correct item response when any required attributes are not mastered. Next, the main effect of δ_{ik} represents the increased probability by adding each attribute, and the interaction effect of $\delta_{ikk'}$ as a first-order interaction represents the changed probability of a correct response by interaction of α_k and $\alpha_{k'}$. Lastly, $\delta_{i12 \ldots K_j^*}$ represent the interaction effect

occurred when all required attributes were mastered. The logit link function of the G-DINA model provides a similar form of equation as other log-linear CDMs. The logit link is represented as below:

$$\text{logit}[\text{P}(\alpha_i^*)] = \lambda_{i0} + \sum_{k=1}^{K_i^*} \lambda_{ik}\alpha_k + \sum_{k'=k+1}^{K_i^*} \sum_{k=1}^{K_i^*-1} \lambda_{ikk'}\alpha_k\alpha_{k'} \ldots + \lambda_{i12\ldots K_i^*} \prod_{k=1}^{K_i^*} \alpha_k. \quad (2.67)$$

Next, the log link function is represented as below:

$$\text{log}\text{P}(\alpha_i^*) = \upsilon_{i0} + \sum_{k=1}^{K_i^*} \upsilon_{ik}\alpha_k + \sum_{k'=k+1}^{K_i^*} \sum_{k=1}^{K_i^*-1} \upsilon_{ikk'}\alpha_k\alpha_{k'} \ldots + \upsilon_{i12\ldots K_i^*} \prod_{k=1}^{K_i^*} \alpha_k. \quad (2.68)$$

These three link functions are very similar except for the difference in the way that attribute mastery affects the probability of a correct item response (i.e., additive vs. multiplicative impact). Regardless of the type of functions, they represent that mastering only a few of the required attributes may increase the probability of a correct response. For parameter estimation of the G-DINA model, MMLE estimation was developed. For more details on special cases of CDMs by the G-DINA model, see de la Torre (2011).

Measurement of the psychometric properties of subscale scores

Measurement of Reliability

A reliable test allows precise measurement for the construct being measured. There are different approaches of defining test reliability across different measurement models. However, regardless of types of measurement models, reliability must assess the consistency of measurement so that test scores are trustworthy and precise. This section presents how each measurement model examines the concept of test reliability.

Measurement of Reliability in CTT

A test is considered to be reliable if the observed scores and the true scores are highly correlated. Therefore, CTT considers the squared correlation between the observed and true scores, ρ_{XT}^2, as an index of reliability in which the true scores and the observed scores are linearly related. The squared correlation between observed and true score, ρ_{XT}^2, can be also presented as the proportion of true score variance to the observed score variance, $\frac{\sigma_T^2}{\sigma_X^2}$. Also, the squared correlations between observed and true score are identical to the correlation between observed scores on two parallel tests, as proven in the statement *k)* in the section 3.1.1. Thus, reliability can be also expressed as the correlation between observed scores from two parallel tests, $\rho_{XX'}$. In CTT, parallel tests are assumed to have the same true score (i.e., $T = T'$) and the same error variance (i.e., $\sigma_E^2 = \sigma_{E'}^2$), and high correlations between two parallel scores can be evidence that the scores on a test or on its parallel test are reliable. The reliability can be also easily transformed as $1 - \frac{\sigma_E^2}{\sigma_X^2}$ (i.e., see the equation *h)* in the section 3.1.1). Note that given the same error variances in an examinee group, the size of the reliability depends on the variance of observed scores among examinees. The reliability will be large when the observed score variance is large, and it will get small when the observed score variance is small. Thus, reliability will be estimated more highly for a homogeneous group than for a heterogeneous group. However, ρ_{XT}^2 and $\rho_{XX'}$ are easily unobtainable because true scores are unobservable in general cases and it is difficult to create parallel tests. Therefore, they should be indirectly estimated. There are several primary ways for obtaining the reliability coefficients: test-retest reliability, parallel-forms reliability, and internal consistency.

Test-retest reliability is estimated by correlating two observed scores by administering all examinees with the same test twice. If an examinee takes the same test twice and receives the same scores in both testings for all examinees, the correlations will be as high as $r_{t_1 t_2} = 1.0$. Although this method seems to be reasonable and practically useful, it yields many types of carry-over effects. Carry-over effects can occur due to memory, practice, motivation, or maturation in cognitive ability, resulting in overestimating or underestimating the reliability.

A parallel-forms reliability is estimated by correlating an observed score from a test with an observed score from its parallel test. Although the high correlation between observed scores from two parallel tests means that the score on a test is more reliable, it depends on how parallel those two tests are. However, it does not seem possible for tests to be parallel. Instead, they are considered to be parallel when test must exhibit equal observed score means, variances, and also show similar correlations with other criterion measures. This method may yield better estimates than the test-retest reliability because it may reduce carry-over phenomena to some degree by removing memory and practice effects.

Internal consistency measures includes a split-half reliability, Coefficient α (i.e., Cronbach α) , and Kuder-Richardson formula 20 (KR20). These methods compute reliability coefficients based on a single testing occasion. For example, a split-half reliability coefficient is obtained by dividing a test into two parallel parts, and computing the correlation of observed scores from two divided parts. Note that this correlation is based on only half of the test. Because the reliability based on a shorter test is generally smaller than based on a total test, some corrections are needed to estimate the reliability of an entire test. The Spearman-Brown formula was developed to correct these reduced reliability estimates due to the changes in test length. Two versions of Spearman-Brown formula are written as follows:

$$\rho_{XX'} = \frac{2\rho_{YY'}}{1+\rho_{YY'}}, \text{ and} \tag{2.69}$$

$$\rho_{XX'} = \frac{N\rho_{YY'}}{1+(1-N)\rho_{YY'}}, \tag{2.70}$$

where $\rho_{XX'}$ is the corrected reliability based on the entire test, $\rho_{YY'}$ is the reliability based on the half test, and *N* is the number of parallel sections. The first equation is applied only when a test measures a split-half reliability, and the second equation can be generalized to the case when multiple components are existing for reliability estimation. These formulas can be applied only under the assumption that parallel tests are added to form a longer test. Indeed, if a test that is added is not parallel, the reliability could decrease. However, a longer test generally tends to yield high reliability. It occurs from the fact that true-score variance escalates faster than error variance as the number of parallel tests gets bigger. Thus, adding a test that is not parallel with the original test may overestimate the actual reliability. Special care is needed when one tries to apply the Spearman-Brown formula to estimate reliability. See Allen and Yen (2002) for more details. A split-half reliability is available only when split sections are assumed to be parallel. However, it is possible to estimate reliability coefficient in a situation when two split sections are not parallel. Coefficient α, also called Cronbach α, yields a reliability coefficient when the split halves essentially τ-equivalent (i.e., tests differ in their true score mean and observed score variances). Coefficient α for the split haves is formulated as follows:

$$Coefficient\ \alpha = \frac{2\left\{\sigma_X^2 - \left(\sigma_{Y_1}^2 + \sigma_{Y_2}^2\right)\right\}}{\sigma_X^2} = \frac{2\{2\sigma_{Y_1Y_2}^2\}}{\sigma_X^2}, \tag{2.71}$$

where σ_X^2 is the variance of the observed score on the entire test, X, and $\sigma_{Y_1}^2$ and $\sigma_{Y_2}^2$ are, respectively, the variances of the observed scores on split haves, Y_1 and Y_2. As shown in the equation, Coefficient α is the proportion of covariance between two split halves to the variance

of observed score on the entire test, multiplied by a constant, four. As the covariance between the split halves becomes larger, the Coefficient α gets larger. The split-half reliability and Coefficient α have the major advantage of being obtainable through only a single testing occasion. However, they may not appropriately estimate reliability when the split haves are not parallel or not essentially τ-equivalent. KR 20 is a generalized form of the Coefficient α, in which dichotomous items are involved, and formulated as follows:

$$KR20 = \left[\frac{N}{N-1}\right]\left[\frac{\sigma_X^2 - \sum_{i=1}^{N}\sigma_{Y_i}^2}{\sigma_X^2}\right] = \left[\frac{N}{N-1}\right]\left[\frac{\sigma_X^2 - \sum_{i=1}^{N} p_i q_i}{\sigma_X^2}\right], \text{ and}$$
$$q_i = 1 - p_i, \tag{2.72}$$

where N is the number of split tests, σ_X^2 is the variance of a total score, and $\sigma_{Y_i}^2$ is the variance of the *i*th split section of a test. When item responses are dichotomous, taking on only 0 or 1, the variance is same as the product of the proportion of examinees answering an item correctly (i.e., p_i) and the proportion of examinees missing an item (i.e., q_i). KR20 estimates reliability by dividing a test into more than two sections and correlating observed scores from the multiple sections. When correlations among sections or sets of items get are high, the reliability would be larger. As a result, KR20 produces higher reliability as the test includes homogeneous items.

Measurement of Reliability in IRT

In IRT, the measurement precision of a test score is determined by the amount of information that a test provides, which is the reciprocal of a variance of ability estimates, conditioned on θ. Thus, $SE(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}}$. Although the standard error of measurement in CTT is the same across different score levels, those in IRT vary at different trait levels. Thus, the error of measurement in CTT is given as a fixed value, and the size of errors in IRT are expressed by a function of θ fluctuating across different ability levels. An item that is more informative at a

specific ability level cannot be as informative at another level. Items are generally more informative when item discrimination, α, is high, item guessing, γ, is low, and item difficulty parameter, β, is close to the specific θ level. The item information functions for the 1-PL, 2-PL, and 3-PL are shown in the Table 2.2 below.

Table 2. 2. Item Information Functions for 1PL, 2PL, and 3PL IRT Models

Model	1-PL	2-PL	3-PL
$I(\theta)$	$Dp_i(\theta)q_i(\theta)$	$D\alpha_i^2 p_i(\theta)q_i(\theta)$	$D\alpha^2\left[\frac{q_i(\theta)}{p_i(\theta)}\right][\frac{p_i(\theta)-\gamma^2}{1-\gamma^2}]$

Test information function is defined by adding up all relevant item information functions as follows:

$$T(\theta) = \sum_{i=1}^{N} I_i(\theta), \qquad (2.73)$$

where N is the number of items and $I_i(\theta)$ is the amount of information for item *i* at the level of θ. Test information will increase as the number of items in a test increases, implying that a test with a high length would yield better measurement precision. In addition, as items in the test have better discriminating power among examinee and their difficulties are at the same or similar levels where examinees are located, the information function would be high and provide more precise measurement.

Assuming the normality of their distribution, ability estimates are distributed with the 95% confidence interval ranging from $\hat{\theta} - 1.96SE$ to $\hat{\theta} + 1.96SE$. Given θ, as information increases, the standard errors will be decreased, and the 95% confidence interval will be smaller. Hence, the accuracy of the estimate is increased. On the other hand, as information decreases, the standard errors will be increased, and the 95% confidence interval will be larger. Hence, the

accuracy of the estimate is decreased. This test information function can be used to measure the reliability of a test, but the information is given conditional on θ. That is, some test would be more precise for high trait levels rather than other levels and other would be more precise for low trait levels.

In IRT, small error of estimates indicates the accuracy of estimation, and improved accuracy of estimates increase the reliability of test scores. Andrich (1988) suggested that reliability can be computed for the sample using the average value of squared standard errors and the observed score variance (i.e., the variance of trait level estimates). Specifically, he formulated the reliability as shown below:

$$Empirical\ reliability = 1 - \frac{\sum_{n=1}^{N} \sigma_{error}^2}{N\sigma_{\theta}^2}, \qquad (2.73)$$

where *N* is the number of person, σ_{error}^2 is the squared standard errors of estimates, and σ_{θ}^2 is the variance of estimates across person. As the standard error of estimates is smaller, the reliability is higher.

Measurement of Reliability in CDA

The reliability of attribute estimates approaches two different questions: 1) if estimated attribute profiles and true attribute profiles correspond, and 2) two attribute profiles that are obtained from a test administered at two different time points are consistent. Although the measurement of the reliability is an important aspect of a test, the studies of reliability in diagnostic assessment are very rare. Henson, Roussos, and Templin (2004) measured reliability using multiple datasets simulated from the posterior distributions from an analysis. Specifically, they simulated datasets from the calibrated model, and estimated attribute profiles. These estimated attribute profiles were compared with the known true attribute profile. Otherwise,

estimates from two simulated datasets were compared with regard to the proportion that an examinee is classified into same attribute patterns. Templin and Bradshaw (2013) also discussed how reliability based on attribute estimates from diagnostic models can be measured. They also examined reliability using hypothetically repeated observations. Repeated observations were drawn from an acquired posterior distribution of a marginal attribute probability. Given two hypothetically identical tests, the marginal attribute probabilities from the two occasions should be hypothetically the same. Reliability is attainable by observing the difference of the probabilities that an examinee would get the same attribute mastery pattern estimates from the two occasions.

Measurement of Validity

The traditional concept of validity has been changed, and the newly defined concept of validity demands the establishment of the evidence of validity in items and the test as well as content validity and criterion-related validity. This section describes the changes in the concept of validity over several decades, and how a test can be developed to be valid in the perspective of the modern concept of validity.

Changes in the Concept of Validity

Cronbach and Meehl (1955) originally framed the concept of construct as "some postulated attribute of people assumed to be reflected in test performance" and argued that identifying the structural network through relationships between a test and other measures (i.e., nomothetic span) is the most vital consideration for construct validation. However, Embretson (1983) argued that construct validity must be also proven with direct evidence related to whether a test should measure what it intends to measure. Thus, she suggested that construct representation be considered another crucial component for demonstrating construct validity.

Construct representation was defined as a process of identifying the cognitive processes, strategies, and knowledge structure that underlie task or item performance. While the nomothetic span accounts for the communalities of all possible components between a test and other measures, construct representation focuses on a more explicit relationship between theoretical mechanisms (constructs) and task performance by modeling the impact of the constructs on performance. Thus, the process of the construct representation requires knowledge from cognitive psychology that can provide theoretical rationale on cognitive processes (e.g., working memory, logic), strategies, and knowledge structure related to a construct. Subsequently, Messick (1989) reformulated the traditional concept of validity by Cronbach and Meehl (1955) and emphasized the importance of the substantive evidence in validity. The traditional concept of validity is mainly divided it into three categories (i.e., content validity, criterion-related validity, and construct validity), but Messick (1989) redefined the traditional concept of validity by unifying the three categories to include six aspects of evidence of validity (i.e., content aspect, substantive aspect, structural aspect, generalizability aspect, external aspect, and consequential aspect). Then, he claimed that these all six aspects individually function in order to establish and support "construct validity" to some degree, defining construct validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment." However, above all things, he emphasized that understanding the substantive structure underlying test performance should be a key factor for construct validity. These changes in the concept of validity suggested by Embretson (1983) and Messick (1989) highlighted the role of cognitive psychology in the validation procedure, test design, and test interpretation, attracting the interest and attention of many researchers.

Development of New Test Design Systems

Unintended constructs in a test can decrease test validity because the test does not measure the intended constructs. Once a test is developed, it seems of importance to ensure that test items are designed based on intended constructs. In such meaning, the following two test design systems aid in the establishment of test validity.

Cognitive Design System

Cognitive design system (CDS; Embretson, 1998) is a test development framework that centralizes cognitive theory in a test design. CDS is processed within two separate frameworks: Conceptual framework vs. Procedural framework. The conceptual framework deals with the expanded concept of construct validation, and the procedural framework explains a series of stages in which cognitive theory is grafted.

The conceptual framework focuses on the expanded concept of construct validation including both nomothetic span and construct representation. Nomothetic span and construct representation concerns different aspects of test scores (see Embretson, 1983). Specifically, nomothetic span concerns the significance of test scores through empirical relationship among measures, whereas construct representation primarily concerns their meaning by identifying underlying cognitive constructs for item solving. Because in construct representation phase, the impacts of cognitive variables involved in items on item properties are revealed, it is possible to freely manipulate item properties by including or removing specific cognitive variables. Also, although nomothetic span does not seem to provide direct rationale for the meaning of test scores, it contributes to establishing the meaning of test scores by the nomological network to some level. Thus, the conceptual framework emphasizes considering both nomothetic span and construct representation in designing CDA.

The procedural framework specifies the overall development procedures of CDA, focusing on the role of cognitive psychology in each stage. This framework consists of seven stages involving a) specifying measurement goals, b) identifying design features in the task domains, c) developing a cognitive model, d) generating items, e) evaluating models, f) creating an item bank by cognitive complexity, and g) test validation.

Specifying the goals of measurement refers to describing the purpose of measurement. The purpose includes exactly what specific cognitive variables would be measured and how the relationships among the variables look like. The next stage, identifying design features, concern the decision of item features (e.g., mode, format, conditions, etc.). Because item features can affect cognitive processes, strategies, and knowledge structures for item-solving, knowledge for detecting these item features influencing item-solving is required. Then, cognitive theory models are determined and may be applied to available items. Cognitive models provide theory or knowledge on the processes underlying item solving. They may be obtained from the literature review of cognitive psychology or by altering the existing cognitive theory fitting the item type chosen. The chosen models are evaluated with regard to their plausibility with available test items using the overall fit test of a mathematical model. At this point, the impact of each stimulus feature on item properties can be examined as well. Based on the evaluation of the cognitive model, an item is generated through determination of regarding whether specific features would be included or not. Generating items requires work for operationalizing cognitive processing variables into stimulus features involving items as identifying item structures and substitution rules. Then, cognitive models are again tested with data from items developed. In this stage, the impact of cognitive variables on item discrimination, item difficulty, or response time is predicted. The results are utilized in item banking. An item bank can be easily created by

differing the cognitive processing demand across items. For example, if multiple stimulus features demanding similar or equal cognitive process are found, using many stimulus features within the same structure would not largely change item properties. In the stage of validation, nomothetic span is established. Nomothetic span must be supported by evidence of construct representation which is obtained in the previous stages. In this procedure, correlations with other tests are collected to ensure that the new test is valid. It should be confirmed that measures being compared represent similar cognitive processing demand. The applications of CDS in assessment development are illustrated in the development of the abstract reasoning test (Embretson, 1998), the spatial learning ability test (SLAT; Embretson, 1994), and the standardized mathematics achievement tests for middle school students (Embretson, 2014).

Evidence Centered Design

Similar to the CDS, evidence centered design (ECD; Mislevy, 1994; Mislevy, Almond, & Steinberg, 2003) provides the design framework for cognitive assessment development. The ECD focuses on maximally accumulating evidence to help inferences about an individual. Specifically, the goal of the ECD is to help test developers in designing a test, creating items, and reporting scores with the purpose of the tests so that they can appropriately make reasoning of what an individual really knows, can achieve, and can do in the real life. The ECD specifies five different layers pertaining to test design, task development, and score reporting: 1) domain analysis, 2) domain modeling, 3) conceptual assessment framework, 4) assessment implementation, and 5) assessment delivery. The first two layers are relevant to test designs, and the other three layers are related to task development and scoring or score reporting.

The domain analysis layer concerns the comprehensive investigation of contents or subjects to be assessed. In this layer, information of the concept, terminology, and knowledge

related to a domain, and information of how the domain is applied in a real life are identified. The domain modeling layer organizes information from the domain analysis and creates assessment arguments in narrative form. Assessment argument includes three components: 1) the claim that one wants to make about an individual, 2) the data that are evidence that can support these claims, and 3) the warrant that is rationale of how particular data can be connected with particular claims. Domain experts, teachers, and assessment specialists cooperate to find specific attributes seen as claims, data, and warrant components. Next, the conceptual assessment framework layer concerns technical specification for designing a task. In this layer, various components are formalized in terms of student model, task model, and evidence model. The student model, also called the proficiency model, specifies what an assessment designer is trying to measure and make inference about an individual. The student model identifies variables reflecting knowledge, skills, or abilities that an individual might have. The task model concerns the forms in which an inference of a student performance (i.e., what a student say, do, or make) would be made and describes the important features of task materials and the presentation methods. The evidence model focuses on verifying the link of the task model and the student model. The evidence model is processed by two reasoning steps: Evaluation and measurement modeling. Evaluation involves how one identifies and evaluates student performance. For example, whether automated scoring procedures would be used or whether other methods should be considered are determined in this step. The measurement modeling steps concern the considerations of measurement models for dealing with task responses. Next, the assessment implementation layer is much related to item writing and the assembly of test forms in traditional test development. Although tasks are generated based on the task model in the conceptual assessment framework layer, they require additional analysis and preparation for

implementation. Observing model fits with pilot test data is also an activity relevant to this layer. Lastly, the assessment delivery layer concerns test administration, evaluation of test performance, and feedback reports. The Cisco System's Networking Performance Skill System (NetPass; Behrens, Mislevy, Bauer, Williamson, & Levy, 2004) used ECD for designing simulation-based learning and assessment for training engineers.

CHAPTER 3

REAL-WORLD DATA STUDY

Real-world data from math achievement tests were used to obtain subscale scores based on different psychometric methods. The first section includes detailed information about examinees and test material from which the real-world data are drawn. Next, subscale scores based on different methods are computed and their reliabilites are compared among subscale scores. Results are interpreted and summarized.

Method

Examinees

Approximately, 5,000 examinee response data were randomly sampled from 33,000 Grade 8 students in a Midwestern state who administered a math test. All responses were recoded into two categories, zero for wrong answers and one for right answers.

Testing Materials

The Grade 8 math achievement tests were designed for the purpose of measuring examinees' general math skills and evaluating students based on their achievement goal in the math area. The math test consists of 71 items with four answer options. Items were originally written based on a test blueprint from the state, in which mathematical contents, skills, or knowledge that Grade 8 students have to accomplish are specified. The test blueprint represents a hierarchical structure in which three different levels exist: Standards, Benchmarks, and Indicators. Figure 3.1 illustrates the hierarchical structure of Standards, Benchmarks, and Indicators. The four Standards represent Number and Computation, Algebra, Geometry, and

Data. Each standard includes two or more specific benchmarks; in turn, each benchmark involves one or two indicator skills, although they are not specified in the figure. For the current study, only the four Standards were employed to define the subscales. Thus, four subscale scores were available. The test has 23 Number and Computation, 17 Algebra, 17 Geometry, and 14 Data items with a total of 71 items. Each item was relevant to one out of four standards.

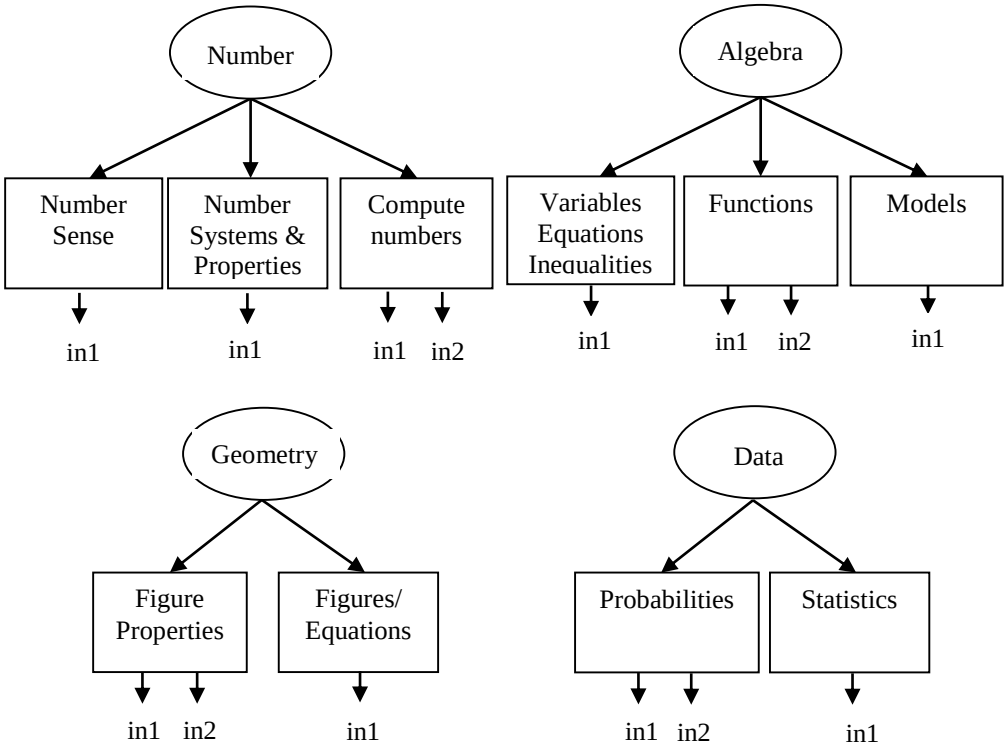

Figure 3.1. Grade 8 Math Test Blueprint with Hierarchical Structure

Real-World Data Analysis and Results

The current study employs seven different psychometric models based on CTT and IRT scaling frameworks. CTT-based methods include raw subscale scores, Kelley's regressed method, Holland and Hoskens' method, and Haberman's weighted average method, and IRT-based methods include unidimensional IRT model, the objective performance index (OPI), and multidimensional 2PL model. For understanding data structure, descriptive statistics and

reliability values are computed with total items and subscale items. Then, correlations between observed subscale scores and observed total scores are calculated. CTT-based subscale scores are computed and compared regarding the reliability of estimation with the RMSE-MBs and the PRMSE-MBs. IRT-based subscale scores are first compared in terms of goodness-of-fit index for evaluating estimation accuracy. In all procedures, computer software of SPSS 22.0 and FlexMirt Version 2 (Cai, 2013) are used.

Descriptive Statistics

Table 3.1 presents the means, the standard deviations, KR-20, and correlations between observed subscale scores and the total score. All statistics were based on the item responses of 4,959 examinee. The total number of items was 71, and the number of items on each subscale, Number, Algebra, Geometry, and Data, was 23, 17, 17, and 14, respectively. The average proportion passing for the 71 items was 0.71, showing that items were moderately easy. The average proportion passing for the standards was 0.69 for the Number subscale, 0.74 for the Algebra subscale, 0.71 for the Geometry subscale, and 0.69 for the Data subscale. From the results, students seem to answer Algebra items somewhat more accurately than items in the other standards. The standard deviation was the largest in the Number subscale which had the highest number of items, while the smallest standard deviation was for the Data subscale in which the least number of items are included, as expected. The reliabilities of subscale scores and total score were computed by KR-20 index. Specifically, reliability based on all 71 items was 0.92, while those based on subscales ranged between 0.74 and 0.76. Although the sizes of subscale score reliabilities are still acceptable (i.e., reliable), they dropped by a considerable degree compared to the total items. The correlations of subscale scores with total scores were very high,

ranging from $r = 0.84$ to $r = 0.90$. From these results, it is possible that additional information from total score can yield better estimates for each subscale score.

Table 3. 1. Summary Statistics for a Math Test Scores Based on the CTT Model

Subscale	# of items	Mean(p)	SD	KR-20	Correlation with total score
Number	23	15.76 (0.69)	4.06	0.75	0.90
Algebra	17	12.68 (0.74)	3.27	0.76	0.85
Geometry	17	12.04(0.71)	3.46	0.76	0.89
Data	14	9.69(0.69)	3.02	0.74	0.84
Total	71	50.17(0.71)	12.07	0.92	1.00

Similar statistical results were obtained from subscale θs. Table 3.2 presents the means and the standard deviations for scale subscaores, empirical reliability, and the correlations among subscale score θs. These subscale scores θs were estimated based on the responses of the corresponding subscale items. The means of the subscale θs were close to zero, and the standard deviations were ranged between 0.29 and 0.51. The standard deviation of θs was the smallest in the total test. As expected, the empirical reliability was the highest for the whole test, and was reduced for the subscale scores θs. Correlations between subscale scores and total scores were usually high, similar to the results of the CTT based scores. However, the number subscale scores showed different patterns: high correlations based on the CTT model and moderate correlations based on the IRT model.

Table 3.2. Summary Statistics for a Math Test Scores on the IRT Model

Subscale	# of items	Mean(p)	SD	Empirical reliability	Correlation between θs
Number	23	0.02	0.46	0.78	0.76
Algebra	17	0.00	0.48	0.76	0.83
Geometry	17	-0.01	0.46	0.78	0.87
Data	14	0.00	0.51	0.74	0.81
Total	71	0.09	0.29	0.92	1.00

Correlation Structures

Correlations among subscale scores can be examined. If correlations among subscale scores are too high, it may not be reasonable to yield and report subscale scores, because it cannot provide additional information over a total test score. Table 3.3 presents the correlations among subscale scores for raw observed subscale scores based on CTT. Correlations between subscale scores ranged from 0.59 to 0.71, presenting moderately high correlations.

Table 3.3. Correlations among Raw Subscale Scores

	Subscale			
	Number	Algebra	Geometry	Data
Number	1.00			
Algebra	0.66	1.00		
Geometry	0.71	0.71	1.00	
Data	0.70	0.59	0.69	1.00

Similar results were found in the correlations among Subscale scores θs. In specific, correlations between subscale scores θs ranged between 0.53 and 0.69, which are rather smaller than those among raw subscale scores. The correlations structure among subscale θs are shown in Table 3.4.

Table 3.4. Correlations among Subscale Score $\boldsymbol{\theta}$s from the Unidimensional 2PL model

	Subscale			
	Number	Algebra	Geometry	Data
Number	1.00			
Algebra	0.53	1.00		
Geometry	0.57	0.69	1.00	
Data	0.63	0.57	0.64	1.00

Subscale score Estimates and their Reliability

CTT-based Subscale Scores

Various psychometric methods for overcoming the reduced reliability in the shorter length subtests were introduced in the previous sections. For computing the subscale scores in the math achievement subscales, three CTT-based methods, using the regression technique, were employed. These methods approximate true subscale scores by using one of the following predictors: 1) the observed subscale score, 2) the observed total score, and 3) the combination of the observed subscale score and the total score.

In these methods, regression coefficients and intercepts are calculated using information of the means, the SDs, the reliability measures, the correlations of observed subscale scores and observed total scores). Then, resulting regression equations are used to compute subscale scores. See Chapter 2 for the detailed information about procedures of computing the regression coefficient. Table 3.5 presents descriptive statistics for the raw subscale scores and the estimated true subscale scores from math achievement data. The means of all the raw subscale scores and the estimated true subscale scores remain equal, but the standard deviations differ across subscale scores. In particular, raw subscale scores had the largest standard deviations, and the true subscale scores approximated by corresponding observed subscale scores had the smallest standard deviations. Using the regression technique for approximating the true subscale scores general generally decreased the variability among examinee subscale scores.

Table 3. 5. Descriptive Statistics for Estimated True Subscale Scores

		Type of predictors (Method)									
		No predictors		Observed subscale score		Total score		Subscale score & Total score		OPI	
Subscale	Subscale length	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Number	23	15.85	4.06	15.85	3.05	15.85	3.30	15.85	3.30	16.79	3.91
Algebra	17	12.72	3.28	12.72	2.48	12.72	2.56	12.72	2.63	12.92	3.23
Geometry	17	12.09	3.48	12.09	2.66	12.09	2.86	12.09	2.87	12.07	3.4
Data	14	9.71	3.02	9.71	2.23	9.71	2.34	9.71	2.39	9.38	2.38
Total	71	50.36	12.05	50.36	9.08	50.36	11.06	50.36	11.07	-	-

Each subscale scoring method differs in the degree to which it increases the reliability of subscale scores. Table 3.6 presents the standard error of measurement from the observed subscale scores and the amount of errors (e.g., RMSE-MB) from approximations for true subscale scores using different predictors. The root mean squared error (RMSE-MB) is obtained by the average of the squared difference of the observed subscale scores and the predicted subscale scores. See Table 2.1 in Chapter 2 for the details of the RMSE-MB computations

Table 3.6. Root Mean Squared Errors for Approximations for Estimated True Subscale Scores

Subscale	$\sigma(e_x)$	$\mathrm{RMSE-MB}_{Kelley}$ $\sigma(R(\tau_s \mid S_s)$	$\mathrm{RMSE-MB}_{HH}$ $\sigma(R(\tau_s \mid S_T)$	$\mathrm{RMSE-MB}_{Haberman}$ $\sigma(R(\tau_s \mid S_s, S_T)$
Number	2.02	1.75	1.22	0.84
Algebra	1.62	1.41	1.27	1.01
Geometry	1.69	1.47	1.03	0.69
Data	1.54	1.32	1.10	0.83

The terms used in the table above, S_s and S_T are the observed subscale score and the observed total score from the math data, respectively, and τ_s is the estimated true subscale score. $\sigma(e_x)$ is the term of the standard error of measurement for raw subscale scores. In turn, $\sigma(R(\tau_s|S_s)$, $\sigma(R(\tau_s|S_T)$, and $\sigma(R(\tau_s|S_s, S_T)$ are, respectively, the RMSE-MB values from the approximation of true subscale scores by predictors: observed subscale scores, observed total scores, and the combination of observed subscale scores and observed total scores. The standard error of measurements for raw subscale scores were always larger than the RMSE-MB values from approximations by predictors, regardless of the kinds of predictors. Number subscale scores showed the highest standard error of measurement, and Data subscale scores scores showed the lowest standard error of measurement. The smaller RMSE-MB was obtained when the true subscale scores were approximated by the observed total score rather than the observed subscale scores or the combination of observed subscale scores and observed total scores. The largest

decrease in errors was found in Number, and the smallest decrease in errors was found in Algebra. In the meantime, using the linear regression of the true subscale score on the observed subscale score and the combination with the observed total score produced only a slight reduction in errors, in which these predictors led reduction in errors to the similar degree. Generally, subscale scores obtained by regression techniques seem to be stable and accurate than raw subscale scores.

Table 3.7 presents the proportional reduction in MSE from different subscale scores. The PRMSE-MBs were used as an indicator for the amount of added-value. As described earlier, the PRMSE-MB measures the proportion of MSE reduced by using a predictor relative to a standard value. Here, the standard value is the MSE of mean observed subscale scores, which is obtained by approximating the true subscale score by the expected value of observed subscale scores across examinees approximates the true subscale score. The higher the PRMSE-MB, the more added-value the corresponding subscale scores have. Results show that PRMSE-MB values were largest in Haberman subscale scores, and smallest in Kelley's regressed subscale scores. See Table 2.1 in Chapter 2 for more detailed descriptions of PRMSE-MB computations

Table 3.7. Proportional Reduction in Mean Squared Errors for Four Math Subscale Scores

Subscale	$PRMSE - MB_{Kelley}$ $\sigma(R(\tau_s \mid S_s)$	$PRMSE - MB_{HH}$ $\sigma(R(\tau_s \mid S_T)$	$PRMS - MB_{Haberman}$ $\sigma(R(\tau_s \mid S_s, S_T)$
Number	0.75	0.87	0.88
Algebra	0.76	0.80	0.85
Geometry	0.76	0.88	0.89
Data	0.74	0.82	0.85

Table 3.8 shows the partial regression coefficients used in Haberman's method. $\beta(\tau_s \mid S_s \cdot S_T)$ is the partial regression coefficient of the true subscale score on the observed subscale score given the observed total score, and $\beta(\tau_s \mid S_T \cdot S_s)$ is the partial regression of the

true subscale score on the observed total score given the observed subscale score. These indicate the relative strength of weights by using predictors of the observed total score or the observed subscale score. In the results, the observed total score seems to have more influence on estimating the true subscale score than the observed subscale scores. A linear regression of true subscale score on both observed subscale score and observed total score has high weight of the total score and seems to provide better predictor of true subscale score than observed subscale scores.

Table 3.8. Partial Correlation Coefficients of Four Math Subscales

Subscale	$\beta(\tau_s \mid S_s \cdot S_T)$	$\beta(\tau_s \mid S_T \cdot S_s)$
Number	0.12	0.24
Algebra	0.34	0.13
Geometry	0.15	0.20
Data	0.29	0.13

IRT-based subscale scores

Prior to the subscale scoring, overall fits for multiple models were compared. The given dataset was analyzed using multiple IRT models: unidimensional 1PL, 2PL, 3PL, and the multidimensional 2PL model. In all models, the MML-EM method for item parameters and the EAP method for person parameter estimation were used, with fifteen quadrature points. Table 3.9 presents the overall goodness-of-fit statistics, Akaike Information Criterion (AIC) and the -2loglikelihood (-2lnL). In both indices, the smaller the values, the better fit the model. The statistical significance of fit difference can be also examined through the -2lnL difference because the difference of -2lnL is considered to be asymptotically distributed chi-square with the difference of the degrees of freedom. From the table below, the -2lnL values were obtained through the comparisons with 1PLM. The best fitting model was the 3PLM with Δ-2lnL of 4611.86. The multidimensional 2PLM showed better fit than the 1PLM, but poorer fit than the

2PLM and the 3PLM, supporting the unidimensionality of math data. From the results, applying multidimensional 2PLM to the math data seems to be inappropriate.

Table 3.9. Overall Goodness-of-Fit Comparison among IRT Models from a Math Test

Type of Models (# of parameter)	AIC	-2lnL	Δ-2lnL	Comparing models
1PL (72)	370,757.3	370,613.3		
2PL (142)	369,786.7	369,502.7	1629.0*	1PLvs.2PL
3PL (213)	366,427.4	366,001.4	2982.9*	2PLvs.3PL
MIRT-2PL (148)	369,273.8	368,977.8	6.5	2PLvs.MIRT-2PL

Summary and Discussion

The structure of data based on raw subscale scores was examined. The reliabilities of raw subscale scores were quite smaller than that of the total test score. Averaged correlation between subscale scores and total test scores was 0.82, indicating the possibility that subscales and total test may measure quite similar constructs. From the comparisons of RMSE-MBs of subscale scores, all three methods, Kelley's, HH's, and Haberman's methods, yielded lower RMSE-MBs than raw subscale scores, and the lowest RMSE-MBs were found in Haberman's method, having the highest accuracy of true subscale score prediction.

However, PRMSE-MBs from the Haberman's method showed that subscale scores did not provide much improvement relative to the PRMSE-MBs from the HH method, indicating that subscale scores from the Haberman's method do not give added-value over the total scores. The results from IRT-based subscale scores also supported the CTT-based results. In order to determine whether subscale scores are valid for reporting, dimensionality of data was examined through the comparisons of overall goodness-of-fit from unidimensional models and a multidimensional 2PL model. The results of Δ-2lnL showed that the multidimensional 2PL model did not show significantly better fit than the unidimensional 2PL. That is, the

multidimensionality of the math test was not supported, indicating that subscale scores are not appropriate for reporting.

CHAPTER 4

SIMULATION DATA STUDY

This chapter begins with the description of simulation procedures. Simulated data, varied in 1) subscale lengths, 2) the amount of subscale consistency, 3) between-subscales correlations, and 4) test types, are used to estimate subscale scores, using seven different psychometric methods. The resulting scores are evaluated with respect to their accuracy based on several criteria. The criteria include measurement-based root mean square error (RMSE-MB; Haberman, 2008), simulation-based root mean square error (RMSE-SB), and correlation between estimated and true subscale scores.

Simulation Procedures

Data Generation

Data were simulated with various conditions under the multidimensional 2PL IRT model (MIRT-2PL) that is one of the most complex models in the study, using a SAS 9.4 macro. The number of subscales in all tests was fixed to four. In the MIRT model, simple structures were assumed; thus, each item was loaded on a single dimension. True trait level scores, θs, for four dimensions were generated for 3,000 samples. The scale of measurement for the MIRT models was set by fixing the means and the variances of θs as 0 and 1, respectively. Four subscales scores (i.e., the θs for the four dimensions) were distributed from a multinormal population distribution with the mean of (0, 0, 0, 0) and the variance of (1, 1, 1, 1), $\theta_i \sim MVN(0, \sum)$. The off-diagonal elements in the variance-covariance matrix were set to vary, depending on the correlations defined in the specific condition.

Varied Simulation Conditions

Simulation conditions were varied in subscale lengths, test types, between-subscales correlations, and subscale consistency. The simulation was designed to meet several goals of the study: 1) to understand the impacts of subscale length, correlation among subscales, subscale consistency, and item difficulty level on subscale score estimation, and 2) to demonstrate the accuracy of subscale score estimation under various data conditions based on various psychometric models.

First, the Subscale Length condition was defined by the number of items within each subscale. Tests with 10 and 20 items per subscale were simulated. Tests with 10 items within a subscale, I = 10, are generally expected to be less reliable or less accurate than those with 20 items, I = 20, although the impact of the subscale length may differ somewhat across different subscale scoring methods.

Second, the Test Type condition was defined by item difficulty values that are typical of ability vs. achievement tests. By considering practical testing situations in which achievement tests are relatively easy and ability tests are difficult, two different item difficulty sets were generated. Specifically, items in the ability test type were randomly generated from a normal distribution of ~N (0.0, 0.5), which correspond to a mean p-value of 0.5, and those in the achievement test type were obtained from ~N (-1.2, 0.5), which correspond to a mean p-value of 0.7.

Third, the Between-Subscales Correlation condition was defined by the correlations between subscale scores in a test. Three different between-subscales correlation conditions of r = [0.3, 0.6, 0.9] were simulated, which, respectively, correspond to low, medium, and high

correlation level. Correlations between any two subscales within a test were set to be equal as below:

$$\Sigma = \begin{bmatrix} 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 \end{bmatrix}, \ \Sigma = \begin{bmatrix} 1 & 0.6 & 0.6 & 0.6 \\ 0.6 & 1 & 0.6 & 0.6 \\ 0.6 & 0.6 & 1 & 0.6 \\ 0.6 & 0.6 & 0.6 & 1 \end{bmatrix}, \ \Sigma = \begin{bmatrix} 1 & 0.9 & 0.9 & 0.9 \\ 0.9 & 1 & 0.9 & 0.9 \\ 0.9 & 0.9 & 1 & 0.9 \\ 0.9 & 0.9 & 0.9 & 1 \end{bmatrix}.$$

Fourth, the Subscale Consistency condition was defined by consistency of responses within a subscale. High subscale consistency indicates that there are high correlations among item responses within the subscale. On the other hand, low subscale consistency indicates low correlations among items within the subscale. Two different subscale consistency conditions, high vs. low, were simulated. Because subscale consistency is manipulable by the amount of item discrimination, high and low subscale consistency conditions were generated by simulating items with high and low discrimination values, respectively. Specifically, high item discrimination value sets were generated from a log normal distribution of ~*lnN* (0.0, 0.03), whose item discrimination value mean was 1.2. In turn, low discrimination value sets were generated from ~*lnN* (-0.2, 0.08), and their mean discrimination was 0.8. High subscale consistency is expected to increase the score reliability in a subscale, whereas low subscale consistency is expected to decrease the subscale score reliability.

These four simulation conditions described above yield a total of 24 conditions (i.e., 2 Subscale Length x 2 Test Type x 3 Between-Subscales Correlation x 2 Subscale Consistency = 24). Each condition was repeated with 100 replications. Thus, a total of 2,400 datasets were generated. Table 4.1 below presents all possible study conditions.

Table 4.1. Simulation Study Conditions

Test type	Subscale length	Subscale consistency	Between-subscales correlation		
			r=0.3	*r*=0.6	*r*=0.9
Achievement	I=10	High	x	x	x

		Low	x	x	x
	I=20	High	x	x	x
		Low	x	x	x
Ability	I=10	High	x	x	x
		Low	x	x	x
	I=20	High	x	x	x
		Low	x	x	x

Analysis of Simulated Data

Simulated data were used to estimate subscale scores. For the computation of the subscale scores, seven subscale scoring methods, including raw subscale scoring, Kelley's, Holland-Hoskens', and Haberman's regressed subscale scoring, unidimensional 2PL, multidimensional 2PL, and OPI, were used. The accuracy of subscale scores from each method was evaluated according to the root mean square errors (RMSE) and correlations of estimated subscale scores with their true scores. Two different types of RMSE values, one based on true scores and the other based on observed scores, are available. Throughout the paper, the RMSEs based on true scores in the simulation are termed as RMSE-SB, and the RMSEs based on observed scores are termed as RMSE-MB.

RMSE-SB can be defined as rooted mean of squared deviations between estimated scores and their true scores. RMSE-SB is the square root of the averaged squared deviations between estimated and true scores across a sample. The equation of RMSE-SB for subscale *j* can be written as follows:

$$RMSE - SB_j = \sqrt{\frac{\sum_{n=1}^{N}(\theta_{ij} - \theta'_{ij})^2}{N}}, \qquad (4.1)$$

where *N* is the total number of examinees, θ_{ij} and θ'_{ij} are, respectively, the true trait score and the estimated score in subscale *j* of examinee *i*.

RMSE-MB is a method of evaluating the reliability of subscale scores in CTT methods that Haberman (2008) suggested. It is available by obtaining the deviations between the observed subscale scores and the estimated subscale scores from a model. See Chapter 2 for more details.

Results of Simulated Data

Descriptive Statistics on Simulations

Descriptives on the simulation were computed for two reasons. First, the simulations were analyzed to determine the adequacy of the parameter specifications, such as having the predicted impact on descriptive statistics. Second, the simulations were analyzed to determine the plausibility of the overall properties of a test based on the specifications within each condition. The results will be presented in two sections: true item and person parameters, and summary statistics of simulated item responses.

True Item and Person Parameters

Item and person parameters were randomly sampled from a specified distribution, as described above. Table 4.2 below shows the resulting means and standard deviations of true item and true person parameters of datasets under 24 different conditions (i.e., 100 datasets for each condition). All means and standard deviations were averaged over 100 replication data. In Achievement Test Type condition, the means of item difficulty values were mostly -0.90, indicating that item difficulties are set to be easy, as expected. In turn, in Ability Test Type condition, item difficulty means were -0.01, which is somewhat higher than those in achievement test type condition. The standard deviations of item difficulty were around 0.50 across all test type conditions. Also, the means of item discrimination values were 1.20 in the High Subscale Consistency condition, and 0.80 in Low Subscale Consistency condition, as expected. Their

standard values were approximately 0.1, showing that discrimination values were generated in a narrow range, as intended. Each Subscale Length condition was represented in the simulated data, as well.

Table 4.2. Means and Standard Deviations for True Item Parameters of Simulated Data

Data condition				IRT Item parameters			
				α		β	
Test type	Subscale length	Subscale consistency	Correlations between θs	Mean	SD	mean	SD
Achievement	I=10	High	0.3	1.20	0.10	-0.90	0.50
			0.6	1.20	0.10	-0.90	0.50
			0.9	1.20	0.10	-0.89	0.50
		Low	0.3	0.80	0.08	-0.90	0.50
			0.6	0.80	0.08	-0.90	0.50
			0.9	0.80	0.08	-0.90	0.50
	I=20	High	0.3	1.20	0.10	-0.90	0.50
			0.6	1.20	0.10	-0.90	0.51
			0.9	1.20	0.10	-0.90	0.50
		Low	0.3	0.80	0.08	-0.90	0.50
			0.6	0.80	0.08	-0.90	0.50
			0.9	0.80	0.08	-0.89	0.50
Ability	I=10	High	0.3	1.20	0.10	-0.01	0.51
			0.6	1.20	0.10	-0.01	0.50
			0.9	1.20	0.10	-0.01	0.51
		Low	0.3	0.80	0.08	-0.01	0.50
			0.6	0.80	0.08	-0.01	0.50
			0.9	0.80	0.08	0.00	0.50
	I=20	High	0.3	1.20	0.10	-0.01	0.49
			0.6	1.20	0.10	-0.00	0.50
			0.9	1.20	0.10	0.00	0.49
		Low	0.3	0.80	0.08	0.01	0.50
			0.6	0.80	0.08	0.01	0.50
			0.9	0.80	0.08	0.01	0.49

True person parameters, θs, were generated with means and standard deviations of approximately 0.0 and 1.0 in all conditions. The results of correlations among subscale θs, although not shown in the table, showed that Between-subscale Correlation conditions were well represented in simulated data, as intended. Table A1 presents the resulting means and standard

deviations of subscale score, θs, in the simulation conditions. Overall, the simulated data seem to appropriately represent test properties as defined in the specification of each condition.

Summary Statistics of Simulated Item Responses

Summary statistics based on CTT and IRT were computed with the simulated data. The results include raw subscale scores from CTT and IRT, their standard deviations, subscale score based KR-20, and their correlation with the total test scores. Especially, the impact of specified conditions on reliability of subscale scores will be discussed.

CTT-based Descriptive Statistics

The descriptive statistics of item responses were computed from 3,000 examinee data simulated in each condition and were averaged over 100 replication data. The summary statistics for two Test Types, Achievement and Ability, are, respectively, shown in Table A2 and A3. The tables include the means, the standard deviations, and KR-20 values for four subscales and a total test score for each condition. Correlations between scores from a total test and the corresponding subscale are also included in the tables. Table 4.3 below includes the summary of KR-20 means within tests across four raw subscale scores.

The means of raw subscale scores in the Achievement Test Type condition were higher than those in the Ability Test Type condition. Specifically, *p*-values ranged from 0.65 to 0.69 in the achievement test condition, but from 0.50 to 0.51 in the ability test condition, as specified in the simulation conditions of Achievement and Ability Tests.

KR-20 values were computed for the reliabilities, and the results show that KR-20 values broadly ranged between 0.55 and 0.84 across different simulation conditions. Different test types showed small differences in the KR-20 values of subscale scores. Specifically, the mean KR-20 in subscale scores was on average, 0.71 in the Achievement test, and 0.72 in the Ability test,

showing slightly higher reliabilities in the Ability test. The greatest differences in subscale score reliability were found in different Subscale Length and Subscale Consistency conditions.

KR-20 values were larger when Subscale Length is I = 20, rather than I = 10, and Subscale Consistency is High rather than Low. Specifically, KR-20 values of subscale scores were on average, 0.78 in I = 20 Subscale Length condition, and 0.65 in the I = 0. In turn, KR-20 values were greater in High Subscale Consistency condition than in Low Subscale Consistency condition. KR-20 values of subscale scores were on average, 0.78 in High Subscale Consistency condition, and 0.65 in Low Subscale Consistency condition. Also, the amount of reliability in different Subscale Consistency conditions largely differed depending on the length of the subscale. In the I = 10 Subscale Length condition, KR-20 values from subscale scores were averagely 0.72 in the High Subscale Consistency condition and 0.57 in the Low Subscale Consistency condition, representing large difference between High vs. Low Subscale Consistency conditions. In the I = 20 Subscale Length condition, KR-20 values were averagely 0.83 in the High Subscale Consistency condition, and 0.72 in the Low Subscale Consistency condition, also showing large difference between two Subscale Consistency conditions, although the amount of the difference was smaller than in the I = 10 Subscale Length condition. This presenting higher reliability and less difference across Subscale Consistency conditions in I = 20.

Table 4.3. KR-20 Means within Tests for the Four Raw Subscale Scores

Data condition			KR-20 mean	
Subscale length	Subscale consistency	Between-subscales correlation	Achievement	Ability
I = 10	High	0.3	0.71	0.73
		0.6	0.71	0.73
		0.9	0.71	0.73
	Low	0.3	0.56	0.58
		0.6	0.56	0.58
		0.9	0.56	0.58

I = 20	High	0.3	0.83	0.84
		0.6	0.83	0.84
		0.9	0.83	0.84
	Low	0.3	0.72	0.73
		0.6	0.72	0.73
		0.9	0.72	0.73

KR-20 values of raw subscale scores were the same across all three Between-subscales Correlation conditions, $r = 0.3$, $r = 0.6$, and $r = 0.9$, which shows that Between-subscales Correlation conditions do not have a direct impact on the reliability of raw subscale score across various Between-subscales Correlation conditions. Tables A6 and A7 present correlations among these raw subscale scores, respectively from achievement and ability tests. Specifically, the correlations among raw subscale scores ranged between 0.16 and 0.24 where $r = 0.3$ in the Between-subscales Correlation condition, between 0.33 and 0.49, where $r = 0.6$ condition, and between 0.50 and 0.74, where $r = 0.9$ for both achievement and ability tests.

However, the amount of Between-subscales Correlation had a positive relationship with that of correlations between subscale scores and their total score. Specifically, in the Between-subscales Correlation condition of $r = 0.3$, 0.6 and 0.9, the average subscale-total score correlations were, respectively, 0.64, 0.75, and 0.85, showing that as correlations among subscale scores get larger, correlation between subscale and total scores became greater. In turn, subscale-total score correlations were pertinent to the size of reliability of the total test. As the subscale-total score correlations were high, the reliability of the total test was high, and as the subscale-total score correlations are low, the reliability of the total test was relatively low. Specifically, in the Between-subscales Correlation condition of $r = 0.3$, KR-20 values based on the total test were, on average, 0.80, making relatively small difference in reliabilities between the subscale and the total test. However, in the Between-subscales Correlation conditions of $r = 0.9$, KR-20

values based on the total test were, on average, 0.92, making large difference in reliabilities between the subscale and the total test.

In summary, the highest subscale score reliabilities were found in I = 20 Subscale Length and High Subscale Consistency conditions across test types. That is, KR-20 values seem to be large enough, if a subscale has sufficient number of items, and is internally consistent. Two other conditions, I = 10 Subscale Length and High Subscale Consistency, and I = 20 Subscale Length and Low Subscale Consistency, yielded marginally acceptable levels of reliability of 0.72 on average. However, the condition of I = 10 and Low Subscale Consistency yielded very low subscale score reliability, which may not be acceptable in practical tests.

IRT-based Descriptive Statistics

Subscale scores, θs, were obtained using the unidimensional 2PL-IRT model for the simulated data, and their summary statistics are shown in Table A4 and A5, including the means, the standard deviations, and empirical reliabilities for both subscale scores and total score, respectively from achievement and ability tests. These results also include the correlations of the subscale scores with total score. In all cases, empirical reliability was computed based on the ratio of true score variance to the sum of true score variance and error variance from score estimation, because all IRT-based scores were estimated based on Expected A Posteriori (EAP) method. From the results, the means and the standard deviations of subscale θs were, respectively, close to 0.0 and 0.85 across all conditions. The amount of empirical reliability substantially varied across conditions. Table 4.4 includes the summary of IRT-based empirical reliability under each simulation condition. Empirical reliabilities ranged between 0.55 and 0.84 across all conditions being considered. Generally, the large variance in empirical reliabilities

were due to the differences in Subscale Consistency and Subscale Length conditions, rather than Between-subscales Correlation and Test Type conditions.

First, the mean empirical reliabilities of subscale scores were 0.63 in the I = 10 Subscale Length condition, and 0.77 in the I = 20 Subscale Length condition, showing substantially increased amount of reliability in the greater subscale length condition. The empirical reliability had a broad range between 0.55 and 0.74 in the I = 10 Subscale Length condition, and between 0.71 and 0.83 in the I = 20 Subscale Length condition. The amount of the empirical reliability substantially varied across different subscale consistency and test type conditions.

Second, the amount of empirical reliability largely differed in the different subscale consistency conditions. The average empirical reliability was 0.76 in the High Subscale Consistency condition, and 0.64 in the Low Subscale Consistency condition, indicating that subscale scores are more reliable when responses within a subscale are highly correlated. However, the empirical reliability varied across different test types and subscale lengths, ranging between 0.68 and 0.84 in the High Subscale Consistency condition, and between 0.55 and 0.74 in the Low Subscale Consistency condition, showing large variation in reliability across different subscale consistency conditions.

Different test types showed only small differences in the empirical reliabilities of subscale scores. The average empirical reliability in the subscale scores was 0.69 in the Achievement test, and 0.72 in the Ability test, showing slightly higher reliability in the Ability test. The ranges of empirical reliabilities were very similar across test types, ranging between 0.55 and 0.80 in the Achievement test, and 0.58 and 0.84 in the Ability test.

Table 4.4. Empirical Reliability Means within Tests for the Four Subscale score $\boldsymbol{\theta}$s

Data condition	Empirical reliability mean

Subscale length	Subscale consistency	Between-subscales correlation	Achievement	Ability
I = 10	High	0.3	0.68	0.72
		0.6	0.68	0.72
		0.9	0.68	0.74
	Low	0.3	0.55	0.57
		0.6	0.55	0.58
		0.9	0.55	0.58
I = 20	High	0.3	0.80	0.83
		0.6	0.80	0.83
		0.9	0.80	0.84
	Low	0.3	0.71	0.74
		0.6	0.72	0.74
		0.9	0.71	0.73

Similar to the results from CTT, the amount of empirical reliabilities for the subscale scores were constant across three different between-subscales correlations, which shows that Between-subscales Correlation conditions do not have any direct impact on the reliability of raw subscale scores. Tables A8 and A9 present correlations among subscale θs. The correlations among scale scores of subscale scores ranged between 0.16 and 0.25, where $r = 0.3$ in the Between-subscales Correlation condition, between 0.33 and 0.5, where $r = 0.6$ in the Between-subscales Correlation condition, and between 0.49 and 0.76, where $r = 0.9$ in the Between-subscales Correlation condition across both achievement and ability tests.

However, correlations among subscale scores had positive relationship with the correlation of subscale scores with their total score. As correlation among subscale scores is large, the correlation between subscale scores and their total score is expected to be large, denoting the indirect impact of the Between-subscales Correlation on subscale score reliability. Empirical reliabilities based on the total test were as low as those based on subscales in the Between-subscales Correlation condition of $r = 0.3$, thus showing little difference in reliabilities between the subscale and the total test. In contrast, in Between-subscales Correlation conditions

of $r = 0.6$ and $r = 0.9$, empirical reliabilities based on the total test were much higher than those based on the subscales.

In summary, the highest subscale score reliabilities were found in the I = 20 Subscale Length and High Subscale Constancy conditions across test types (i.e., empirical reliability = 0.82). Empirical reliabilities were large enough if a subscale has a sufficient number of items, and the subscale is internally consistent. Two other conditions, I = 10 and High Subscale Consistency, and I = 20 and Low Subscale Consistency, yielded marginally acceptable levels of reliabilities of 0.71 on average. However, the condition comprising I = 10 and Low Subscale Consistency yielded a very low subscale score reliability of 0.56, which probably would not be deemed sufficient for an operational test.

Summary

Generally, the simulated data represented appropriate test properties, based on the specification in each condition. First, the resulting means and standard deviations of true item and true person parameters supported the plausibility of the simulation. The mean of true item difficulty values was, on average, 0.00 in Ability Test Type conditions, and -0.90 in Achievement Test Type conditions, as expected. Similarly, the mean of true item discrimination values was, on average, 1.2 in High Subscale Consistency condition, and 0.8 in Low Subscale Consistency condition, as expected as well. In addition, three different conditions of correlations among true subscale θs were also appropriately represented in the simulated data.

The reliability or accuracy of subscale scores based on CTT and IRT was influenced by conditions specified in the simulation for the study. The most dominant two factors on the accuracy of subscale scores were subscale length and subscale consistency. As expected, as the length of the subscale is I = 20 rather than I = 10, and subscale consistency is High rather than

Low, the accuracy or reliability of subscale scores was greater. Further, subscale score reliability or accuracy only slightly differed between the different test types (i.e., $\Delta = 0.02$). Lastly, the Between-subscales Correlation conditions did not appear to have an impact on the amount of reliability. However, between-subscale correlations were positively correlated with the amount of subscale-total correlation, indicating the possibility that between-subscale correlations may indirectly influence on reliabilities. From these results, the four factors that have been considered as variables affecting subscale score accuracy in the simulation, appear to be reasonable. Thus, the impact of these factors on various subscale estimates is worthy of being examined.

Subscale Score Estimates Based on the CTT and IRT Methods

Various CTT- and IRT-based subscale scoring methods are available. Seven subscale scoring methods were chosen for the study and used to compute subscale scores, with the simulated data. This section briefly summarizes the descriptives from the resulting subscale scores, and discuss how they differ from raw subscale scores.

CTT-based Subscale Scores

Subscale scores were estimated using four different types of CTT subscale scoring methods: raw subscale scoring, Kelley's regression, Holland-Hoskens' (HH) regression, and Haberman's weighted average methods (i.e., multiple regression). As explained earlier, the regression methods approximate the predicted values (i.e., subscale score estimates) with one of three types of predictive variables: observed subscale score (i.e., raw subscale score), the observed total score, and the weighted combination of observed subscale score and the total score.

Descriptive Statistics of the CTT-based Subscale Scores

The means and the standard deviations of the resulting subscale score estimates in achievement and ability tests are shown in Tables A10 and A11, respectively. These means and standard deviations were obtained by averaging subscale score means and standard deviations from the 100 replications within each condition. The means from all three regression scores were the same as the raw subscale scores, but their standard deviations were somewhat smaller than those of the raw subscale scores.

Measurement of Reliability in CTT-based Subscale Scores: PRMSE-MB

PRMSE-MB was suggested by Haberman (2008) as a measure of evaluating the added-value of subscale score over total score. However, as described earlier, the PRMSE-MBs are known to be mathematically equal to traditional reliability estimates. Thus, the larger the PRMSE-MB values, the more reliable the subscale scores. PRMSE-MB values were computed from Kelley's, Holland-Hoskens' and Haberman's methods, and compared across various conditions, and the results are shown in Tables A12 and A13, respectively for achievement and ability tests. See Chapter 2 for the details about the computation.

The square root of PRMSE-MB values from different methods were compared with the correlations between true subscale score θs and estimated subscale scores. Note that the correlations between true subscale scores and estimates are conceptually same as the square root of reliability of the subscale scores. The following three figures, Figures 4.1, 4.2, and 4.3 show high consistency between the square roots of PRMSE-MB from three different methods and correlations of true and estimated subscale scores.

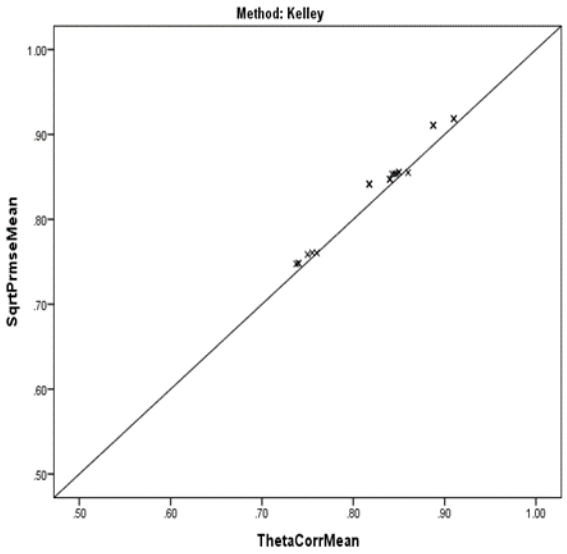

Figure 4.1. Consistency between the Square Roots of PRMSE-MB from the Kelley's Method and Correlations of True and Estimated Subscale Scores

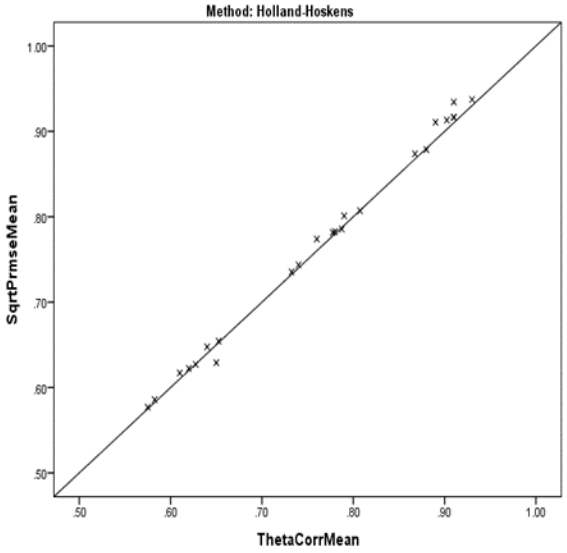

Figure 4.2. Consistency between the Square Roots of PRMSE-MB from the Holland-Hoskens' Method and Correlations of True and Estimated Subscale Scores

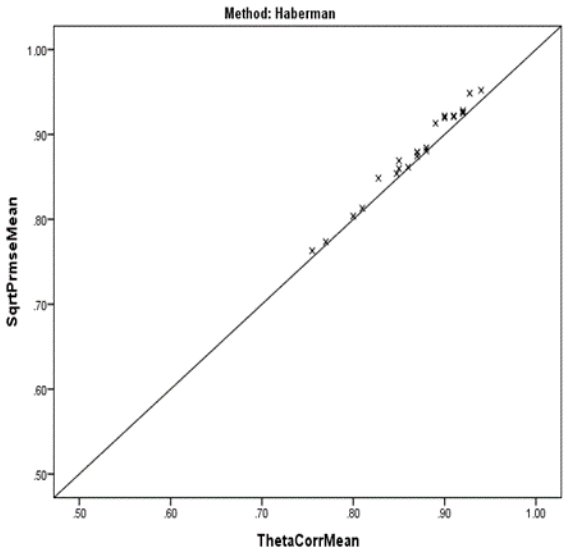

Figure 4.3. Consistency between the Square Roots of PRMSE-MB from the Haberman's Method and Correlations of True and Estimated Subscale Scores

First, the PRMSE-MBs did not show much difference between Achievement and Ability tests. The mean PRMSE-MBs in Achievement test condition were 0.71, 0.60, and 0.77 in Kelley's, HH, and Haberman's methods, respectively, and those in Ability test condition were 0.72, 0.61, and 0.78 in order.

Second, the PRMSE-MBs were generally larger in I = 20 than I = 10 conditions and in the High Subscale Consistency condition than in the Low Subscale Consistency condition, with any other conditions fixed. In specific, P$RMSE - MB_{Kelley}$ ranged in 0.56~0.73 in I = 10 conditions, and 0.72~0.84 in I = 20 conditions, representing higher reliability in the Subscale Length of I = 20. With other conditions fixed, P$RMSE - MB_{HH}$ ranged in 0.33~0.84 in I = 10 conditions and 0.39~0.88 in I = 20 conditions, which are widely spread. P$RMSE - MB_{Haberman}$ ranged between 0.58~0.86 in I = 10 conditions, and 0.73~0.91 in I = 20 conditions. In average, $PRMSE - MB_{Kelley}$ had averages of 0.64 in I = 10, and 0.78 in I = 20, PRMSE $-$ MB$_{HH}$ had averages of 0.58 in I = 10, and 0.63 in I = 20, and P$RMSE - MB_{Haberman}$ had averages of 0.73

in I = 10, and 0.83 in I = 20. With all other conditions held, the three different methods had high PRMSE-MB values in I = 20 condition, and the largest difference between I = 10 and I = 20 conditions was observed in Kelley's method.

Third, the PRMSE-MBs were generally larger in the High Subscale Consistency condition than in the Low Subscale Consistency condition, with any other conditions fixed. Also, P$RMSE - MB_{Kelley}$ were 0.71~0.84 in the High Subscale Consistency condition, and 0.56~0.73 in the Low Subscale Consistency condition. P$RMSE - MB_{HH}$ were 0.38~0.88 in the High Subscale Consistency condition, and 0.33~0.84 in the Low Subscale Consistency condition. P$RMSE - MB_{Haberman}$ were 0.72~0.91 in the High Subscale Consistency condition, and 0.58~0.86 in the Low Subscale Consistency condition. In average, P$RMSE - MB_{Kelley}$ had averages of 0.78 in the High Subscale Consistency condition, and 0.65 in the Low Subscale Consistency condition, and P$RMSE - MB_{HH}$ had averages of 0.63 in the High Subscale Consistency condition, and 0.58 in the Low Subscale Consistency condition, and P$RMSE - MB_{Haberman}$ had averages of 0.82 in the High Subscale Consistency condition, and 0.73 in the Low Subscale Consistency condition. With all other conditions held, the three different methods had high PRMSE-MB values in the High Subscale Consistency condition.

Fourth, PRMSE-MBs varied in their amount in different Between-subscales conditions. With other conditions fixed, P$RMSE - MB_{Kelley}$ were constant across three Between-subscales conditions of r = 0.3, r = 0.6, and r = 0.9, and exactly same as raw subscale score reliability. P$RMSE - MB_{HH}$ ranged between 0.33 and 0.43 in r = 0.3, between 0.54 and 0.65 in r = 0.6, and between 0.76 and 0.88 in r = 0.9, with all other conditions fixed. P$RMSE - MB_{Haberman}$ ranged between 0.58 and 0.85 in r = 0.3, between 0.64 and 0.86 in r = 0.6, and between 0.77 and 0.91 in r = 0.9 across all other conditions. In average, P$RMSE - MB_{Kelley}$ was 0.71 across all Between-

subscales Correlation conditions, P$RMSE - MB_{HH}$ was averagely 0.39 in r = 0.3, 0.60 in r = 0.6, and 0.83 in r = 0.9, and P$RMSE - MB_{Haberman}$ was averagely 0.72 in r = 0.3, 0.76 in r = 0.6, and 0.85 in r = 0.9. When the Between-subscales correlation is 0.3, the highest average PRMSE-MB was found in Haberman's method (i.e., 0.72), and the lowest average PRMSE-MB was found in Holland-Hoskens' method (i.e., 0.39). When the Between-subscales Correlation is 0.6, the highest PRMSE-MB was found also in Haberman's method (i.e., 0.76), and the lowest average PRMSE-MB was found in Kelley's method (i.e., 0.71). In turn, when the Between-subscales Correlation is 0.9, the highest PRMSE-MB was found in Haberman's method (i.e., 0.85), and the lowest average PRMSE-MB was found in Kelley's method (i.e., 0.71).

From these results, the highest average PRMSE-MBs were found in Haberman's method in each Between-subscales Correlation condition. The large variance in average PRMSE-MBs under three Between-subscales Correlation conditions was observed in the HH method. The HH method yielded the lowest PRMSE-MBs means in Between-subscales Correlations of 0.3 or 0.6, but as high as Haberman's method in Between-subscales Correlation of 0.9. Kelley's method, as noted above, yielded the constant PRMSE-MBs means across different Between-subscales Correlations, which is the same as in KR-20 of raw subscale scores.

In summary, three out of four conditions, Subscale Length, Subscale Consistency, and Between-subscale Correlation influenced the amount of PRMSE-MBs. That is, subscale scores generally had higher PRMSE-MBs in Subscale Length of I = 20, High Subscale Consistency, and High Between-subscales Correlation conditions. Among four CTT-based methods, Haberman's method yielded higher PRMSE-MBs than any other method across all conditions. HH method yielded improved PRMSE-MBs compared to raw subscale scores only when

Between-subscales Correlation is high (i.e., 0.9). Further, Kelley's method yielded the equal level of reliability. Figure 4.4 below presents these findings.

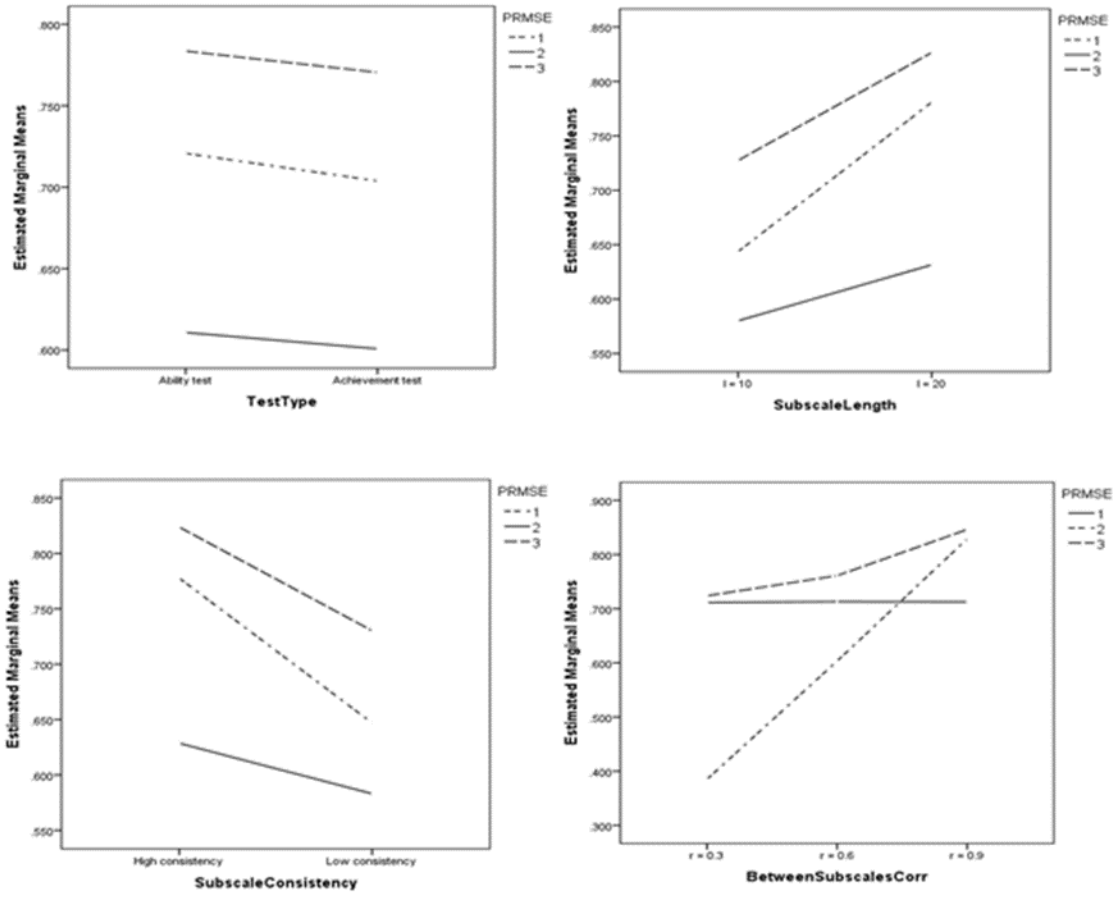

Figure 4.4. PRMSE-MB means from Different CTT-based Subscale Scoring Methods across Test Type, Subscale Length, Subscale Consistency, and Between-subscale Correlation Conditions (Line 1=Kelley's Method, Line 2 = HH Method, and Line 3 = Haberman's Method)

IRT-based Subscale Scores

Subscale scores were computed using unidimensional IRT models (1PL, 2PL, and 3PL), multidimensional IRT (MIRT-2PL), and the OPI method, with the simulated data. The results included descriptive statistics from these methods, overall goodness-of-fit comparisons, and empirical reliability values. For the IRT analyses, the marginal maximum likelihood estimation

based on the expectation-maximization (EM) algorithm and the Expected A Posteriori (EAP) were used, respectively for item and person parameter estimation, in which fifteen quadrature points were used.

Descriptive Statistics of the IRT-based Subscale Scores

Because descriptives from the unidimensional 2PL IRT model were shown in the previous section, they would not be described in the current section. Note that the mean and the standard deviations of subscale trait scores θs based on the unidimensional 2PL IRT model were 0.00 and 0.85, respectively.

OPI values are adjusted p-values based on the raw subscale scores and global trait scores from the unidimensional IRT analysis (i.e., mainly 3PL). Tables A14 and A15 include results of the means and the standard deviations of OPI scores in achievement and ability tests. OPI subscale scores showed only a slight difference from the raw subscale scores with slightly bigger score variances.

Measurement of Multidimensionality

Measuring the dimensionality of a test may be an important consideration in order for subscale scores to be reported. Three unidimensional IRT models (i.e., 1PL, 2PL, and 3PL) and the MIRT-2PL model were compared with respect to the overall goodness-of-fit. Tables A16 and A 17 presents two types of overall goodness-of-fit statistics, Akaike Information Criterion (AIC) and the -2loglikelihood (-2lnL), respectively with achievement and ability tests. The Δ-2lnL values based on the likelihood difference between models are also shown in the last column of respective tables. The statistical significance of fit difference may also be examined through the -2lnL difference (i.e., Δ-2lnL), because the difference of -2lnL is considered to be asymptotically distributed chi-square with the difference of the degrees of freedom. These values were obtained

for 1PL vs. 2PL, 2PL vs. 3PL, and 2PL vs. MIRT-2PL. Note that in both AIC and -2lnL indices, the smaller values indicated better model fit.

In most comparisons among unidimensional models, the 3PL models showed a better fit than 2PL models. However, the 3PL models generally had largely increased number of parameters relative to the 2PL models, which often led to a larger AIC in the comparisons, which indicates worse fit. In the comparisons of unidimensional and multidimensional models, evidently, in both achievement and ability data, the MIRT-2PL model showed a better fit than the 2PL model in all conditions. However, the amount of the -2lnL difference greatly differed in different Between-subscales correlation conditions. Namely, in the Between-subscales correlation condition where $r = 0.3$, the multidimensional 2PL model usually showed much better fit than the unidimensional 2PL model, as expected, since the resulting total test would be heterogeneous. In the Between-subscales correlation condition where $r = 0.9$, the difference of -2lnL between unidimensional and multidimensional models was not so large.

Measure of Empirical Reliabilities

The empirical reliabilities between the unidimensional 2PL and the MIRT-2PL model scores were compared. Tables A18 and A19 show the empirical reliabilities, respectively from achievement and ability tests. In order to compare reliabilities of subscale scores from 2PL and MIRT-2PL, the mean reliabilities within tests were computed. Table 4.5 shows average empirical reliability among subscale scores in Tests. Empirical reliabilities in the MIRT-2PL scores were higher than in the unidimensional 2PL subscale scores, for all conditions. The large increase in empirical reliabilities was found when correlations among subscales get high. Especially, as correlations among subscale scores are as high as 0.9, empirical reliability of MIRT-2PL scores were comparable to that of unidimensional 2PL scores.

In general, the total test reliability for all conditions could be deemed as acceptable, as all were above .70. However, the subscale reliabilities often fell below .70 in the conditions with shorter tests and low subscale consistency.

Table 4.5. Empirical Reliability Means from Total and Subscale Scores in Tests Based on the Unidimensional and the multidimensional 2PL Models

Test type	Subscale length	Subscale consistency	Between-subscales correlation	Empirical reliability means		
				Total test	Subscale	
				2PL	2PL	MIRT-2PL
Achievement	I = 10	High	r = 0.3	0.79	0.68	0.70
			r = 0.6	0.85	0.68	0.74
			r = 0.9	0.88	0.68	0.87
		Low	r = 0.3	0.69	0.55	0.58
			r = 0.6	0.77	0.55	0.64
			r = 0.9	0.81	0.55	0.80
	I = 20	High	r = 0.3	0.88	0.80	0.81
			r = 0.6	0.92	0.80	0.83
			r = 0.9	0.93	0.80	0.93
		Low	r = 0.3	0.82	0.71	0.72
			r = 0.6	0.87	0.71	0.76
			r = 0.9	0.90	0.71	0.89
Ability	I = 10	High	r = 0.3	0.81	0.72	0.73
			r = 0.6	0.87	0.72	0.77
			r = 0.9	0.90	0.72	0.90
		Low	r = 0.3	0.70	0.57	0.60
			r = 0.6	0.78	0.58	0.66
			r = 0.9	0.83	0.58	0.81
	I = 20	High	r = 0.3	0.90	0.83	0.74
			r = 0.6	0.93	0.83	0.78
			r = 0.9	0.95	0.84	0.90
		Low	r = 0.3	0.83	0.73	0.84
			r = 0.6	0.88	0.73	0.86
			r = 0.9	0.91	0.73	0.94

Measuring Accuracy of Subscale Score Estimation: RMSE-MB and RMSE-SB

In the current study, the measurement of the subscale score accuracy was carried out by observing essentially two criteria: 1) measuring root mean square errors (RMSE) including RMSE-MB and RMSE-SB, and 2) examining correlations of estimated subscale scores with their true subscale scores. Results from the two criteria are interpreted in order.

Approximation Errors from CTT-based Subscale Scores

As explained earlier, RMSE-MB measures the difference between subscale scores predicted by a model and observed raw subscale scores. Applied to the CTT-based subscale scoring methods, the RMSE-MB may be a worthy measure of subscale score accuracy (Haberman, 2008). However, in a simulation approach, RMSE can be defined in a somewhat different way, as the difference between true scores and estimated scores. In this section, CTT-based subscale scores are assessed by both RMSE-MB and RMSE-SB for their accuracy. The proportional reduction in MSE (PRMSE), which is also suggested by Haberman (2008) as an index of examining the added-value, are computed with the resulting subscale scores. In turn, for measuring the accuracy of IRT-based subscale scores, RMSE-SB is used.

CTT-based Subscale Score Accuracy: RMSE-MB

The estimation of scores yields errors, in which the errors are expected to be small for the more accurate prediction. One index of the accuracy of subscale score estimation is root mean square error (RMSE-MB), which measures the error of approximation. RMSE-MB is obtainable by computing the squared mean of residual of observed subscale scores and their true subscale scores (i.e., predicted scores based on a linear regression model for subscale scoring) and rooting the mean residual. Tables A20 and A21 include RMSE-MB values based on different subscale scoring methods under various conditions, respectively for achievement and ability tests. The RMSE-MB means were obtained by averaging RMSE-MBs over 100 replications in each

condition, and for the comparisons of accuracy among raw subscale scores and the other subscale scores, the results of the standard error of measurement of the subscale score, $\sigma(e_x)$, is presented as well. See Chapter 2 for more information on the details of the RMSE-MB computation.

The means of SEs and RMSE-MBs across various subscale scores are shown in Table 4.6. Comparing the different data conditions within a method, subscale scoring methods had different amount of approximation error in different data conditions. It should be noted that the true test score variability is directly impacted by the conditions. That is, the larger the score variability for these conditions, the larger the RMSE-MB becomes. However, in general, they were large in the I = 20 Subscale Length, and the Ability Test Type conditions, with the other conditions fixed. Specifically, the highest standard errors (SE) of measurement in raw subscale scores were found in the Ability Test Type, the I = 20 Subscale Length, the Low Subscale Consistency condition (i.e., 2.07), and the lowest SEs were found in the Achievement Test Type, the I =10 Subscale Length, and the High Subscale Consistency condition (i.e., SE = 1.28). The Kelley's method yielded the highest RMSE-MBs in the Ability Test Type, the I = 20 Subscale Length, and the High Subscale Consistency condition (i.e., 1.78), and the lowest RMSE-MBs in Achievement Test Type, the I = 10 Subscale Length, and the Low Subscale Consistency condition (i.e., 1.05). The Holland-Hoskens's (HH) method yielded the highest RMSE-MBs in the Ability Test Type, the I = 20 Subscale Length, the High Subscale Consistency, and the Between-subscales Correlation of r = 0.9 condition (i.e., 3.40), and the lowest RMSE-MBs were observed in the Achievement Test Type, the I = 10 Subscale Length, the Low Subscale Consistency, and High Between-subscale Correlation condition of r = 0.9 (i.e., 0.77). In turn, the Haberman's weighted average method yielded the greatest RMSE-MBs in the Ability Test Type,

High Subscale Consistency, the I = 20 Subscale Length, the Between-subscales Correlation of r = 0.3 condition (i.e., 1.75), and the lowest RMSE-MBs in the Achievement Test Type, the I = 10 Subscale Length, the Low Subscale Consistency, and High Between-subscales Correlation of *r* = 0.9 conditions (i.e., 0.76).

Compared to the raw subscale scoring method, Kelley's and Haberman's methods yielded less approximation error with smaller RMSE-MBs in all simulation conditions. In particular, the Haberman's method considerably reduced amount of error, although the amount of reduction in error varied across different simulation conditions. Specifically, the large reduction in RMSE-MBs were observed in the Low Subscale Consistency and the Between-subscales Correlation of *r* = 0.9 conditions. In the meantime, the HH method yielded even higher approximation error on average than the raw subscale scoring method, indicating that the HH method may not provide more reliable estimates. However, their performance substantially varied depending on the size of correlations among subscales. For example, in the Between-subscales Correlation conditions where *r* = 0.9, the HH method yielded considerably decreased RMSE-MBs compared to those from the raw subscale method, whereas they poorly performed in the Between-subscales Correlation conditions where *r* = 0.3 or *r* = 0.6.

Table 4.6. RMSE-MB Means from CTT Subscale Scores

Test type	Subscale Length	Subscale consistency	Between-subscales correlation	$\sigma(e_x)$	$RMSE - MB_{Kelley}$	$RMSE - MB_{HH}$	$RMSE - MB_{Haberman}$
Achievement	I = 10	High	*r* = 0.3	1.28	1.08	1.57	1.06
			r = 0.6	1.28	1.08	1.26	0.99
			r = 0.9	1.28	1.08	0.83	0.79
		Low	*r* = 0.3	1.41	1.05	1.29	1.02
			r = 0.6	1.40	1.05	1.07	0.94
			r = 0.9	1.40	1.05	0.77	0.76
	I = 20	High	*r* = 0.3	1.81	1.65	3.04	1.63
			r = 0.6	1.81	1.65	2.40	1.55

			r = 0.9	1.81	1.65	1.43	1.27
		Low	*r* = 0.3	1.98	1.68	2.48	1.65
			r = 0.6	1.98	1.68	1.98	1.54
			r = 0.9	1.98	1.68	1.29	1.22
Ability	I = 10	High	*r* = 0.3	1.37	1.17	1.74	1.14
			r = 0.6	1.37	1.17	1.40	1.07
			r = 0.9	1.37	1.17	0.90	0.85
		Low	*r* = 0.3	1.46	1.11	1.38	1.08
			r = 0.6	1.46	1.11	1.15	1.00
			r = 0.9	1.46	1.11	0.81	0.80
	I = 20	High	*r* = 0.3	1.93	1.78	3.40	1.75
			r = 0.6	1.93	1.78	2.66	1.67
			r = 0.9	1.94	1.78	1.58	1.38
		Low	*r* = 0.3	2.07	1.77	2.65	1.73
			r = 0.6	2.07	1.77	2.12	1.62
			r = 0.9	2.07	1.77	1.36	1.29

The following two figures, Figure 4.5 and Figure 4.6, present these differences among RMSE-MB means for subscale scores, from different methods across different data conditions of Test Type, Subscale Consistency, and Between-subscales correlation conditions, respectively in I = 10 and I = 20 Subscale Length conditions. The means of error from Raw Subscale scores and Kelley Method were consistent across different Between-subscales Correlation conditions, but lower error means were found in Kelley's method. Overall, Haberman's method yielded the lowest RMSE-MBs regardless of test conditions. However, the HH method presented large changes in RMSE-MBs across Between-subscales Correlation conditions. The amount of RMSE-MBs were larger than the SEs in Low Between-subscales Correlation condition of r = 0.3, and similar to or smaller than the SEs in High Between-subscales Correlation condition of r = 0.9. However, this pattern differed in different Subscale Consistency conditions. The largest difference across different Subscale Consistency conditions were found in Raw Subscale scores and HH method. Raw subscale scores had larger RMSE-MBs in high Subscale Consistency, and HH method showed the larger variance of the RMSE-MBs across different Between-subscales

Correlation conditions. Different Test Type conditions did not have significant difference with other conditions fixed. Similar patterns are found in the Subscale Length conditions of I = 20, but they generally showed higher RMSE-MBs, due to the increased variance in longer test length conditions.

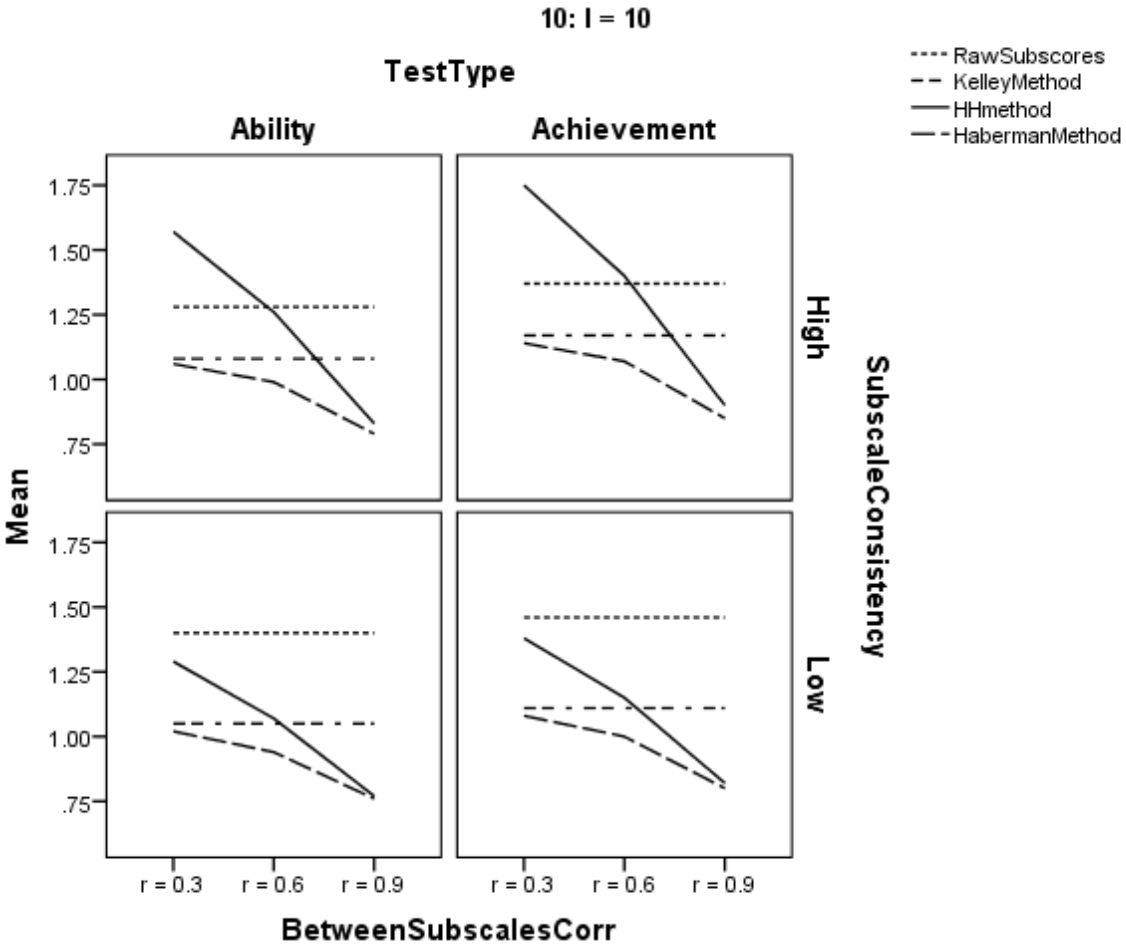

Figure 4.5. RMSE-MB Means from Different Four Methods across Test Type, Subscale Consistency, and Between-subscales Correlation Conditions in Subscale Length of I = 10

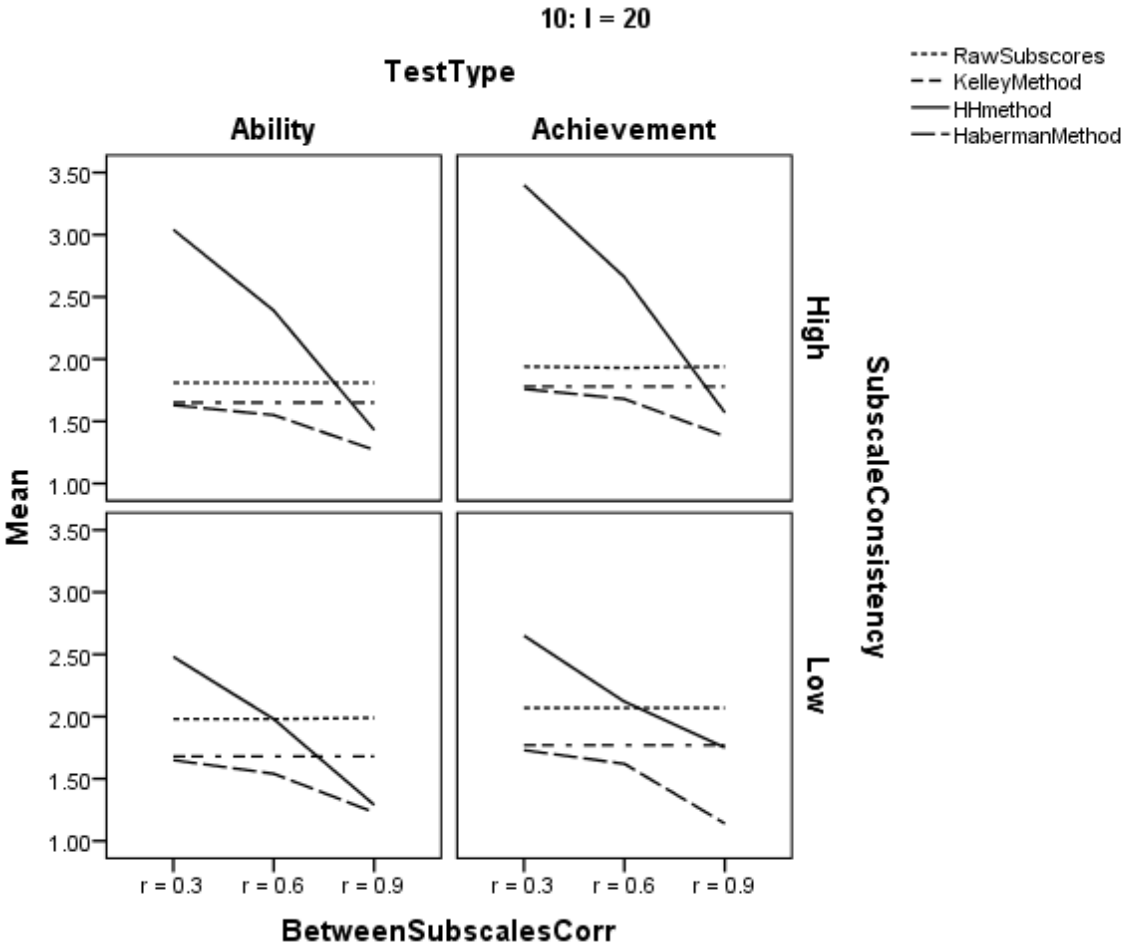

Figure 4.6. RMSE-MB Means from Different Four Methods across Test Type, Subscale Consistency, and Between-subscales Correlation Conditions in Subscale Length of I = 20

Repeated Measures ANOVA: Comparisons of RMSE-MBs

The repeated measures procedure was employed to examine the mean differences in RMSE-MB values. RMSE-MBs from different subscale scoring methods were repeated measures, and Test type, Subscale Consistency, and Between-subscales correlation variables were between-group factors. Because subscale items within a test were generated under the same conditions and their resulting subscale scores were mostly the same across subscales, they yielded the equivalent RMSE-MB means across subscales. Thus, RMSE-MB means were

averaged over subscales, and the average RMSE-MB means were used to conduct the repeated measures analysis.

Table 4.7 below presents results of repeated measures for RMSE-MBs based on different methods across different simulation conditions. The results show that all within- and between-groups effects were statistically significant. First, RMSE-MBs based on different methods (i.e., raw, Kelley, HH, and Haberman methods) were statistically and significantly different. That is, RMSE-MB values were statistically significantly different among subscale scoring methods. Also, the amount of RMSE-MB from different methods varied across all four simulation conditions. By comparing effect size (i.e., squared partial eta, η_p^2), the influential effects on the amount of RMSE-MB were found in the main effect of Method (i.e., 0.99), two-way interaction of Method x Test Type (i.e., 0.56), Method x Subscale Consistency (i.e., 0.98), Method x Subscale Length (i.e., 0.99), and Method x Between-subscales Correlation (i.e., 0.99), and three way interaction effect of Method x Subscale Consistency x Subscale Length (i.e., 0.87), Method x Subscale Consistency x Between-subscales Correlation (i.e., 0.90), and Method x Subscale Length x Between Subscale Correlation (i.e., 0.97). Although the four-way and five-way interactions were also statistically significant, their effect sizes were not so large. The following four figures, Figure 4.7 through Figure 4.10 correspond to the results of two-way interactions in which the Test Type, the Subscale Consistency, the Subscale Length, and the Between-subscale Correlation conditions are, respectively, involved.

Table 4.7. Test of Repeated Measures of RMSE-MBs for Subscale Score Estimates

Source	df	SS	MS	F_{obs}	p	η_p^2
Method	3	329.93	109.98	648,332.4	<0.01	**0.99**
2 way interaction						
Method*TestType	3	1.55	0.52	3,040.95	<0.01	**0.56**
Method*SubscaleConsistency	3	63.48	21.16	124746.89	<0.01	**0.98**

Method*SubscaleLength	3	83.81	27.94	164,689.25	<0.01	**0.99**
Method*BetweenSubscaleCorr	6	312.66	52.11	307,200.1	<0.01	**0.99**
3 way interaction						
Method*TestType*SubscaleConsistency Mixture	3	0.37	0.12	732.40	<0.01	0.24
Method*TestType *SubscaleLength	3	0.43	0.14	835.98	<0.01	0.26
Method*TestType*BetweenSubscaleCorr	6	0.95	0.16	929.03	<0.01	0.44
Method*SubscaleConsistency* SubscaleLength	3	8.26	2.76	16,239.24	<0.01	**0.87**
Method*SubscaleConsistency* BetweenSubscaleCorr	6	11.03	1.84	10,837.15	<0.01	**0.90**
Method*SubscaleLength* BetweenSubscaleCorr	6	46.11	7.69	45,301.86	<0.01	**0.97**
4 way interaction						
Method*TestType*SubscaleConsistency* SubscaleLength	3	0.09	0.03	168.97	<0.01	0.07
Method*TestType*SubscaleConsistency* BetweenSubscaleCorr	6	0.15	0.02	144.25	<0.01	0.11
Method*TestType*SubscaleLength* BetweenSubscaleCorr	6	0.15	0.02	142.05	<0.01	0.11
Method* SubscaleConsistency* SubscaleLength*BetweenSubscaleCorr	6	1.37	0.23	1,349.48	<0.01	0.53
5 way interaction						
Method*TestType*SubscaleConsistency * SubscaleLength*BetweenSubscaleCorr	6	0.02	0.00	18.71	<0.01	0.02
Error (Method)	7,128	1.21	0.00			

Figure 4.7 compares the RMSE-MB means based on different methods in the Ability vs. Achievement Test Type conditions. The Ability Test Type was expected to have larger RMSE-MB means because it has higher variance than the Achievement Test Type, due to item difficulty mean of 0.5. The actuall results show that RMSE-MB means were somewhat higher in the Ability Test Type than in the Achievement Test Type. The lowest RMSE-MB were found in the Haberman's method, and the highest RMSE-MB means were observed in the HH method. RMSE-MB were lower in order of Haberman < Kelley < Raw < HH in both test type conditions. Specifically, in the Ability Test Type condition, the Haberman, the Kelley, the Raw, and the HH methods had the RMSE-MB means of, 1.28, 1.46, 1.71, and 1.76 respectively, and in the

Achievement Test Type condition, they had 1.20, 1.37, 1.62, and 1.62, in order. Notice that only the Kelley's and the Haberman's methods yielded less RMSE-MBs than the Raw subscale scoring in both test types, indicating that their subscale scores from these methods produces better approximation, with more accuracy.

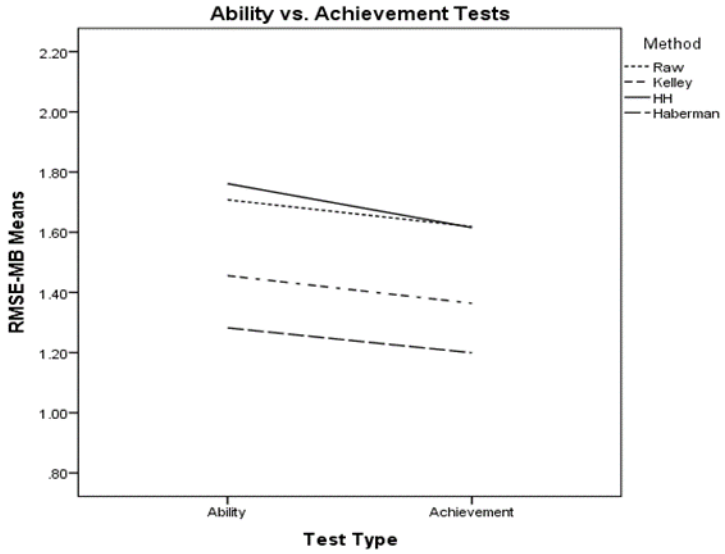

Figure 4.7. RMSE-MB Means in Distinct Test Types: Ability vs. Achievement Tests

Figure 4.8 below shows the RMSE-MB means across different methods depending on the degree of subscale consistency: High vs. Low Subscale Consistency conditions. The high internal consistency of a subscale was expected to have larger RMSE-MB means because it has higher standard deviations than low subscale consistency condition. As expected, while Raw subscales scoring method yielded greatly decreased RMSE-MB means in the Low Subscale Consistency condition, other three subscale scoring methods performed in the other way. That is, they showed lower RMSE means in the Low Subscale Consistency condition than in the High Subscale Consistency condition. Such pattern was especially noticeable in the HH method. Also, the Kelley's and the Haberman's methods yielded less RMSE-MBs than the raw subscale scoring methods, with more accuracy.

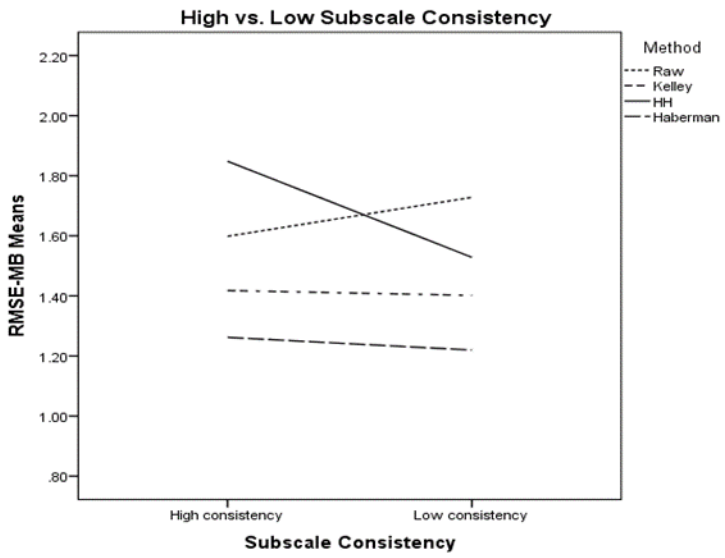

Figure 4.8. RMSE-MB Means in Different Subscale Consistency Conditions: High vs. Low Subscale Consistency

Next, Figure 4.9 present the amount of RMSE-MB means from different methods, depending on different subscale lengths. The I = 20 subscale length condition was expected to yield larger RMSE-MB means than the I = 10 subscale length condition because it includes more items, increasing variability of subscale scores. As expected, RMSE-MB means were smaller in I = 10 than in I =20. Similar to results for the other conditions, the smallest RMSE-MB means were observed in the Haberman's method, and the largest RMSE-MB means were found in the raw subscale scoring methods in the I = 10 Subscale Length condition, and in the HH method in the I = 20 Subscale Length condition. Also, regardless of subscale lengths, Kelley's and Haberman's methods performed better than the raw subscale scoring.

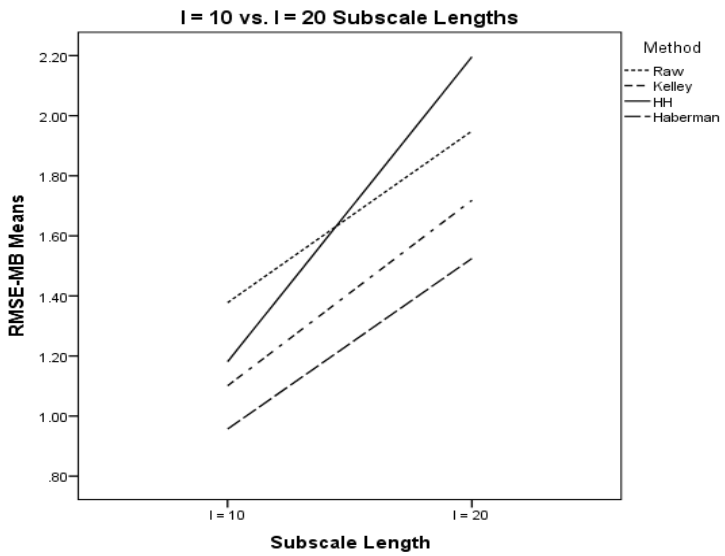

Figure 4.9. RMSE-MB Means in Different Subscale Length Conditions: I = 10 vs. I = 20 Subscale Lengths

Figure 4.10 shows RMSE-MB values from different methods depending on the three Between-subscales Correlation conditions: r = 0.3, r = 0.6, and r =0.9. The lowest RMSE-MBs were observed in Haberman's method regardless of the size of Between-subscales Correlation, and the next low RMSE-MBs were observed in Kelley's method. Also, raw subscale scores and Kelley's method yielded consistently high RMSE-MBs. Although HH method yielded the highest RMSE-MBs among methods in r = 0.3 conditions, their RMSE-MBs were as low as those in Haberman's method in r = 0.9 condition.

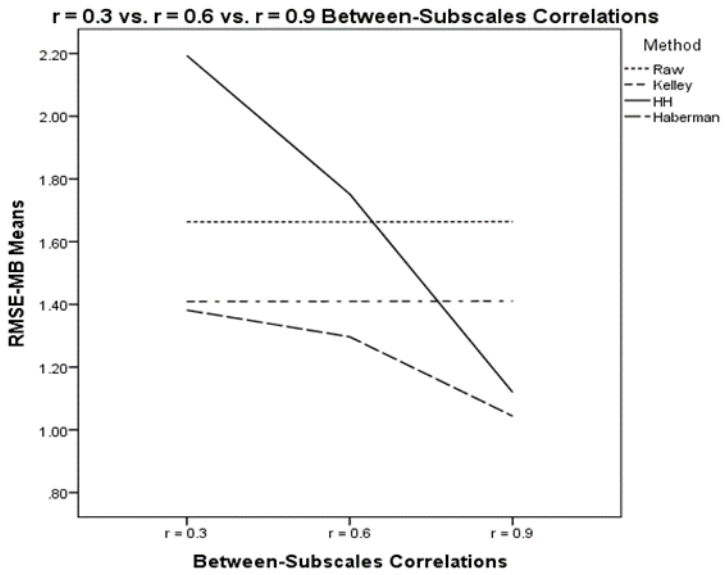

Figure 4.10. RMSE-MB Means in Different Between-subscales Correlation Conditions: r = 0.3, 0.6, vs., 0.9

The following figures, Figure 4.11 through Figure 4.14, present results of three-way interactions, respectively of Method x different Subscale Consistency x Between-subscales Correlation, Method x Subscale Length x Between Subscales Correlation, and Method x Subscale Length x Subscale Consistency on RMSE-MB means. They present how RMSE-MBs values based on different methods differently perform depending on the combinations of other two between-group factors.

Figure 4.11 and Figure 4.12 show the RMSE-MBs from different methods across Between-subscales Correlations, respectively in the High vs. Low Subscale Consistency conditions. The distinct appearance in different subscale-correlations among different subscale consistency conditions were mostly observed in the performance of the HH methods. In both subscale consistency conditions, RMSE-MB from the HH methods was very large in $r = 0.3$, and was dramatically dropped in $r = 0.9$, resulting in less accuracy in the Low Between-subscale

Correlation condition, and greater accuracy in the High Between-subscale Correlation condition than the raw subscale scoring method. The degree of how well the HH performs differed in the Between-subscales Correlation, where r = 0.6. The HH method poorly performed by having larger error in the High Subscale Consistency condition than the raw subscale scoring method, but it performed better in the Low Subscale Consistency condition. That is, HH method showed different appearance in different Subscale Consistency conditions. In High Subscale Consistency condition, the RMSE-MB mean from the HH method were higher than that from raw subscale scores in *r* = 0.3, but in the Low Subscale Consistency condition, it was lower than that of raw subscale scores.

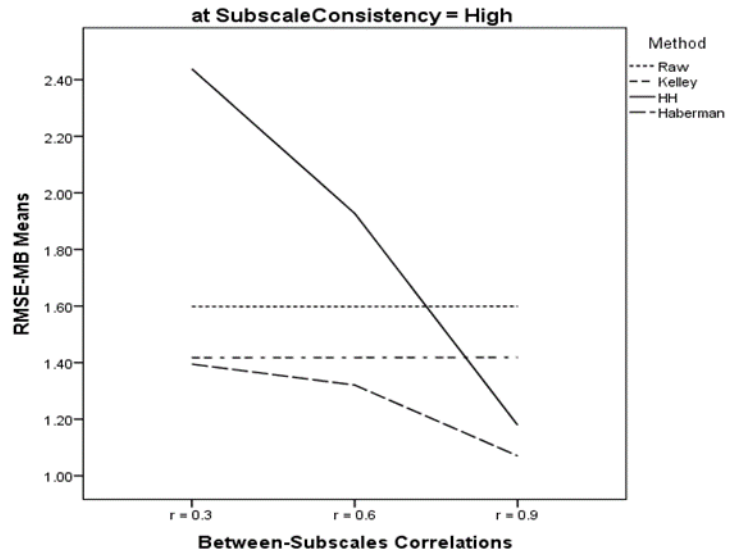

Figure 4.11. RMSE-MB Means across Different Between-subscales Correlation in the High Subscale Consistency Condition.

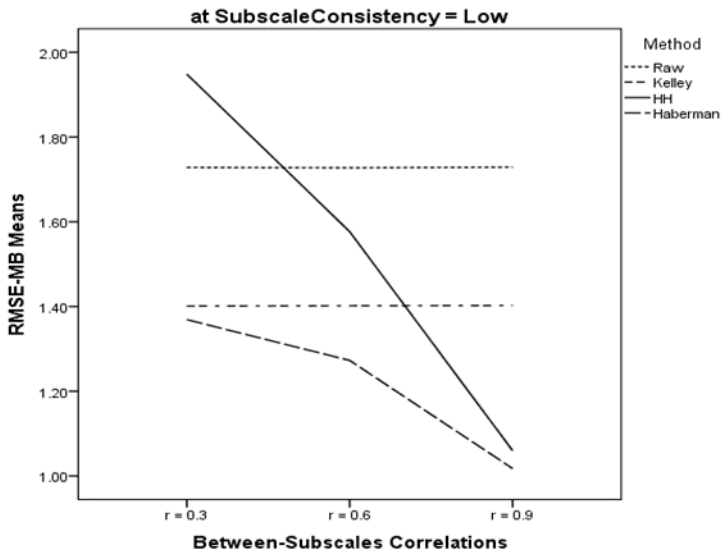

Figure 4.12. RMSE-MB Means across Different Between-subscales Correlation in the Low Subscale Consistency Condition

The following figures, Figure 4.13 and Figure 4.14, shows the RMSE-MB means across different between-subscales correlations, respectively in I =10 Subscale Length and I = 20 Subscale Length conditions. In both subscale length conditions, the distinction of RMSE-MB means across different between-subscales correlations was mainly found in the relationship in RMSE-Means between the raw subscale scoring and the HH methods. The HH method performed worse than the raw subscale scoring method in r = 0.3, but it performed better in r = 0.9 in both subscale length conditions. However, the distinguishable difference in pattern was found in r = 0.6. Specifically, the HH method had lower RMSE-MB than the raw subscale scoring method in the I = 10 Subscale Length condition, whereas it had much higher RMSE-MB than the raw subscale scoring method in the I = 20 Subscale Length condition.

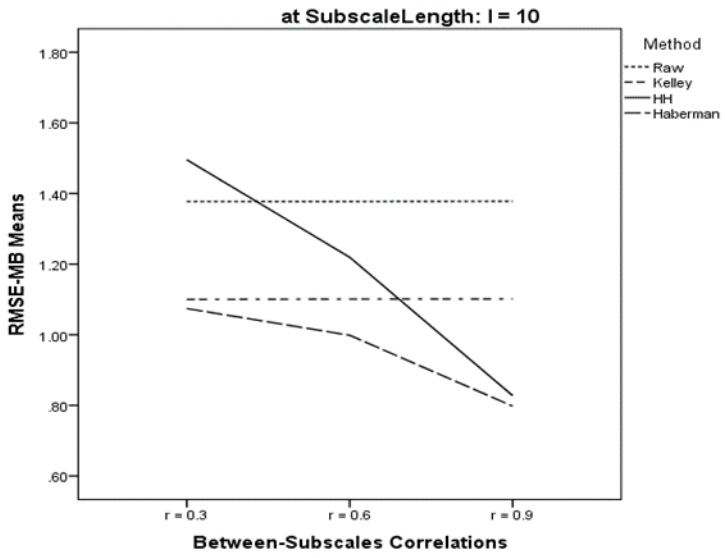

Figure 4.13. RMSE-MB Means across Different Between-subscales Correlation in the I = 10 Subscale Length Condition

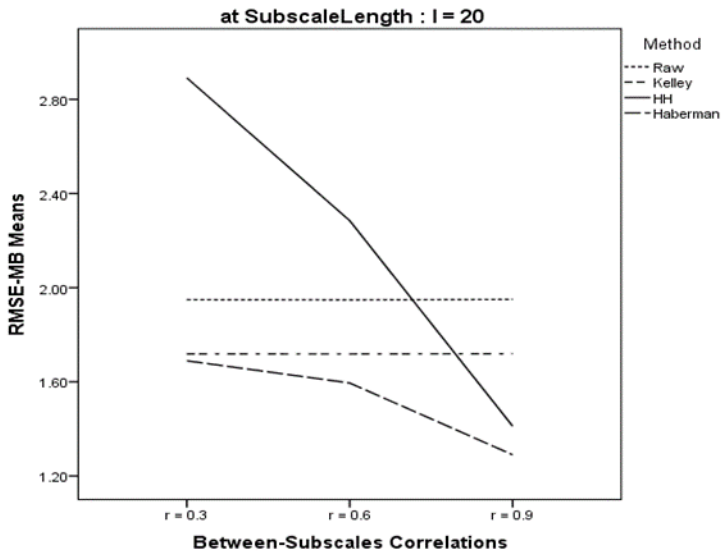

Figure 4.14. RMSE-MB Means across Different Between-subscales Correlation in the I = 20 Subscale Length Condition

Summary. The Kelley's and Haberman's methods generally yielded less error than the raw subscale scoring method. However, the HH method varied in error across the simulation

conditions, especially across between-subscale correlation conditions. All data conditions, being considered, were influencing the magnitude of RMSE-MB values, although the size of the impact was different. However, the impact of data conditions on the magnitude of RMSE-MB values were seemingly different from results expected in the research hypothesis. Such inconsistency was relevant to the magnitude of standard deviations (SDs). In general, RMSE values tend to increase, as SDs become larger. Thus, the High Subscale Consistency, the Ability Test Type, and the I = 20 Subscale Length conditions with large SDs were expected to have larger RMSE than the Low Subscale Consistency, the Ability Test Type, and the I = 10 Subscale Length conditions, respectively, simply due to the SD effect. In fact, RMSE-MB from different methods was greater error in I = 20 than I = 10, in the High Subscale Consistency than the Low Subscale Consistency. Unlike the other three conditions, there were no differences in the magnitude of SDs across different between-subscale correlations. Between-subscales correlation conditions also affected the amount of errors, although the impact substantially varied across different methods. For example, although the Haberman and the Kelley methods yielded less error than raw subscale scores in all between-subscales correlations, the HH method yielded less error than raw subscale scores only in r = 0.9. Overall results of RMSE-MB demonstrate that different data conditions may yield different amount of error at different degree depending on the subscale scoring methods. Although the Kelley's method yielded less error, it was caused by decreased SDs. Kelley's simply shrank scores toward the mean, resulting in reduced SDs and yielding less RMSE-MB.

PRMSE-MB, excluding the impact of SDs, could be used to compare among various simulation conditions with different SDs. Based the PRMSE-MB results above, long subscale length and high subscale consistency conditions resulted in less error (i.e., high reliability) than

short subscale length and low subscale consistency conditions, respectively. As expected from classical test theory developments, accuracy of subscale scores was obtained in the I = 20, the High Subscale Consistency, and the Ability Test Type conditions.

CTT-based Subscale Score Accuracy: RMSE-SB

True subscale θs were transformed into the expected true subscale scores so that they are comparable to the CTT-based subscale scores. Then, RMSE-SB were obtained by computing difference between the expected true subscores and the CTT-based subscale scores. The expected true subscale scores (i.e., the mean in a replication), S_{ET}, are computed by using the following equation:

$$S_{ET} = \frac{1}{N}\sum_{i=1}^{I} P_i(\theta_j),$$

where *N* is the sample size, *I* is the number of items in a subscale, and $P_i(\theta_j)$ is the probability that an examinee with the trait subscale score θ_j answers item *i* correctly. The means and the standard deviations of expected true subscale scores across replications for achievement and ability tests are, respectively, shown in Tables A22 and A23. The following Table 4.8 includes the expected true subscale scores averaged across four subscale scores. The mean of expected true subscale scores were the same as that of raw subscale scores, and the standard deviations of expected true subscale scores were somewhat low relative to those of the raw subscale scores.

Table 4.8. Descriptive Statistics for Expected True Subscale Scores in Various Simulation Conditions

Test type	Subscale length	Subscale consistency	Between-subscales correlation	Mean	SD
Achievement	I = 10	High	*r* = 0.3	6.92	2.00
			r = 0.6	6.92	2.00
			r = 0.9	6.90	2.01
		Low	*r* = 0.3	6.50	1.58

			r = 0.6	6.49	1.59
			r = 0.9	6.48	1.59
		High	r = 0.3	13.82	3.99
			r = 0.6	13.79	4.01
	I = 20		r = 0.9	13.80	4.01
		Low	r = 0.3	12.96	3.17
			r = 0.6	12.99	3.17
			r = 0.9	12.95	3.17
Ability		High	r = 0.3	5.02	2.25
			r = 0.6	5.01	2.25
	I = 10		r = 0.9	5.02	2.25
		Low	r = 0.3	5.02	1.71
			r = 0.6	5.00	1.71
			r = 0.9	5.00	1.71
		High	r = 0.3	10.03	4.50
			r = 0.6	10.02	4.50
	I = 20		r = 0.9	9.99	4.50
		Low	r = 0.3	9.95	3.42
			r = 0.6	9.97	3.42
			r = 0.9	9.99	3.42

In order to determine the plausibility of using expected true subscale scores in place of true subscale θs for the RMSE-SB computation, the correlation between the true subscale θs and the expected true subscale scores was calculated. Tables A24 and A25 include the correlations of raw subscale scores (i.e., summed subscale scores) and true subscale scores θs with expected true subscale scores, respectively from achievement and ability tests. The results showed that the true subscale θs and expected true subscale scores are highly correlated in the corresponding subscales, ranging between 0.97 and 1.00. Moreover, the structure of correlations among expected true scores from different subscales were highly consistent with that of correlations among true trait subscale θs. For example, in the simulation condition where correlations among trait subscale θs are assumed to be 0.3, the off-diagonal correlations among the expected true subscale scores and the trait subscale scores were also 0.3. Therefore, the use of the expected true subscale scores in the RMSE-SB computation seems to be reasonable.

Due to the difference of measurement in among subscale scores from different subscale length or subscale consistency, all subscale scores for computing the RMSE-SB were standardized with their subscale score means and standard deviations in each condition and each replication. The RMSE-SB means were obtained by averaging RMSE-SBs over 100 replications in each condition. Tables A26 and A27 include standardized RMSE-SB means based on different subscale scoring methods under varied conditions, respectively from achievement and ability tests. The results showed that the amount of standardized RMSE-SB varied across different methods and different data conditions. Also, they seemed to have interactive impacts on RMSE-SB. Table 4.9 below present the results of standardized RMSE-SB means for four subscales. The results were very similar to the results from the RMSE-MB.

First, subscale scoring methods yielded different amount of approximation error, depending on various data conditions within a method. In general, RMSE-SB means were smaller in the I = 20 Subscale Length, and the Ability Test Type conditions, with the other conditions fixed. Specifically, the lowest RMSE-SB means from the raw subscale scores were found in the Ability Test Type, I = 20 Subscale Length, High Subscale Consistency condition (i.e., 0.40), and the highest RMSE-SBs were found in the Achievement Test Type, I =10 Subscale Length, and Low Subscale Consistency condition (i.e., 0.71). The Kelley's method yielded the exactly same results as the Raw Subscale Scoring Method. The Holland-Hoskens' (HH) method yielded the lowest RMSE-SBs in the I = 20 Subscale Length, the High Subscale Consistency and the Between-subscales Correlation of $r = 0.9$ condition (i.e., 0.36) regardless of test types. Also, the HH method yielded the highest RMSE-SBs in the Achievement Test Type, the I = 10 Subscale Length, the Low Subscale Consistency, and Low Between-subscale Correlation condition of $r = 0.3$ (i.e., 0.92). Lastly, the Haberman's weighted average method

yielded the lowest RMSE-SB in the Ability Test Type, High Subscale Consistency, I = 20 Subscale Length, Between-subscales Correlation of r = 0.9 condition (i.e., 0.31), and the highest RMSE-SB in the Achievement Test Type, I = 10 Subscale Length, Low Subscale Consistency, and Low Between-subscales Correlation of *r* = 0.3 condition (i.e., 0.69).

Compared to the results from raw subscale scoring method, Only Haberman's methods consistently yielded less approximation error with smaller RMSE-SBs in all simulation conditions. The Haberman's method considerably reduced the amount of error, although the amount of reduction in error varied across different simulation conditions. Specifically, most large reduction in RMSE-MBs were observed in the Low Subscale Consistency and the Between-subscales Correlation of *r* = 0.9 conditions. In most simulation conditions, the HH method yielded even higher approximation error on average than the raw subscale scoring method, indicating that the HH method may not provide more reliable estimates. However, their performance substantially varied depending on the size of correlations among scores from different subscale scales. In the Between-subscales Correlation conditions where *r* = 0.9, the HH method yielded considerably decreased RMSE-MBs compared to those from the raw subscale method, whereas they yielded much higher RMSE-MB in the Between-subscales Correlation conditions where *r* = 0.3, or *r* = 0.6. The Kelley's method yielded the same amount of RMSE-SB as the raw subscale scoring method.

Table 4.9. RMSE-SB Means from CTT Subscale Scores

Test type	Subscale length	Subscale consistency	Between-subscales correlation	$RMSE - SB_{Raw}$	$RMSE - SB_{Kelley}$	$RMSE - SB_{HH}$	$RMSE - SB_{Haberman}$
Achievement	I = 10	High	*r* = 0.3	0.56	0.56	0.88	0.55
			r = 0.6	0.56	0.56	0.67	0.51
			r = 0.9	0.56	0.56	0.42	0.40

		Low	$r = 0.3$	0.71	0.71	0.92	0.69
			$r = 0.6$	0.71	0.71	0.73	0.63
			$r = 0.9$	0.71	0.71	0.50	0.49
	I = 20	High	$r = 0.3$	0.42	0.42	0.84	0.42
			$r = 0.6$	0.42	0.42	0.63	0.39
			$r = 0.9$	0.42	0.42	0.36	0.32
		Low	$r = 0.3$	0.55	0.55	0.87	0.54
			$r = 0.6$	0.55	0.55	0.66	0.50
			$r = 0.9$	0.55	0.55	0.42	0.39
Ability	I = 10	High	$r = 0.3$	0.54	0.54	0.86	0.53
			$r = 0.6$	0.54	0.54	0.66	0.49
			$r = 0.9$	0.54	0.54	0.41	0.39
		Low	$r = 0.3$	0.69	0.69	0.91	0.67
			$r = 0.6$	0.69	0.69	0.72	0.61
			$r = 0.9$	0.69	0.69	0.49	0.48
	I = 20	High	$r = 0.3$	0.40	0.40	0.83	0.40
			$r = 0.6$	0.40	0.40	0.62	0.38
			$r = 0.9$	0.40	0.40	0.36	0.31
		Low	$r = 0.3$	0.54	0.54	0.86	0.53
			$r = 0.6$	0.54	0.54	0.65	0.49
			$r = 0.9$	0.54	0.54	0.41	0.39

Repeated Measures ANOVA: Comparisons of RMSE-SB

The mean differences in RMSE-SB values across various simulation conditions within methods were examined using the Repeated Measure procedure. Subscale scoring methods were repeated measure factors, and Test type, Subscale Consistency, and Between-subscales correlation variables were between-group factors. Because subscale items within a test were generated under the same conditions and their resulting subscale scores were mostly the same across subscales, they yielded the equivalent RMSE-SB means across subscales. Thus, RMSE-SB means were averaged over subscales, and the average RMSE-SB means were used to conduct the repeated measures analysis.

Table 4.10 below presents results of repeated measures for RMSE-SBs based on different methods across different simulation conditions. Because raw subscale scores and Kelley's scores

are the exactly same, Kelley's scores were excluded in the analyses. The results show that all within- and between-groups effects were statistically significant. First, RMSE-SBs based on different methods (i.e., raw, HH, and Haberman methods) were statistically and significantly different (i.e., main effect), and the amount of RMSE-SB from different methods varied across all four simulation conditions (i.e., two-way interaction effects). The different effect sizes of factors by squared partial eta, η_p^2 were observed. The most influential effects on the amount of RMSE-SB were, particularly, found in the main effect of Method (i.e., 0.99), two-way interaction of Method x Subscale Consistency (i.e., 0.98), Method x Subscale Length (i.e., 0.98), and Method x Between-subscales Correlation (i.e., 0.99), and three way interaction effect of, Method x Subscale Consistency x Between-subscales Correlation (i.e., 0.85), and Method x Subscale Length x Between Subscale Correlation (i.e., 0.85). Although several four-way and five-way interactions were also statistically significant, their effect sizes were trivial. The following four figures, Figure 16, through Figure 19, correspond to the results of two-way interactions in which the Test Type, the Subscale Consistency, the Subscale Length, and the Between-subscale Correlation conditions are, respectively, involved.

Table 4.10. Test of Repeated Measures of RMSE-SBs for Subscale Score Estimates

Source	df	SS	MS	F_{obs}	p	η_p^2
Method	2	37.079	18.539	1,494,700.83	<0.01	**0.99**
2 way interaction						
Method*TestType	2	0.019	0.010	774.98	<0.01	0.25
Method*SubscaleConsistency	2	2.701	1.350	108,862.08	<0.01	**0.98**
Method*SubscaleLength	2	2.597	1.299	104,707.16	<0.01	**0.98**
Method*BetweenSubscaleCorr	4	42.300	10.575	852,578.46	<0.01	**0.99**
3 way interaction						
Method*TestType*SubscaleConsistency	2	0.000	0.000	13.43	<0.01	0.00
Method*TestType *SubscaleLength	2	0.000	0.000	1.61	>0.01	0.00
Method*TestType*BetweenSubscaleCorr	4	0.002	0.001	50.09	<0.01	0.04
Method*SubscaleConsistency*	2	0.004	0.002	180.11	<0.01	0.07

SubscaleLength						
Method*SubscaleConsistency* BetweenSubscaleCorr	4	0.338	0.084	6,808.39	<0.01	**0.85**
Method*SubscaleLength* BetweenSubscaleCorr	4	0.342	0.085	6,891.30	<0.01	**0.85**
4 way interaction						
Method*TestType*SubscaleConsistency* SubscaleLength	2	0.000	0.000	2.28	>0.01	0.00
Method*TestType*SubscaleConsistency* BetweenSubscaleCorr	4	0.000	0.000	3.72	<0.01	0.00
Method*TestType*SubscaleLength* BetweenSubscaleCorr	4	0.000	0.000	0.69	>0.01	0.00
Method* SubscaleConsistency* SubscaleLength*BetweenSubscaleCorr	4	0.003	0.001	63.92	<0.01	0.05
5 way interaction						
Method*TestType*SubscaleConsistency * SubscaleLength*BetweenSubscaleCorr	4	0.000	0.000	1.08	>0.01	0.00
Error (Method)	4,752	0.059	0.000			

Figure 4.15 compares the standardized RMSE-SB means based on different methods in the Ability vs. Achievement Test Type conditions. Regardless of test types, RMSE-SB means were somewhat higher in the Achievement Test Type than in the Ability Test Type, although the RMSE-SB difference seems to be small. The lowest RMSE-SB was found in the Haberman's method, and the highest RMSE-SB means was observed in the HH method. RMSE-SB were lower in order of Haberman < Raw < HH in both test type conditions. Notice that only Haberman's method yielded less RMSE-SBs than the Raw subscale

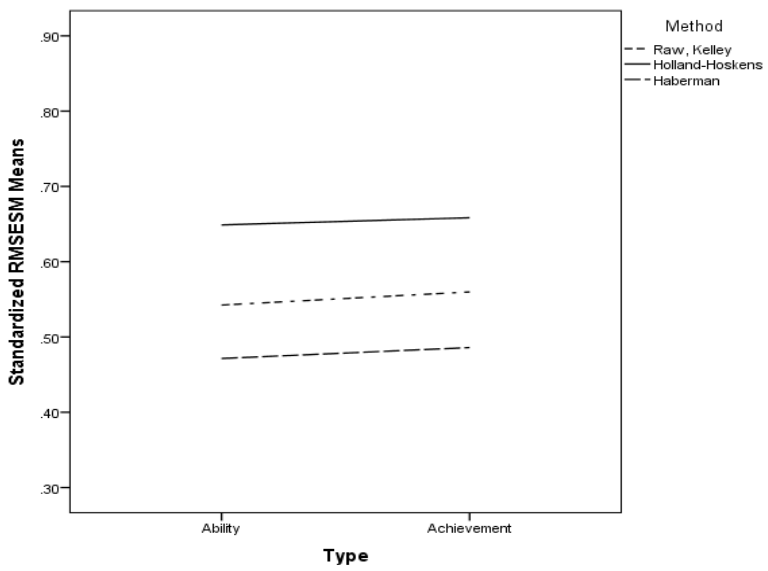

Figure 4.15. RMSE-SB Means in Distinct Test Types: Ability vs. Achievement Tests

Figure 4.16 below compares the standardized RMSE-SB means across different methods depending on the different subscale consistency: High vs. Low Subscale Consistency conditions. All subscales scoring method yielded lower RMSE-SB means in the High Subscale Consistency condition than in the Low Subscale Consistency condition. Although Haberman method yielded less RMSE-SB than the raw subscale scoring method, HH method made higher RMSE-SB means than the raw subscale scoring.

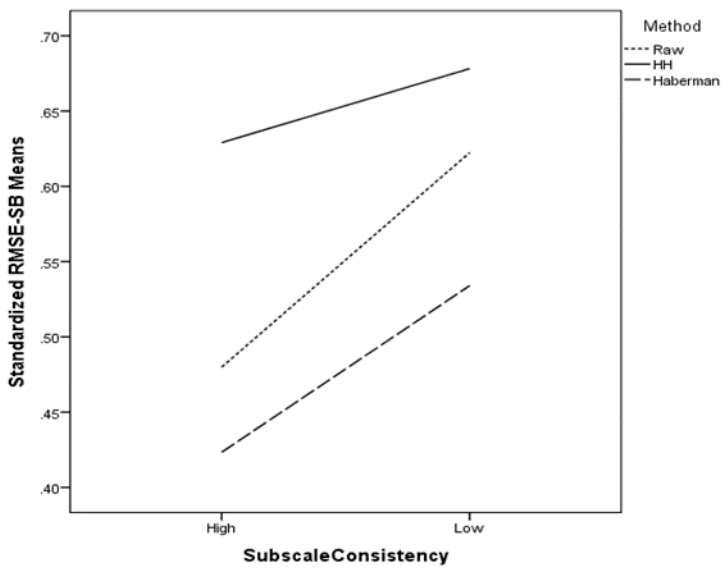

Figure 4.16. RMSE-SB Means in Different Subscale Consistency Conditions: High vs. Low

Next, Figure 4.17 presents the standardized RMSE-SB means from different methods, depending on different subscale lengths. In all three methods, RMSE-SB means were smaller in I = 20 than in I = 10. Haberman method yielded lowest RMSE-SB than the raw subscale scoring method in both subscale length conditions, and HH method yielded highest RMSE-SB than raw subscale scoring.

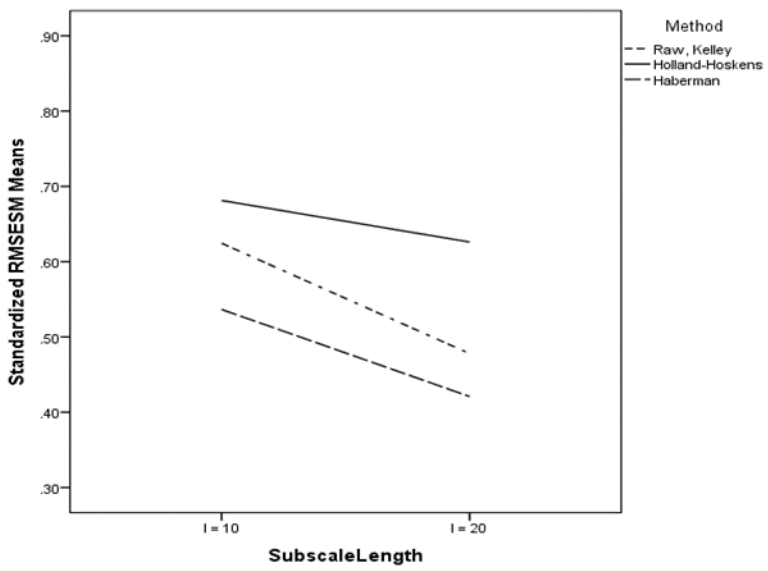

Figure 4.17. RMSE-SB Means in Different Subscale Length Conditions: I = 10 vs. I = 20

Figure 4.18 shows the standardized RMSE-SB means from different methods depending on the three Between-subscales Correlation conditions: r = 0.3, r = 0.6, and r =0.9. Across three types of between-subscales correlation conditions, Haberman's methods always yielded lower RMSE-SB means than the raw subscale scoring method, and the decrease in RMSE-SB was the largest in the Between-subscales Correlations of r = 0.9. Although HH method yielded substantially higher RMSE-SBs among methods in r = 0.3 conditions than raw subscale scoring, it yielded RMSE-SBs as low as those from Haberman's method in the Between-subscales Correlation condition of r = 0.9.

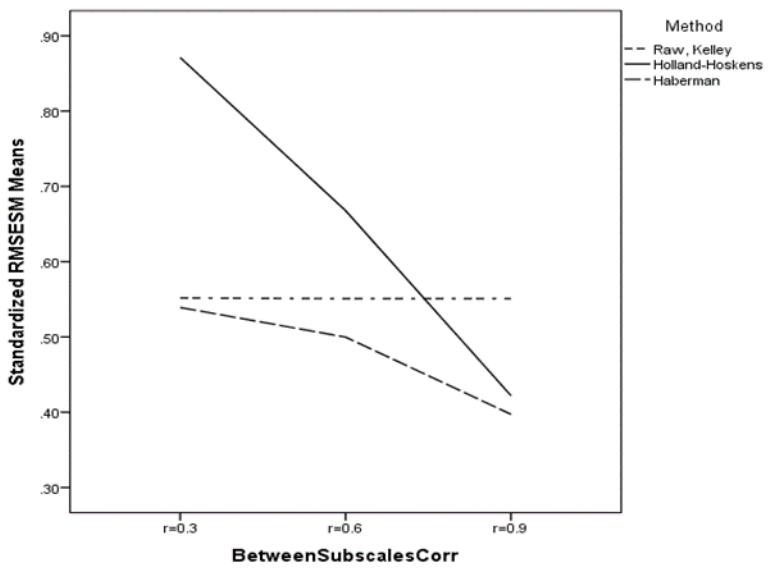

Figure 4.18. RMSE-SB Means in Different Between-subscales Correlation Conditions: r = 0.3 vs. v = 0.6 vs. r = 0.9

The following figures, Figure 4.19 through Figure 4.22, present results of three-way interactions, respectively of Method x different Subscale Consistency x Between-subscales Correlation, and Method x Subscale Length x Between Subscales Correlation on RMSE-SB means. They present how the standardized RMSE-SB means based on different methods perform depending on the combinations of other two between-group factors.

Figure 4.19 and Figure 4.20 show the standardized RMSE-SBs from different methods across Between-subscales Correlations, respectively in the High vs. Low Subscale Consistency conditions. The noticeable difference in patterns between two subscale consistency conditions was especially found in the performance of the HH method. From both graphs, RMSE-SB from the HH methods was very large in r = 0.3, but low when r = 0.9. However, the patterns in different subscale consistency conditions differed in the Between-subscales Correlations. The RMSE-SB from the HH method were considerably high in r = 0.3 in the High Subscale

Consistency condition, but the difference in RMSE-SB among different methods were relatively small in the Low Subscale Consistency condition.

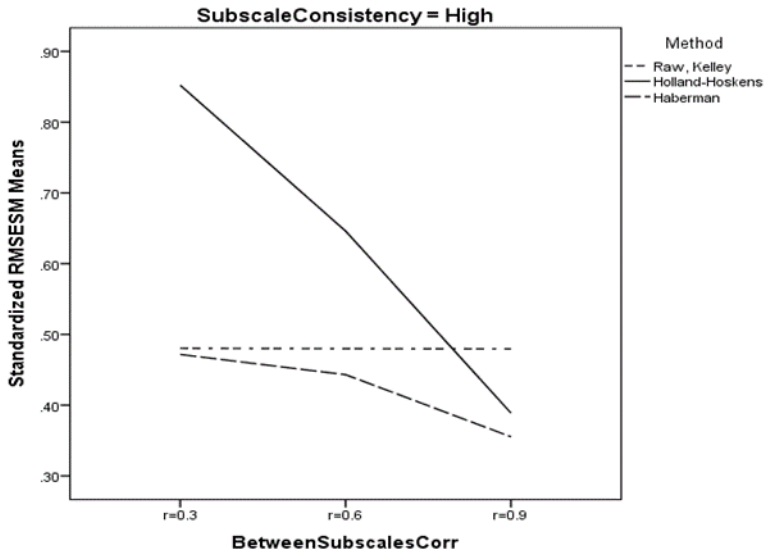

Figure 4.19. RMSE-SB Means across Different Between-subscales Correlation in the High Subscale Consistency Condition

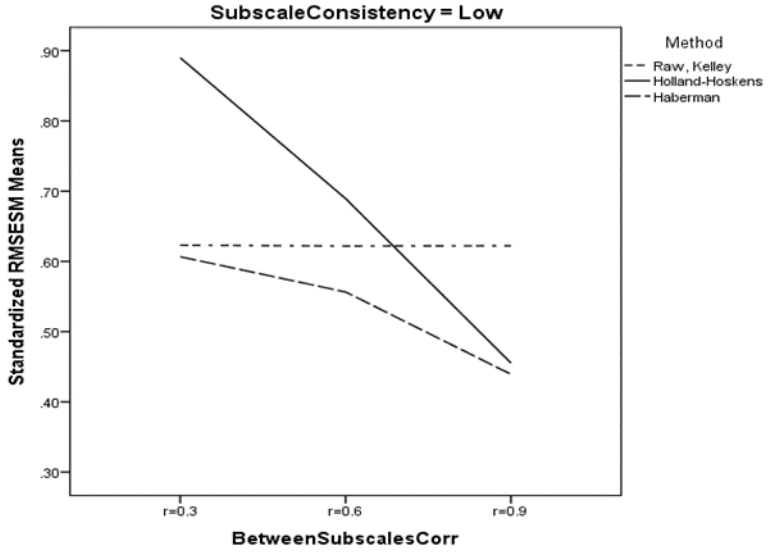

Figure 4.20. RMSE-SB Means across Different Between-subscales Correlation in the Low Subscale Consistency Condition

The following figures, Figure 21 and Figure 22, shows the standardized RMSE-SB means across different between-subscales correlations, respectively in I =10 and I = 20 Subscale Length conditions. From the graphs, RMSE-SB means from the HH method greatly differed across between-subscale correlations. Although the patterns were very similar in both subscale length conditions: I = 10 vs. I = 20. The deviations of RMSE-SB means across different subscale lengths seemed to be large in Subscale length of I = 10.

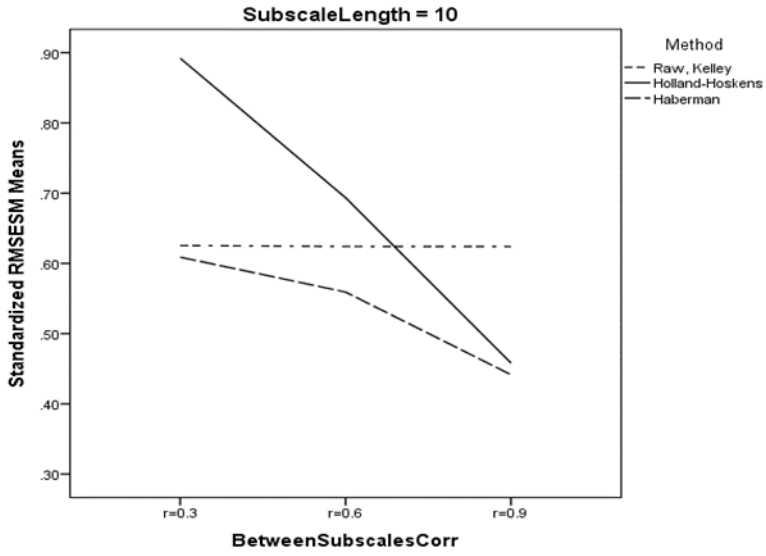

Figure 4.21. RMSE-SB Means across Different Between-subscales Correlation in the I = 10 Subscale Consistency Condition.

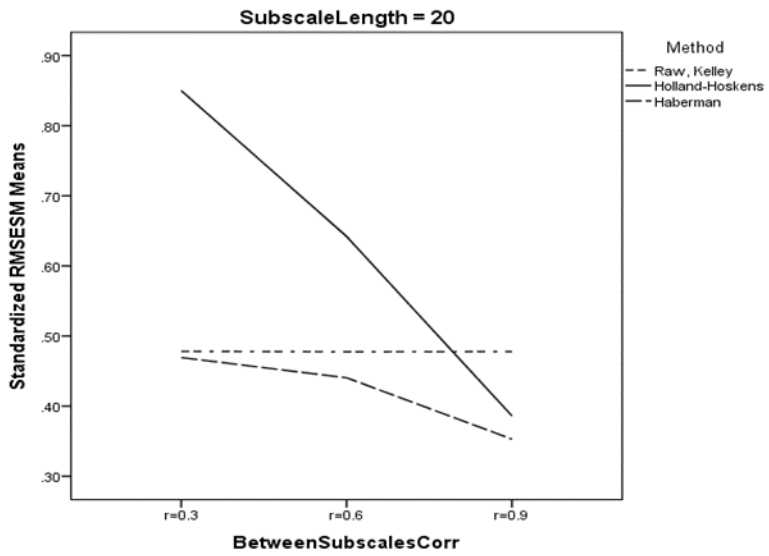

Figure 4.22. RMSE-SB Means across Different Between-subscales Correlation in the I = 20 Subscale Consistency Condition.

Approximation Errors from IRT-based Subscale Scores

As an index of the accuracy of estimation from IRT, RMSE-SBs were compared. RMSE-SB was computed by obtaining the squared mean difference of true subscale θs and estimated IRT subscale θs in each replication. Tables A28 through A29 include the results of RMSE-SBs from two different methods across the simulation conditions.

Table 4.11 includes RMSE-SB Means for IRT subscale scores. In all conditions, MIRT-2PL yielded smaller RMSE-SBs than UIRT-2PL, indicating that MIRT-2PL models estimate subscale scores with more precision. Comparing two Test Type conditions, ability test yielded somewhat lower RMSEs than achievement test, although the difference was very trivial. Other three conditions, Subscale Length, Subscale Consistency, and Between-subscales correlation, showed more substantial difference in RMSE-SBs. Large difference in the amount of RMSE-SBs across models were, especially, found in the Subscale Length and the Subscale Consistency conditions. For example, RMSE-SB values were averagely 0.57 in the I = 10 Subscale Length

condition and 0.46 in the I = 20 Subscale Length condition, with other conditions fixed. Also, RMSE-SB were, on average, 0.57 in the Low Subscale Consistency condition, and 0.47 in the High Subscale Consistency condition, with other conditions fixed, as well. However, without considering the type of methods, the lowest RMSE-SBs were found in the r = 0.9 Between-subscales Correlation conditions, with an average of 0.49 across simulation conditions. In contrast, the highest RMSE-SBs were found in the r = 0.3 Between-subscales Correlation conditions, with an average of 0.54 across simulation conditions. However, this difference in the amount of RMSE-SB was wholly due to the RMSE-SB from the multidimensional IRT model. While RMSE-SB means based on the unidimensional IRT scores were constant across (i.e., 0.55), the RMSE-SB means based on the multidimensional IRT scores varied depending on the type of Between-subscales Correlation conditions. For example, the RMSE means were 0.43 in the r = 0.3 Between-subscales Correlation condition, whereas they were 0.53 in the r = 0.9 Between-subscales Correlation condition. When the MIRT-2PL model is applied to the data whose correlations get high, the MIRT-2PL model may perform better than the unidimensional model.

Table 4.11. RMSE-SB Means from IRT scores in Various Simulation Conditions

Test type	Subscale length	Subscale consistency	Between-subscales Correlation	UIRT-2PL	MIRT-2PL
Achievement	I = 10	High	r = 0.3	0.57	0.56
			r = 0.6	0.57	0.52
			r = 0.9	0.57	0.45
		Low	r = 0.3	0.68	0.66
			r = 0.6	0.68	0.61
			r = 0.9	0.68	0.50
	I = 20	High	r = 0.3	0.45	0.44
			r = 0.6	0.45	0.42
			r = 0.9	0.45	0.40
		Low	r = 0.3	0.54	0.53
			r = 0.6	0.54	0.50

			r = 0.9	0.54	0.44
Ability	I = 10	High	r = 0.3	0.54	0.53
			r = 0.6	0.54	0.49
			r = 0.9	0.53	0.42
		Low	r = 0.3	0.66	0.64
			r = 0.6	0.66	0.59
			r = 0.9	0.66	0.48
	I = 20	High	r = 0.3	0.41	0.40
			r = 0.6	0.41	0.39
			r = 0.9	0.42	0.36
		Low	r = 0.3	0.53	0.51
			r = 0.6	0.53	0.48
			r = 0.9	0.52	0.41

Repeated Measures ANOVA: Comparisons of RMSE-SBs

The mean differences of RMSE-SBs from two IRT-based scores: Unidimensional IRT-2PL vs. Multidimensional IRT-2PL (i.e., UIRT-2PL vs., MIRT-2PL) were examined using the Repeated Measure analysis. Table 4.12 presents test results for repeated measures of the RMSE-SB means. The results indicate that RMSE-SB means from the two models were statistically significantly different, with $F_{(2,2376)} = 38{,}382.27, p < 0.01, \eta_p^2 = 0.94$. Specifically, RMSE-SBs were lower in the MIRT-2PL model than the unidimensional 2PL model, indicating that the MIRT-2PL scores have more accuracy.

RMSE-SB means from the two models were statistically significantly different, depending on three simulation conditions: Subscale Length ($F_{(1,2376)} = 2{,}704.43, p < 0.01, \eta_p^2 = 0.53$), Subscale Consistency ($F_{(1,2376)} = 1{,}955.42, p < 0.01, \eta_p^2 = 0.45$), or Between-subscales correlation ($F_{(1,2376)} = 10{,}134.79, p < 0.01, \eta_p^2 = 0.90$). These results show that RMSE-SB means for subscale scores from different models significantly differ across the Subscale Length, the Subscale Consistency, and the Between-subscales Correlation conditions. However, RMSE-SB means were not statistically significantly different across Test Type conditions. All types of three-way interaction effects were statistically significant, but only two

of six sources had significantly large effect size, Method * Subscale Consistency * Between-subscale Correlation (i.e., η_p^2 = 0.36) and Method * Subscale Length * Between-subscale Correlation (η_p^2 = 0.43). Although two sources of four-way interaction effect were also statistically significant, their effect sizes were ignorable, providing only a little affects the total variance of RMSE-SBs.

Table 4.12. Test of Repeated Measures of PRMSE-SBs for Subscale Score $\boldsymbol{\theta}$s

Source	df	SS	MS	F_{obs}	p-value	η_p^2
Method	1	4.152	4.15	38,382.27	<0.01	**0.942**
2 way interaction						
Method*TestType	1	0.001	0.001	6.14	>0.01	0.003
Method*SubscaleConsistency	1	0.212	0.21	1,955.42	<0.01	**0.451**
Method* SubscaleLength	1	0.293	0.29	2,704.43	<0.01	**0.532**
Method*BetweenSubscaleCorr	2	2.193	1.10	10,134.79	<0.01	**0.895**
3 way interaction						
Method* TestType * SubscaleConsistency	1	0.002	0.001	19.47	<0.01	0.008
Method* TestType * SubscaleLength	1	0.001	0.001	12.17	<0.01	0.005
Method* TestType * BetweenSubscaleCorr	2	0.002	0.001	7.93	<0.01	0.007
Method* SubscaleConsistency * SubscaleLength	1	0.001	0.001	9.87	<0.01	0.004
Method* SubscaleConsistency * BetweenSubscaleCorr	2	0.145	0.07	669.97	<0.01	**0.361**
Method* SubscaleLength * BetweenSubscaleCorr	2	0.196	0.10	906.62	<0.01	**0.433**
4 way interaction						
Method* TestType * SubscaleConsistency * SubscaleLength	1	0.001	0.001	6.78	<0.01	0.003
Method* TestType * SubscaleConsistency * BetweenSubscaleCorr	2	0.001	0.001	6.76	<0.01	0.006
Method* TestType * SubscaleLength * BetweenSubscaleCorr	2	0.003	0.001	11.78	<0.01	0.010
Method* SubscaleConsistency * SubscaleLength * BetweenSubscaleCorr	2	0.001	0.000	4.28	>0.01	0.004
5 way interaction						
Method* TestType * SubscaleConsistency *	2	0.002	0.001	7.20	>0.01	0.006

SubscaleLength * BetweenSubscaleCorr			
Error (Method)	2,376	0.257	0.000

Figure 4.23 presents graphs for four types of two-way interaction effects: Method * Test Type, Method * Subscale length, Method * Subscale Consistency, and Method * Between-subscales Correlation. RMSE-SB means are not seen as having much difference across different test types, but RMSE-SB means seem to substantially differ in different Subscale Length, Subscale Consistency, and Between-subscale Correlation conditions. Specifically, the RMSE-SBs were a little bit lower in ability test than achievement test. Also, the RMSE-SBs were lower in the I = 20 Subscale Length condition than the I = 20 Subscale Length condition across models, where the MIRT-2PL had lower RMSE-SB. When it comes to the Subscale Consistency condition, RMSE-SB means were lower in the High Subscale Consistency condition than in the Low Subscale Consistency condition. Similarly, the MIRT-2PL had lower RMSE-SB values than the unidimensional model in both subscale consistency conditions. Comparing the RMSE-SBs across different between-subscales correlations, as mentioned earlier, the unidimensional 2PL model performed worse than the MIRT-2PL in all between-subscales correlations. Although the MIRT-2PL model always performed better than the unidimensional 2PL model, the degree of how better it performed differed across different between-subscales correlations. In addition, the MIRT-2PL model yielded lower RMSE-SBs in all conditions. However, RMSE-SBs from the MIRT-2PL model were dramatically reduced, as Between-subscales Correlations get higher. That is, the MIRT-2PL had greatly lowered RMSE-SBs relative to the UIRP-2PL in the $r = 0.9$ Between-subscales Correlation condition. On the other hand, the difference for RMSE-SB was very small in the $r = 0.3$ Between-subscales Correlation condition.

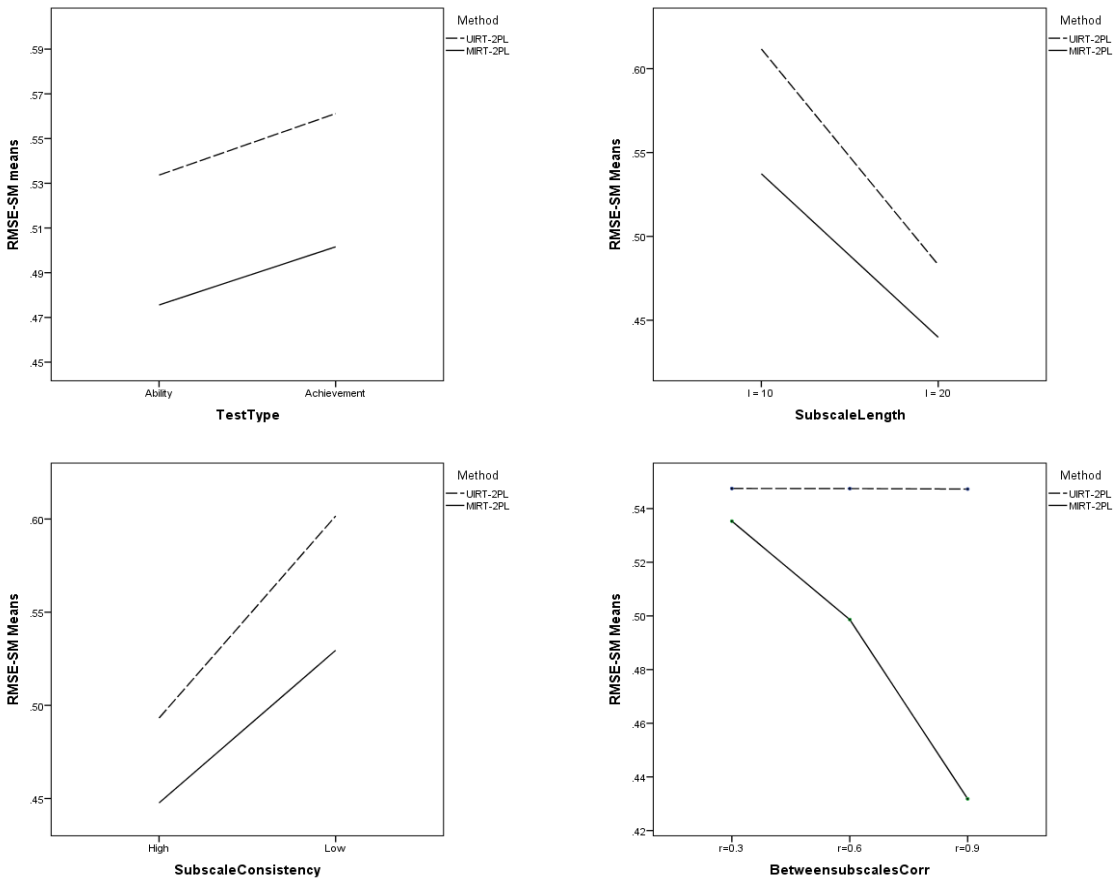

Figure 4.23. Two-way Interaction Effects of RMSE-SBs in Each Between-group factor: Test Type, Subscale Length, Subscale Consistency, and Between-subscale Correlation

Correlations between Estimated and True Subscale Scores

Correlations between simulated true θs and estimated subscale scores were observed. Table A30 through Table A33 present the correlations between true θs and subscale scores estimated through multiple scoring methods in each condition. Observing correlations with true subscale scores will be assisted in determining which method can yield better estimates which approximate true subscale scores.

Table A30 shows correlations between estimated and true subscale scores in Achievement tests and I = 10. The size of correlation generally was larger in High internal consistency than in Low internal consistency with Between-Subscale scores correlations and Subscale scoring methods fixed. Within the same internal consistency conditions, different subscale scoring methods seem to fluctuate based on the size of Between-subscale scores correlations. In r = 0.3, HH and OPI methods had the lowest correlations with true subscale scores and all other methods had relatively high correlations with true subscale scores. Low correlations in HH and OPI methods makes sense if considering that the first two types of subscale scoring method (i.e., HH and OPI methods) combine information from the total test scores. But such tendency seemed to be different in other between-subscale scores correlations. For example, in r = 0.6, in which correlations between subscale scores gets somewhat higher, HH and OPI method showed improved correlations, and Haberman and MIRT-2PL methods showed better correlations than those from other methods. Such high correlations seem to come from the fact that Haberman method borrows some information from total scores and the MIRT-2PL considers correlations among subscale scores. Lastly, in r = 0.9, HH, Haberman, MIRT-2PL, and OPI methods showed high correlations compared to other methods. It also makes sense

when we consider that HH, Haberman, and OPI methods combine some or whole information from total test scores and the MIRT-2PL consider relationships between subscale scores.

Table A31 shows correlations between estimated and true subscale scores in Achievement tests and I = 20. Most of all, correlations between estimated and true scores in I = 20 conditions were generally higher than in those in I = 10 conditions. For example, while average correlations between subscale scores were 0.79 in I = 10 conditions, they were 0.85 in I = 20 conditions, indicating that estimates from a subscale including more item is more accurate. Similar as in I = 10 conditions, the size of correlation was slightly larger in High internal consistency than in Low internal consistency with Between-Subscale scores correlations and Subscale scoring methods fixed. In both High and Low internal consistency conditions, when r = 0.3, HH and OPI methods had the lowest correlations with true subscale scores and all other methods had relatively high correlations with true subscale scores. Such tendency was different in other between-subscale scores correlation conditions. For example, in r = 0.6, in which correlations between subscale scores gets somewhat higher, HH and OPI method showed improved correlations, and Haberman and MIRT-2PL methods showed better correlations than those from other methods. In turn, in r = 0.9, HH, Haberman, MIRT-2PL, and OPI methods showed high correlations compared to other methods.

Similar patterns were also observed in Ability test conditions. Correlations between estimated and true scores from Ability test conditions were similar to those from Achievement test conditions. Specifically, average correlations between subscale scores were 0.82 in Achievement test conditions, and they were 0.83 in Ability test conditions, indicating that the accuracy of estimates was similar regardless of difficulty levels of items in a subscale. Table A32 and A33, present correlations between true subscale scores and estimates from different models.

The relative size and patterns of correlations between true subscale scores and estimated subscale scores based on different method was analogous as in the Achievement test conditions. For example, correlations between estimated and true subscale scores were larger (r = 0.86) in I = 20 than in I = 10 (r = 0.80). Also, HH and OPI scores showed high correlations with true subscale scores in r = 0.9 condition, but they had low correlations in r = 0.3 condition. In sequence, subscale score estimates from the Haberman method and the MIRT-2PL generally showed high consistency with true subscale scores in most conditions.

The following four figures, Figure 4.24 through 4.27, compare true and estimated subscale scores in different Test type, Subscale length, Between-subscales correlation, and Subscale consistency. Figure 4.23 presents the magnitude of correlations between true subscale scores and estimated scores from different methods with separate lines in different test types. Generally, Ability tests showed higher correlations than Achievement tests. However, the difference between both test types was not large.

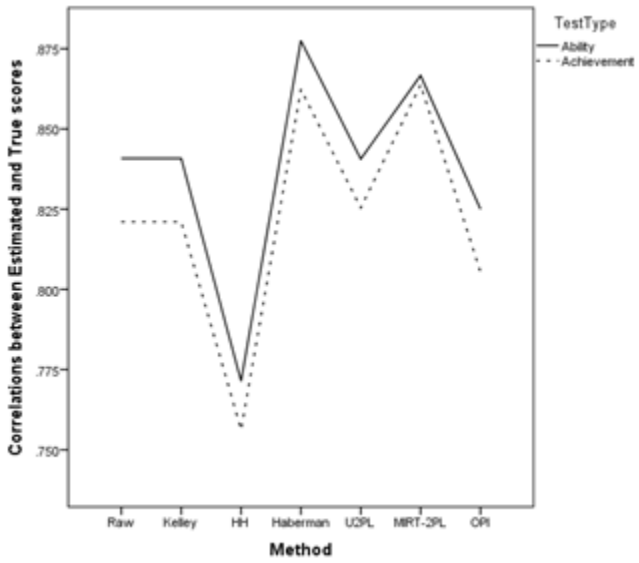

Figure 4.24. Correlations between True Subscale Scores and Estimated Subscale Scores from Different Methods in Distinct Test Types

Figure 4.25 compares correlations of true and estimated subscale scores based on different methods in different subscale length conditions. Generally, the correlations were higher in I = 20 conditions than in I = 10, showing high estimation accuracy when more items are included in a subscale. Comparing among methods, Haberman's method yielded the highest correlations, and the HH method yielded the lowest correlation.

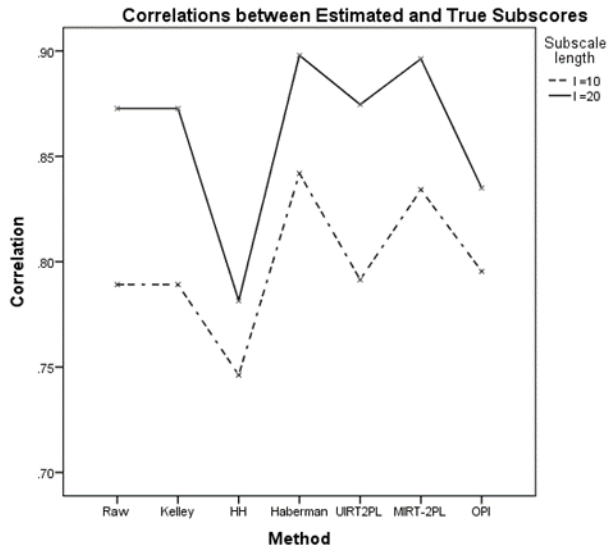

Figure 4.25. Correlations between True Subscale Scores and Estimated Subscale Scores from Different Methods in Different Subscale Length Conditions

The MIRT-2PL also showed similar level of correlation as Haberman's method. In both Subscale Length conditions, the correlations were high in order of Haberman> MIRT-2PL > UIRT-2PL > Raw > Kelley > OPI > HH methods. The Haberman's method and the MIRT-2PL yielded better correlations than raw subscale score and the UIRT-2PL.

Next, correlations in High vs. Low Subscale Consistency conditions were compared, which is shown in Figure 4.26. In each method, the size of the correlation was larger in the High Subscale Consistency condition. The highest correlations were observed in the Haberman's and the MIRT-2PL models, and the lowest correlations were observed in the HH method. Only the

Haberman method and the MIRT-2PL model yielded higher correlations than raw subscale scores and UIRT-2PL, respectively. High correlations between estimated and true subscale scores were found, in order of Haberman > MIRT-2PL > UIRT-2PL > Raw > Kelley > OPI > HH methods.

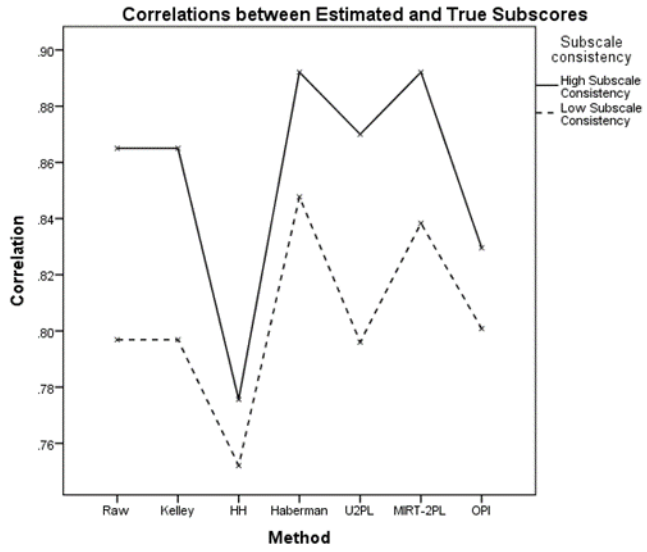

Figure 4.26. Correlations between True Subscale Scores and Estimated Subscale Scores from Different Methods in Different Subscale Length Conditions

Correlations in r = 0.3, 0.6, vs., 0.9 were also compared across different methods as shown in Figure 4.27. The Haberman's method yielded highest correlations, and HH method the lowest correlations. Kelley's and Haberman's methods showed higher correlations than Raw subscale scores. Similarly, the MIRT-2PL yielded higher correlation than the MIRT-2PL. The degree to which the estimated and the true subscale scores are correlated differed among different Between-subscales Correlation conditions. Both the MIRT-2PL and the Haberman's methods showed largely increased correlations in high Between-subscales Correlation condition (i.e., r = 0.9). On the other hand, although the HH and the OPI methods showed lower correlations than raw and UIRT-2PL models, the degree to which the estimated and the true

subscale scores are correlated largely differed across Between-subscale correlations. That is, the HH and the OPI showed high variance of the correlations among different Between-subscale Correlation conditions. Specifically, the HH and the OPI methods showed very low correlations when Between-subscales Correlation is 0.3 and high correlations when Between-subscales Correlation is 0.9. Thus, when Between-subscales Correlation is 0.9, they yielded high correlations over Raw subscale scores and UIRT-2PL subscale scores, High correlations were observed in order of Haberman > MIRT-2PL > UIRT-2PL > Kelley > Raw > OPI > HH methods.

Generally, the results from correlations between the estimated and the true subscale scores showed that high correlations over the Raw subscale score and UIRT-2L were only found in Haberman's and MIRT-2PL models. Although the HH method and OPI method averagely showed low correlations between the estimated subscale scores and the true subscale scores in average, they yielded higher correlations than the raw or the UIRT-2PL subscale scores in the High Between-subscales correlation (i.e., $r = 0.9$).

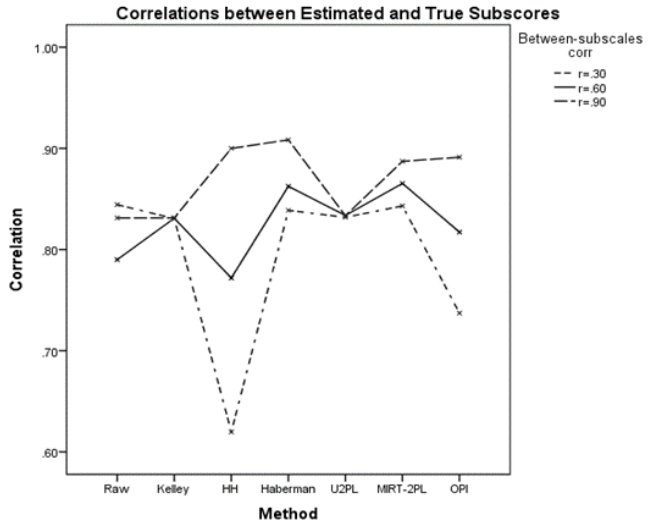

Figure 4.27. Correlations between True Subscale Scores and Estimated Subscale Scores from Different Methods in Different Subscale Correlation Conditions

CHAPTER 5

DISCUSSION

This chapter summarizes the major findings of the study. The primary focus is on both identifying specific data conditions in which subscale scores are more reliable and determining which subscale scoring methods can provide more accurate and reliable than others under various data structures. This chapter also discusses the implications of this study, followed by its limitations and future research.

Findings and Discussion

Alternative subscale scoring methods were employed to yield subscale scores for both real world data and simulation data. These methods were evaluated with respect to their reliability in CTT or accuracy in IRT, primarily by two criteria: root mean square error (RMSE) and correlation of subscale score estimates with true subscale θs. Specifically, two types of RMSE indices were used as evaluation criteria of subscale score accuracy, including the measurement-based RMSE (RMSE-MB; Haberman, 2008) and the simulation-based RMSE (RMSE-SB). RMSE-MB is an index of difference between observed subscale scores and subscale scores estimated based on a model, whereas RMSE-SB computes the difference between subscale score estimated from a model and true subscale scores obtained from simulations. Further, proportional reduction in mean square error (PRMSE-MB) was also used as an additional index for evaluating the reliability of CTT subscale scores. Note that RMSE-SB for CTT subscale scores were standardized with the sample means and standard deviations within each replication and each condition, because RMSE-SB values fluctuate by the unit of measurement.

The accuracy of CTT subscale scores were evaluated using both RMSE-MB and RMSE-SB, where RMSE-MB was obtained by residual of true subscale scores and the true subscale scores based on linear regression models (i.e., $S_t - L(S_t|S_x)$) and RMSE-SB was obtained by residual of model-based subscale scores and expected subscale scores based on the true simulation values. For both indices, the low RMSE indicated better accuracy of subscale score estimation. On the other hand, the low PRMSE-MB indicated low reliability of subscales scores. All RMSE results (i.e., PRMSE-MB and RMSE-SB for CTT subscale scores and RMSE-SB for IRT subscale scores) showed that methods led very similar results in the measurement of subscale score accuracy across simulation conditions.

In general, the RMSE-SB from IRT based scores indicated that the Haberman method and the multidimensional item response model (MIRT-2PL) performed better than the raw subscale scoring in estimating subscale scores. For example, Haberman's method and the MIRT-2PL model yielded the lowest RMSE-MB and the lowest RMSE-SB across simulation conditions, which was significantly smaller than in the raw subscale scoring or unidimensional scoring method. It seems reasonable when considering that Haberman's method and the MIRT-2PL model use more information, respectively, through total score or correlation with other subscales. Though Kelley's method also showed the lower RMSE-MB than the raw subscale scoring method, it did not improve subscale score reliability, not providing any additional advantages over the use of the raw subscale scores. However, the Kelley's method is simply a linear transformation of raw subscale scores, shrinking subscale scores toward the mean. The shrunken subscale scores led to reduced SDs, generating the lower MSE. In most simulation conditions, the HH and the OPI methods did not performed well relative to the raw subscale scoring method, except when subscale scores are highly correlated (i.e., between-subscales

correlations of r = 0.9). The HH and the OPI methods produced more reliable subscale score estimates when the between-subscales correlations are very high. In contrast, these methods had less reliable subscale scores, even lower than the raw scores, when the correlations are very low. Similar results were observed from the correlations of subscale score estimates with their true subscale θs.

The reliability or accuracy of subscale scores from different methods (i.e., models) varied across data conditions. From the hypotheses, described in Chapter 1, we expected that subscale score reliability or accuracy would increase in the following conditions: 1) when a subscale has sufficient number of items, and 2) item scores or item responses are consistent within a subscale and 3) item difficulties are close to .5. Also, we expected that the OPI and the HH methods would perform better when correlations among subscales are very high, because they use total score information. The impact of test types consisting of items with different difficulty levels has not been studied previously, so they were explored in the current study.

The conditions in the study impacted both RMSE-SB from IRT subscale scores and standardized RMSE-SB from CTT subscale scores in the expected manner. That is, subscale scores yielded the lower RMSE when subscale length is longer (20 versus 10 items) and subscale consistency is higher. It was found that the MIRT 2PL model performed better than unidimensional 2PL model, in particular when between-subscales correlations are high (i.e., r = 0.9). Similar results were also supported by investigating the correlation of subscale score estimation with their true trait subscale θs. The results also illustrated that ability test yielded lower RMSE-SB than achievement test. RMSE-SB from CTT subscale scores also showed the same results as these results of IRT scores.

However, results of RMSE-MB from CTT-based subscale scores were seemingly inconsistent with the research hypotheses. CTT-based subscale scores had the low RMSE-MB when subscale length is shorter, and when subscale consistency is low rather than high. Similarly, the ability test yielded slightly higher RMSE-MB than the achievement test. However, it should be noted that RMSE depends on the measurement scale units for the test. Thus, when conditions such as test length, internal consistency and mean item difficulty impact score variances, differences in RMSE will also be found. Note that the Kelley's method, for example, yielded lower RMSE values than the raw subscale scoring method, even though they did not improve reliability of subscale score at all. If considering Kelley's subscale scores are shrunk toward the mean, the RMSE from Kelley's were influenced by the decreased standard deviations. Thus, RMSE cannot be meaningfully compared across levels of the three conditions (i.e., subscale length, subscale consistency, and item difficulty) that impact score variances. Hence, it is important to consider reliability, or the proportional reduction in measurement error. PRMSE-MB measures the reliability of subscale scores, and thus free from the measurement scale unit. Actually, PRMSE-MB results showed high consistency with the research hypothesis. That is, the results report that high reliability was obtained when subscale length is rather longer, and subscale consistency is high.

Implications

The current study has several implications. The current study included several potential factors and subscale scoring models as variables to comprehensively understand the impact of various test data structures on subscale score estimation, and offered some insights for further investigation. The results of this study will help determine the most appropriate data structures

for subscale scores to be reported. When researchers are required to make decisions whether they report subscale scores, this study will guide them what they should consider above all things.

Reliable subscale scoring methods can be worthwhile in providing diagnostic information. Diagnostic information can help teachers design the future instruction or adjust their current lesson, and can help students manage their learning time based on the information. That is, they devote more or less time in review areas their weak areas depending on their relative strength. Also, it may be useful for states and educational institutions to consider results from subscale score patterns in evaluating the effectiveness of their current curricular and considering to fixing their curriculum.

Also, resulting reliable subscale scores can be used as supplementary criteria for selecting the best fitting ones among applicants with the same total score in school, personal selection, and placement. If subscale scores can be precisely estimated, the use of subscale scores seem to be an source of valuable information as supplementary criteria.

Limitations

In order to examine the exact impact of variables on subscale score accuracy, the simulation conditions used in the current study were thoroughly designed and controlled according to the research plan. However, practical testing situations may not be always standardized as in the simulation situation. For example, the subscale scores were assumed to have the same p-value across subscales for convenience of study. However, it may not be certainly possible in real life. Therefore, when the given information is used, careful interpretation and applications are required.

The current study fixed the number of subscales to four, which is the same across all conditions. However, considering that many subscale scoring methods utilize some information

from other subscales or total scores, including more or less subscales in a test can be another factor impacting on subscale accuracy, reflecting the effects of subscale length. Future study is recommended to include various number of subscales within a test.

Furthermore, the current study used two types of RMSE indices for evaluating the CTT-based subscale scores. As mentioned earlier, the indices had disadvantage of fluctuating the RMSE scores depending on the sample variances. Standardized RMSE may eliminate the issue. Future research is recommended to compare the results of standardized RMSE values.

APPENDIX A

SIMULATION DATA RESULTS

Table A1. Means and Standard Deviations for True Person Parameters from Simulated Data

Data condition				Person parameter							
				True theta1		True theta2		True theta3		True theta4	
Test type	Subscale length	Subscale consistency	Correlations between θs	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Achievement	I=10	High	0.3	0.00	1.00	0.00	1.00	-0.00	1.00	-0.00	1.00
			0.6	0.00	1.00	0.00	1.00	0.00	1.00	-0.00	1.00
			0.9	-0.00	1.00	-0.00	1.00	-0.00	1.00	-0.00	1.00
		Low	0.3	-0.00	1.00	0.00	1.00	0.00	1.00	-0.00	1.00
			0.6	-0.00	1.00	-0.00	1.00	-0.00	1.00	-0.00	1.00
			0.9	-0.00	1.00	-0.00	1.00	-0.00	1.00	-0.00	1.00
	I=20	High	0.3	0.00	1.00	0.00	1.00	0.00	1.00	-0.00	1.00
			0.6	0.00	1.00	-0.00	1.00	-0.00	1.00	0.00	1.00
			0.9	0.00	1.00	-0.00	1.00	0.00	1.00	0.00	1.00
		Low	0.3	-0.00	1.00	0.00	1.00	0.00	1.00	-0.00	1.00
			0.6	-0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
			0.9	0.00	1.00	-0.00	1.00	0.00	1.00	-0.00	1.00
Ability	I=10	High	0.3	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
			0.6	-0.00	1.00	-0.00	1.00	-0.00	1.00	0.00	1.00
			0.9	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
		Low	0.3	-0.00	1.00	-0.00	1.00	-0.00	1.00	-0.00	1.00
			0.6	0.00	1.00	-0.00	1.00	0.00	1.00	-0.00	1.00
			0.9	-0.00	1.00	-0.00	1.00	0.00	1.00	-0.00	1.00
	I=20	High	0.3	-0.00	1.00	-0.00	1.00	-0.00	1.00	0.00	1.00
			0.6	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
			0.9	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
		Low	0.3	-0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00
			0.6	-0.00	1.00	-0.00	1.00	0.00	1.00	-0.00	1.00
			0.9	0.00	1.00	0.00	1.00	0.00	1.00	0.00	1.00

Table A2. CTT-based Summary Statistics for Achievement Tests

Data condition							
Subscale length	Subscale consistency	Between-subscales correlation	Subscale	Mean (p)	SD	KR-20	Correlation with total score
I=10	High	0.3	1	6.91(0.69)	2.37	0.71	0.63
			2	6.94(0.69)	2.36	0.71	0.63
			3	6.96(0.70)	2.36	0.71	0.63
			4	6.87(0.69)	2.39	0.71	0.64
			Total	27.68(0.69)	6.05	0.79	-
		0.6	1	6.94(0.70)	2.37	0.71	0.75
			2	6.90(0.69)	2.36	0.71	0.75
			3	6.94(0.69)	2.37	0.71	0.75
			4	6.88(0.69)	2.48	0.71	0.75
			Total	27.66(0.69)	7.13	0.86	-
		0.9	1	6.92(0.69)	2.38	0.71	0.85
			2	6.87(0.69)	2.38	0.71	0.85
			3	6.91(0.69)	2.37	0.71	0.85
			4	6.87(0.69)	2.38	0.71	0.85
			Total	27.57(0.69)	8.13	0.90	-
	Low	0.3	1	6.45(0.65)	2.11	0.56	0.61
			2	6.50(0.65)	2.11	0.56	0.61
			3	6.52(0.65)	2.11	0.56	0.61
			4	6.53(0.65)	2.11	0.56	0.61
			Total	26.00(0.65)	5.19	0.69	-
		0.6	1	6.50(0.65)	2.11	0.56	0.71
			2	6.47(0.65)	2.12	0.56	0.71
			3	6.47(0.65)	2.12	0.56	0.71
			4	6.50(0.65)	2.11	0.56	0.70
			Total	25.95(0.65)	6.00	0.77	-
		0.9	1	6.50(0.65)	2.12	0.56	0.79
			2	6.47(0.65)	2.11	0.56	0.79
			3	6.50(0.65)	2.11	0.56	0.79
			4	6.46(0.65)	2.12	0.56	0.79
			Total	25.93(0.65)	6.73	0.82	-
I=20	High	0.3	1	13.81(0.69)	4.38	0.83	0.65
			2	13.80(0.69)	4.39	0.83	0.66
			3	13.83(0.69)	4.39	0.83	0.66
			4	13.86(0.69)	4.37	0.83	0.65
			Total	55.29(0.69)	11.51	0.89	-
		0.6	1	13.84(0.69)	4.39	0.83	0.78
			2	13.80(0.69)	4.38	0.83	0.78
			3	13.76(0.69)	4.40	0.83	0.78
			4	13.77(0.69)	4.40	0.83	0.78

			Total	55.18(0.69)	13.80	0.93	-
		0.9	1	13.77(0.69)	4.41	0.83	0.90
			2	13.81(0.69)	4.40	0.83	0.90
			3	13.86(0.69)	4.38	0.83	0.90
			4	13.77(0.69)	4.40	0.83	0.90
			Total	55.21(0.69)	15.83	0.95	-
	Low	0.3	1	12.96(0.65)	3.73	0.72	0.64
			2	12.94(0.65)	3.74	0.72	0.64
			3	13.03(0.65)	3.73	0.72	0.64
			4	12.92(0.65)	3.74	0.72	0.64
			Total	51.85(0.65)	9.56	0.81	-
		0.6	1	12.93(0.65)	3.74	0.72	0.76
			2	13.03(0.65)	3.72	0.72	0.75
			3	13.02(0.65)	3.74	0.72	0.76
			4	12.98(0.65)	3.74	0.72	0.76
			Total	51.95(0.65)	11.31	0.87	-
		0.9	1	13.01(0.65)	3.73	0.72	0.86
			2	12.91(0.65)	3.74	0.72	0.86
			3	12.92(0.65)	3.74	0.72	0.86
			4	12.93(0.65)	3.74	0.72	0.86
			Total	51.77(0.65)	12.81	0.90	-

Table A3. CTT-based Summary Statistics for Ability Tests

Data condition							
Subscale length	Subscale consistency	Between-subscales correlation	Subscale	Mean (*p*)	SD	KR-20	Correlation with total score
I=10	High	0.3	1	5.02(0.50)	2.62	0.73	0.64
			2	5.04(0.50)	2.63	0.73	0.64
			3	4.98(0.50)	2.63	0.73	0.64
			4	5.01(0.50)	2.62	0.73	0.64
			Total	20.01(0.50)	6.77	0.81	-
		0.6	1	5.00(0.50)	2.63	0.73	0.76
			2	4.99(0.50)	2.63	0.73	0.76
			3	5.05(0.50)	2.63	0.73	0.76
			4	5.01(0.50)	2.63	0.73	0.76
			Total	20.06(0.50)	8.01	0.87	-
		0.9	1	5.01(0.50)	2.63	0.73	0.86
			2	5.00(0.50)	2.63	0.73	0.86
			3	5.00(0.50)	2.62	0.73	0.86
			4	5.07(0.51)	2.62	0.73	0.86
			Total	20.09(0.50)	9.08	0.91	-
	Low	0.3	1	5.03(0.50)	2.25	0.58	0.62
			2	5.00(0.50)	2.24	0.57	0.61
			3	5.04(0.50)	2.24	0.57	0.62
			4	5.00(0.50)	2.25	0.58	0.62
			Total	20.07(0.50)	5.57	0.70	-
		0.6	1	5.00(0.50)	2.25	0.58	0.71
			2	4.98(0.50)	2.25	0.58	0.71
			3	5.04(0.50)	2.25	0.58	0.71
			4	4.98(0.50)	2.25	0.58	0.71
			Total	20.00(0.50)	6.45	0.78	-
		0.9	1	5.01(0.50)	2.24	0.58	0.80
			2	5.00(0.50)	2.25	0.58	0.80
			3	5.01(0.50)	2.25	0.58	0.80
			4	4.98(0.50)	2.25	0.58	0.80
			Total	19.99(0.50)	7.22	0.83	-
I=20	High	0.3	1	10.02(0.50)	4.90	0.84	0.66
			2	10.07(0.50)	4.89	0.84	0.66
			3	9.99(0.50)	4.91	0.84	0.66
			4	10.00(0.50)	4.89	0.84	0.66
			Total	40.01(0.50)	12.97	0.90	-
		0.6	1	10.01(0.50)	4.90	0.84	0.79
			2	9.95(0.50)	4.88	0.84	0.79
			3	10.04(0.50)	4.90	0.84	0.79
			4	10.03(0.50)	4.91	0.84	0.79

			Total	40.11(0.50)	15.52	0.93	-
		0.9	1	9.9(0.50)	4.92	0.84	0.90
			2	10.03(0.50)	4.90	0.84	0.90
			3	10.00(0.50)	4.88	0.84	0.90
			4	10.02(0.50)	4.88	0.84	0.90
			Total	39.95(0.50)	17.74	0.95	-
	Low	0.3	1	9.91(0.50)	4.00	0.73	0.64
			2	9.93(0.50)	4.00	0.73	0.64
			3	9.95(0.50)	3.98	0.73	0.64
			4	9.99(0.50)	3.98	0.73	0.64
			Total	39.79(0.50)	10.28	0.83	-
		0.6	1	9.99(0.50)	4.00	0.73	0.76
			2	9.90(0.50)	3.99	0.73	0.76
			3	9.97(0.50)	3.99	0.73	0.76
			4	10.02(0.50)	4.00	0.73	0.76
			Total	39.87(0.50)	12.17	0.88	-
		0.9	1	10.01(0.50)	3.99	0.73	0.86
			2	10.03(0.50)	4.00	0.73	0.86
			3	9.98(0.50)	3.99	0.73	0.86
			4	9.92(0.50)	3.98	0.73	0.86
			Total	39.93(0.50)	13.78	0.91	-

Table A4. Summary Statistics of Subscale Scores from the unidimensional 2PL IRT Model for Achievement Tests

Subscale length	Subscale consistency	Between-subscales correlation	Subscale	Mean	SD	Empirical Reliability	Correlations with total scores
I=10	High	0.3	1	0.00	0.82	0.68	0.65
			2	0.00	0.82	0.67	0.62
			3	0.00	0.82	0.67	0.61
			4	0.00	0.83	0.69	0.65
			Total	0.00	0.89	0.79	-
		0.6	1	0.00	0.82	0.68	0.75
			2	0.00	0.83	0.68	0.74
			3	0.00	0.82	0.68	0.74
			4	0.00	0.83	0.69	0.75
			Total	0.00	0.82	0.85	-
		0.9	1	0.00	0.82	0.68	0.84
			2	0.00	0.83	0.68	0.85
			3	0.00	0.82	0.68	0.84
			4	0.00	0.82	0.68	0.84
			Total	0.00	0.94	0.88	-
	Low	0.3	1	0.00	0.74	0.55	0.60
			2	0.00	0.74	0.55	0.61
			3	0.00	0.74	0.55	0.61
			4	0.00	0.74	0.55	0.62
			Total	0.00	0.83	0.69	-
		0.6	1	0.00	0.74	0.55	0.70
			2	0.00	0.74	0.55	0.70
			3	0.00	0.74	0.55	0.71
			4	0.00	0.74	0.55	0.70
			Total	0.00	0.88	0.77	-
		0.9	1	0.00	0.74	0.55	0.79
			2	0.00	0.74	0.55	0.79
			3	0.00	0.74	0.55	0.78
			4	0.00	0.74	0.55	0.79
			Total	0.00	0.90	0.81	-
I=20	High	0.3	1	0.00	0.90	0.80	0.64
			2	0.00	0.90	0.81	0.66
			3	0.00	0.90	0.80	0.66
			4	0.00	0.90	0.80	0.64
			Total	0.00	0.94	0.88	-
		0.6	1	0.00	0.90	0.80	0.78
			2	0.00	0.90	0.80	0.78
			3	0.00	0.89	0.80	0.77

		4	0.00	0.90	0.81	0.79
		Total	0.00	0.96	0.92	-
	0.9	1	0.00	0.90	0.80	0.89
		2	0.00	0.90	0.80	0.89
		3	0.00	0.90	0.80	0.90
		4	0.00	0.90	0.80	0.89
		Total	0.00	0.97	0.93	-
Low	0.3	1	0.00	0.84	0.71	0.62
		2	0.00	0.84	0.71	0.64
		3	0.00	0.84	0.71	0.65
		4	0.00	0.84	0.71	0.64
		Total	0.00	0.90	0.82	-
	0.6	1	0.00	0.85	0.72	0.77
		2	0.00	0.84	0.71	0.75
		3	0.00	0.84	0.71	0.75
		4	0.00	0.84	0.71	0.75
		Total	0.00	0.93	0.87	-
	0.9	1	0.00	0.84	0.71	0.85
		2	0.00	0.84	0.71	0.86
		3	0.00	0.84	0.71	0.85
		4	0.00	0.84	0.71	0.85
		Total	0.00	0.95	0.90	-

Table A5. Summary Statistics of Subscale Scores from the unidimensional 2PL IRT Model for Ability Tests

Subscale length	Subscale consistency	Between-subscale correlation	Subscale	Mean	SD	Empirical Reliability	Correlations with total scores
I=10	High	0.3	1	0.00	0.85	0.72	0.63
			2	0.00	0.85	0.72	0.65
			3	0.00	0.85	0.72	0.63
			4	0.00	0.85	0.72	0.65
			Total	0.00	0.90	0.81	-
		0.6	1	0.00	0.85	0.72	0.76
			2	0.00	0.85	0.72	0.75
			3	0.00	0.85	0.72	0.76
			4	0.00	0.85	0.72	0.76
			Total	0.00	0.93	0.87	-
		0.9	1	0.00	0.86	0.74	0.86
			2	0.00	0.85	0.72	0.86
			3	0.00	0.85	0.71	0.86
			4	0.00	0.85	0.71	0.86
			Total	0.00	0.95	0.90	-
	Low	0.3	1	0.00	0.76	0.57	0.62
			2	0.00	0.75	0.57	0.60
			3	0.00	0.76	0.58	0.63
			4	0.00	0.76	0.57	0.60
			Total	0.00	0.84	0.70	-
		0.6	1	0.00	0.76	0.58	0.71
			2	0.00	0.75	0.56	0.70
			3	0.00	0.77	0.59	0.72
			4	0.00	0.76	0.58	0.72
			Total	0.00	0.89	0.78	-
		0.9	1	0.00	0.76	0.58	0.80
			2	0.00	0.76	0.58	0.80
			3	0.00	0.76	0.58	0.80
			4	0.00	0.77	0.59	0.81
			Total	0.00	0.91	0.83	-
I=20	High	0.3	1	0.00	0.91	0.83	0.66
			2	0.00	0.91	0.83	0.66
			3	0.00	0.91	0.83	0.66
			4	0.00	0.91	0.83	0.65
			Total	0.00	0.95	0.90	-
		0.6	1	0.00	0.91	0.83	0.79
			2	0.00	0.91	0.83	0.79
			3	0.00	0.91	0.83	0.79

		4	0.00	0.92	0.84	0.80
		Total	0.00	0.97	0.93	-
	0.9	1	0.00	0.91	0.84	0.90
		2	0.00	0.91	0.83	0.90
		3	0.00	0.92	0.84	0.91
		4	0.00	0.91	0.83	0.90
		Total	0.00	0.97	0.95	-
Low	0.3	1	0.00	0.86	0.74	0.67
		2	0.00	0.85	0.73	0.63
		3	0.00	0.85	0.73	0.64
		4	0.00	0.86	0.73	0.64
		Total	0.00	0.91	0.83	-
	0.6	1	0.00	0.86	0.74	0.77
		2	0.00	0.85	0.73	0.76
		3	0.00	0.85	0.73	0.76
		4	0.00	0.85	0.73	0.76
		Total	0.00	0.94	0.88	-
	0.9	1	0.00	0.86	0.73	0.86
		2	0.00	0.86	0.73	0.86
		3	0.00	0.86	0.73	0.86
		4	0.00	0.86	0.74	0.86
		Total	0.00	0.95	0.91	-

Table A6. Correlations among Raw Subscale Scores from Achievement Tests

Data condition			Subscale				
Subscale length	Subscale consistency	Between-subscales correlation	Subscale	1	2	3	4
I=10	High	0.3	1	1.00	0.20	0.20	0.21
			2		1.00	0.20	0.20
			3			1.00	0.21
			4				1.00
		0.6	1	1.00	0.41	0.41	0.42
			2		1.00	0.41	0.42
			3			1.00	0.41
			4				1.00
		0.9	1	1.00	0.64	0.63	0.63
			2		1.00	0.63	0.64
			3			1.00	0.63
			4				1.00
	Low	0.3	1	1.00	0.17	0.17	0.16
			2		1.00	0.17	0.17
			3			1.00	0.16
			4				1.00
		0.6	1	1.00	0.33	0.33	0.33
			2		1.00	0.33	0.33
			3			1.00	0.33
			4				1.00
		0.9	1	1.00	0.50	0.50	0.50
			2		1.00	0.50	0.50
			3			1.00	0.50
			4				1.00
I=20	High	0.3	1	1.00	0.24	0.24	0.24
			2		1.00	0.24	0.24
			3			1.00	0.24
			4				1.00
		0.6	1	1.00	0.48	0.49	0.49
			2		1.00	0.49	0.49
			3			1.00	0.49
			4				1.00
		0.9	1	1.00	0.74	0.74	0.74
			2		1.00	0.74	0.74
			3			1.00	0.74
			4				1.00
	Low	0.3	1	1.00	0.21	0.21	0.21
			2		1.00	0.21	0.21
			3			1.00	0.21

		4				1.00
	0.6	1	1.00	0.43	0.43	0.43
		2		1.00	0.43	0.43
		3			1.00	0.43
		4				1.00
	0.9	1	1.00	0.64	0.64	0.64
		2		1.00	0.64	0.64
		3			1.00	0.64
		4				1.00

Table A7. Correlations among Raw Subscale Scores from Ability Tests

Data condition			Subscale				
Subscale length	Subscale consistency	Between-subscales Correlation	Subscale	1	2	3	4
I=10	High	0.3	1	1.00	0.21	0.21	0.22
			2		1.00	0.22	0.22
			3			1.00	0.21
			4				1.00
		0.6	1	1.00	0.43	0.43	0.43
			2		1.00	0.43	0.43
			3			1.00	0.43
			4				1.00
		0.9	1	1.00	0.66	0.66	0.66
			2		1.00	0.66	0.65
			3			1.00	0.65
			4				1.00
	Low	0.3	1	1.00	0.17	0.17	0.17
			2		1.00	0.17	0.17
			3			1.00	0.17
			4				1.00
		0.6	1	1.00	0.35	0.35	0.34
			2		1.00	0.35	0.34
			3			1.00	0.34
			4				1.00
		0.9	1	1.00	0.52	0.52	0.52
			2		1.00	0.52	0.52
			3			1.00	0.52
			4				1.00
I=20	High	0.3	1	1.00	0.25	0.25	0.25
			2		1.00	0.25	0.25
			3			1.00	0.25
			4				1.00
		0.6	1	1.00	0.50	0.50	0.50
			2		1.00	0.50	0.50
			3			1.00	0.50
			4				1.00
		0.9	1	1.00	0.76	0.76	0.76
			2		1.00	0.76	0.76
			3			1.00	0.76
			4				1.00
	Low	0.3	1	1.00	0.22	0.22	0.22
			2		1.00	0.22	0.22
			3			1.00	0.22

		4			1.00
	0.6	1	1.00	0.43	0.44	0.44
		2		1.00	0.44	0.44
		3			1.00	0.44
		4				1.00
	0.9	1	1.00	0.66	0.66	0.66
		2		1.00	0.66	0.66
		3			1.00	0.66
		4				1.00

Table A8. Correlations among Subscale θs for Achievement Tests

Data condition			Subscale				
Subscale length	Subscale consistency	Between-subscales correlation	Subscale	1	2	3	4
I=10	High	0.3	1	1.00	0.20	0.20	0.21
			2		1.00	0.19	0.20
			3			1.00	0.20
			4				1.00
		0.6	1	1.00	0.41	0.41	0.40
			2		1.00	0.40	0.40
			3			1.00	0.40
			4				1.00
		0.9	1	1.00	0.61	0.61	0.61
			2		1.00	0.61	0.61
			3			1.00	0.61
			4				1.00
	Low	0.3	1	1.00	0.17	0.17	0.17
			2		1.00	0.16	0.17
			3			1.00	0.17
			4				1.00
		0.6	1	1.00	0.33	0.33	0.33
			2		1.00	0.33	0.33
			3			1.00	0.33
			4				1.00
		0.9	1	1.00	0.50	0.50	0.51
			2		1.00	0.49	0.51
			3			1.00	0.50
			4				1.00
I=20	High	0.3	1	1.00	0.24	0.24	0.24
			2		1.00	0.24	0.24
			3			1.00	0.24
			4				1.00
		0.6	1	1.00	0.48	0.49	0.49
			2		1.00	0.48	0.49
			3			1.00	0.48
			4				1.00
		0.9	1	1.00	0.73	0.74	0.73
			2		1.00	0.74	0.73
			3			1.00	0.74
			4				1.00
	Low	0.3	1	1.00	0.22	0.22	0.22
			2		1.00	0.22	0.22
			3			1.00	0.22

		4				1.00
	0.6	1	1.00	0.43	0.44	0.43
		2		1.00	0.43	0.43
		3			1.00	0.43
		4				1.00
	0.9	1	1.00	0.64	0.64	0.64
		2		1.00	0.64	0.64
		3			1.00	0.64
		4				1.00

Table A9. Correlations among Subscale θs for Ability Tests

Data condition				Subscale			
Subscale length	Subscale consistency	Between-subscales correlation	Subscale	1	2	3	4
I=10	High	0.3	1	1.00	0.21	0.21	0.22
			2		1.00	0.22	0.23
			3			1.00	0.21
			4				1.00
		0.6	1	1.00	0.43	0.44	0.43
			2		1.00	0.43	0.43
			3			1.00	0.43
			4				1.00
		0.9	1	1.00	0.66	0.65	0.65
			2		1.00	0.65	0.65
			3			1.00	0.65
			4				1.00
	Low	0.3	1	1.00	0.17	0.17	0.17
			2		1.00	0.17	0.16
			3			1.00	0.17
			4				1.00
		0.6	1	1.00	0.34	0.35	0.35
			2		1.00	0.35	0.34
			3			1.00	0.34
			4				1.00
		0.9	1	1.00	0.52	0.52	0.52
			2		1.00	0.52	0.52
			3			1.00	0.52
			4				1.00
I=20	High	0.3	1	1.00	0.25	0.24	0.25
			2		1.00	0.24	0.23
			3			1.00	0.24
			4				1.00
		0.6	1	1.00	0.50	0.51	0.51
			2		1.00	0.50	0.51
			3			1.00	0.50
			4				1.00
		0.9	1	1.00	0.75	0.76	0.75
			2		1.00	0.76	0.75
			3			1.00	0.75
			4				1.00
	Low	0.3	1	1.00	0.22	0.23	0.22
			2		1.00	0.22	0.22
			3			1.00	0.21
			4				1.00

	0.6	1	1.00	0.44	0.44	0.44
		2		1.00	0.44	0.44
		3			1.00	0.44
		4				1.00
	0.9	1	1.00	0.66	0.65	0.65
		2		1.00	0.65	0.66
		3			1.00	0.66
		4				1.00

Table A10. Means and Standard Deviations for Estimated CTT Subscale Scores Averaged over 100 Replicated Achievement Data

Data condition			Subscale	Type of Predictor							
				No predictors		Observed subscale score		Total score		Subscale score & Total score	
Subscale length	Subscale consistency	Between-subscales correlation		Mean	SD	Mean	SD	Mean	SD	Mean	SD
I=10	High	0.3	1	6.91	2.39	6.91	1.71	6.91	1.27	6.91	1.73
			2	6.94	2.38	6.94	1.70	6.94	1.26	6.94	1.71
			3	6.96	2.39	6.96	1.71	6.96	1.28	6.96	1.72
			4	6.87	2.41	6.87	1.73	6.87	1.30	6.87	1.75
			Total	27.68	6.05	27.68	4.32	27.68	4.97	27.68	4.97
		0.6	1	6.99	2.39	6.94	1.70	6.94	1.57	6.94	1.76
			2	6.90	2.38	6.90	1.70	6.90	1.57	6.90	1.75
			3	6.94	2.39	6.94	1.71	6.94	1.58	6.94	1.76
			4	6.88	2.41	6.88	1.72	6.88	1.59	6.88	1.78
			Total	27.66	7.13	27.66	5.07	27.66	6.21	27.66	6.21
		0.9	1	6.92	2.40	6.92	1.72	6.92	1.85	6.92	1.87
			2	6.87	2.40	6.87	1.72	6.87	1.86	6.87	1.87
			3	6.91	2.40	6.91	1.72	6.91	1.85	6.91	1.87
			4	6.87	2.41	6.87	1.72	6.87	1.86	6.87	1.87
			Total	27.57	8.13	27.57	5.79	27.57	7.32	27.57	7.32
	Low	0.3	1	6.45	2.13	6.45	1.21	6.45	0.96	6.45	1.24
			2	6.50	2.12	6.50	1.20	6.50	0.94	6.50	1.23
			3	6.52	2.12	6.52	1.21	6.52	0.94	6.52	1.23
			4	6.53	2.13	6.53	1.21	6.53	0.95	6.53	1.24
			Total	26.00	5.19	26.00	2.94	26.00	3.69	26.00	3.69
		0.6	1	6.50	2.13	6.50	1.22	6.50	1.20	6.50	1.30
			2	6.47	2.14	6.47	1.22	6.47	1.20	6.47	1.31
			3	6.47	2.13	6.47	1.21	6.47	1.19	6.47	1.29
			4	6.50	2.12	6.50	1.20	6.50	1.19	6.50	1.29
			Total	25.95	6.00	25.94	3.39	25.95	4.69	25.95	4.69

0.9	1	6.50	2.13	6.50	1.21	6.50	1.40	6.50	1.41
	2	6.47	2.13	6.47	1.21	6.47	1.41	6.47	1.41
	3	6.50	2.13	6.50	1.21	6.50	1.41	6.50	1.41
	4	6.46	2.14	6.46	1.22	6.46	1.41	6.46	1.42
	Total	25.93	6.73	25.93	3.80	25.93	5.56	25.93	1.40
0.3	1	13.81	4.41	13.81	3.66	13.81	2.62	13.81	3.67
	2	13.80	4.41	13.80	3.66	13.90	2.62	13.80	3.67
	3	13.83	4.41	13.83	3.67	13.83	2.63	13.83	3.68
	4	13.86	4.40	13.86	3.65	13.86	2.62	13.86	3.66
	Total	55.29	11.51	55.29	9.56	55.29	10.37	55.29	10.37
0.6	1	13.84	4.14	13.84	3.67	13.84	3.23	13.84	3.72
	2	13.80	4.40	13.80	3.66	13.80	3.23	13.80	3.70
	3	13.76	4.42	13.76	3.67	13.76	3.24	13.76	3.72
	4	13.77	4.42	13.77	3.68	13.77	3.25	13.77	3.72
	Total	55.18	13.80	55.18	11.47	55.18	12.85	55.18	12.85
0.9	1	13.77	4.43	13.77	3.68	13.77	3.78	13.77	3.83
	2	13.81	4.43	13.81	3.69	13.81	3.78	13.81	3.83
	3	13.86	4.40	13.86	3.66	13.86	3.75	13.86	3.81
	4	13.77	4.43	13.77	3.68	13.77	3.78	13.77	3.83
	Total	55.21	15.83	55.21	13.15	55.21	15.00	55.21	15.00
0.3	1	12.96	3.75	12.96	2.71	12.96	2.00	12.96	2.73
	2	12.94	3.76	12.94	2.72	12.94	2.02	12.94	2.74
	3	13.03	3.74	13.03	2.70	13.03	2.00	13.03	2.72
	4	12.92	3.76	12.92	2.72	12.92	2.02	12.92	2.74
	Total	51.85	9.56	51.85	6.89	51.85	7.92	51.85	7.92
0.6	1	12.93	3.76	12.93	2.72	12.93	2.51	12.93	2.80
	2	13.03	3.74	13.03	2.69	13.03	2.49	13.03	2.78
	3	13.02	3.76	13.02	2.72	13.02	2.51	13.02	2.80
	4	12.98	3.75	12.98	2.71	12.98	2.50	12.98	2.79
	Total	51.95	11.31	51.95	8.15	51.95	9.93	51.95	9.93

0.9	1	13.01	3.74	13.01	2.70	13.01	2.90	13.01	2.93
	2	12.91	3.76	12.91	2.72	12.91	2.92	12.91	2.95
	3	12.92	3.76	12.92	2.71	12.92	2.91	12.92	2.94
	4	12.93	3.75	12.93	2.70	12.93	2.91	12.93	2.93
	Total	51.77	12.81	51.77	9.21	51.77	11.58	51.77	11.58

Table A11. Means and Standard Deviations of Estimated CTT Subscale Scores Averaged over 100 Replicated Ability Data

Data condition			Subscale	Type of Predictor							
				No predictors		Observed Subscale score		Total Score		Subscale score & Total Score	
Subscale length	Subscale consistency	Between-subscales correlation		Mean	SD	Mean	SD	Mean	SD	Mean	SD
I=10	High	0.3	1	5.02	2.64	5.02	1.93	5.02	1.44	5.02	1.95
			2	5.04	2.65	5.04	1.95	5.04	1.45	5.04	1.96
			3	4.98	2.66	4.98	1.96	4.98	1.45	4.98	1.97
			4	5.01	2.65	5.01	1.95	5.01	1.46	5.01	1.97
			Total	20.01	6.77	20.06	4.96	20.06	5.67	20.06	5.67
		0.6	1	5.00	2.66	5.00	1.95	5.00	1.80	5.00	2.01
			2	4.99	2.66	4.99	1.95	4.99	1.79	4.99	2.01
			3	5.05	2.66	5.05	1.95	5.05	1.80	5.05	2.01
			4	5.01	2.65	5.01	1.95	5.01	1.79	5.01	2.00
			Total	20.06	8.01	20.06	5.87	20.06	7.07	20.06	7.07
		0.9	1	5.01	2.65	5.01	1.95	5.01	2.09	5.01	2.11
			2	5.00	2.66	5.00	1.95	5.00	2.09	5.00	2.11
			3	5.00	2.65	5.00	1.95	5.00	2.09	5.00	2.11
			4	5.07	2.65	5.07	1.94	5.07	2.08	5.07	2.10
			Total	20.09	9.08	20.09	6.64	20.09	8.26	20.09	8.26
	Low	0.3	1	5.03	2.27	5.03	1.33	5.03	1.04	5.03	1.36
			2	5.00	2.26	5.00	1.32	5.00	1.04	5.00	1.35
			3	5.04	2.26	5.04	1.32	5.04	1.04	5.04	1.35
			4	5.00	2.26	5.00	1.32	5.00	1.03	5.00	1.34
			Total	20.07	5.57	20.07	3.24	20.07	4.04	20.07	4.04
		0.6	1	5.00	2.27	5.00	1.34	5.00	1.31	5.00	1.43
			2	4.98	2.26	4.98	1.32	4.98	1.29	4.98	1.41
			3	5.04	2.27	5.04	1.33	5.04	1.31	5.04	1.42
			4	4.98	2.27	4.98	1.33	4.98	1.30	4.98	1.42
			Total	20.00	6.45	20.00	3.78	20.00	5.13	20.00	5.14

0.9	1	5.01	2.26	5.01	1.32	5.01	1.52	5.01	1.53
	2	5.00	2.27	5.00	1.34	5.00	1.54	5.00	1.55
	3	5.01	2.27	5.01	1.33	5.01	1.53	5.01	1.54
	4	4.98	2.27	4.98	1.33	4.98	1.53	4.98	1.54
	Total	19.99	7.22	19.99	4.20	19.99	6.04	19.99	6.04
0.3	1	10.02	4.92	10.02	4.16	10.02	2.99	10.02	4.17
	2	10.07	4.91	10.07	4.15	10.07	2.97	10.07	4.16
	3	9.99	4.93	9.99	4.16	9.99	2.99	9.99	4.17
	4	10.00	4.93	10.00	4.17	10.00	3.01	10.00	4.18
	Total	40.09	12.97	40.09	10.96	40.09	11.82	40.09	11.82
0.6	1	10.10	4.92	10.10	4.17	10.10	3.67	10.10	4.21
	2	9.95	4.91	9.95	4.15	9.95	3.65	9.95	4.19
	3	10.04	4.92	10.04	4.16	10.04	3.67	10.04	4.21
	4	10.03	4.94	10.03	4.18	10.03	3.69	10.03	4.23
	Total	40.11	15.52	40.11	13.12	40.11	14.56	40.11	14.56
0.9	1	9.90	4.95	9.90	4.19	9.90	4.27	9.90	4.34
	2	10.03	4.92	10.03	4.16	10.03	4.25	10.03	4.31
	3	10.00	4.91	10.00	4.15	10.00	4.24	10.00	4.30
	4	10.02	4.91	10.02	4.14	10.02	4.23	10.02	4.29
	Total	39.95	17.74	39.95	14.97	39.95	16.89	39.95	16.89
0.3	1	9.91	4.02	9.91	2.95	9.91	2.19	9.91	2.98
	2	9.93	4.01	9.93	2.95	9.93	2.19	9.93	2.97
	3	9.95	4.00	9.95	2.93	9.95	2.17	9.95	2.96
	4	9.99	4.00	9.99	2.93	9.99	2.17	9.99	2.95
	Total	39.79	10.28	39.79	7.54	39.79	8.62	39.79	8.62
0.6	1	9.99	4.02	9.99	2.96	9.99	2.72	9.99	3.05
	2	9.90	4.01	9.90	2.95	9.90	2.71	9.90	3.03
	3	9.97	4.01	9.97	2.94	9.97	2.71	9.97	3.03
	4	10.02	4.02	10.02	2.97	10.02	2.73	10.02	3.05
	Total	39.88	12.17	39.88	8.93	39.88	10.77	39.88	10.77

0.9	1	10.01	4.01	10.01	2.95	10.01	3.16	10.01	3.19
	2	10.03	4.02	10.03	2.96	10.03	3.17	10.03	3.20
	3	9.98	4.01	9.98	2.94	9.98	3.15	9.98	3.18
	4	9.92	4.00	9.92	2.94	9.92	3.15	9.92	3.18
	Total	39.93	13.78	39.93	10.09	39.93	12.54	39.93	12.54

Table A12. PRMSE-MBs of CTT Subscale Scores for Achievement Tests

Subscale length	Subscale consistency	Between-subscales correlation	Subscale	$PRMSE - MB_{Kelley}$ $\sigma(R(\tau_s \mid S_s)$	$PRMSE - MB_{HH}$ $\sigma(R(\tau_s \mid S_T)$	$PRMSE - MB_{Haberman}$ $\sigma(R(\tau_s \mid S_s, S_T)$
I=10	High	0.3	1	0.71	0.38	0.72
			2	0.71	0.38	0.72
			3	0.71	0.38	0.72
			4	0.71	0.39	0.73
		0.6	1	0.71	0.60	0.76
			2	0.71	0.60	0.75
			3	0.71	0.60	0.75
			4	0.71	0.60	0.76
		0.9	1	0.71	0.83	0.85
			2	0.71	0.83	0.85
			3	0.71	0.83	0.84
			4	0.71	0.83	0.85
	Low	0.3	1	0.56	0.33	0.58
			2	0.56	0.33	0.58
			3	0.56	0.33	0.58
			4	0.56	0.33	0.58
		0.6	1	0.56	0.54	0.65
			2	0.56	0.54	0.65
			3	0.56	0.54	0.65
			4	0.56	0.54	0.64
		0.9	1	0.56	0.76	0.77
			2	0.56	0.76	0.77
			3	0.56	0.76	0.77
			4	0.56	0.76	0.77
I=20	High	0.3	1	0.83	0.42	0.83
			2	0.83	0.42	0.83
			3	0.83	0.42	0.83
			4	0.83	0.42	0.83
		0.6	1	0.83	0.64	0.85
			2	0.83	0.64	0.85
			3	0.83	0.64	0.85
			4	0.83	0.64	0.85
		0.9	1	0.83	0.87	0.90
			2	0.83	0.87	0.90
			3	0.83	0.87	0.90
			4	0.83	0.87	0.90
	Low	0.3	1	0.72	0.39	0.73
			2	0.72	0.39	0.73
			3	0.72	0.39	0.73

		4	0.72	0.39	0.73
	0.6	1	0.72	0.61	0.77
		2	0.72	0.61	0.76
		3	0.72	0.61	0.77
		4	0.72	0.61	0.76
	0.9	1	0.72	0.84	0.85
		2	0.72	0.83	0.85
		3	0.72	0.83	0.85
		4	0.72	0.83	0.85

Table A13. PRMSE-MBs of CTT Subscale Scores for Ability Tests

Subscale length	Subscale consistency	Between-subscales correlation	Subscale	$PRMSE - MB_{Kelley}$ $\sigma(R(\tau_s \mid S_s)$	$PRMSE - MB_{HH}$ $\sigma(R(\tau_s \mid S_T)$	$PRMSE - MB_{Haberman}$ $\sigma(R(\tau_s \mid S_s, S_T)$
I=10	High	0.3	1	0.73	0.39	0.74
			2	0.73	0.39	0.74
			3	0.73	0.39	0.74
			4	0.73	0.39	0.74
		0.6	1	0.73	0.61	0.77
			2	0.73	0.61	0.77
			3	0.73	0.61	0.77
			4	0.73	0.61	0.77
		0.9	1	0.73	0.84	0.86
			2	0.73	0.84	0.86
			3	0.73	0.84	0.86
			4	0.73	0.84	0.86
	Low	0.3	1	0.58	0.35	0.60
			2	0.57	0.34	0.60
			3	0.57	0.34	0.60
			4	0.58	0.34	0.60
		0.6	1	0.58	0.55	0.66
			2	0.58	0.55	0.66
			3	0.58	0.55	0.66
			4	0.58	0.55	0.66
		0.9	1	0.58	0.77	0.78
			2	0.58	0.77	0.78
			3	0.58	0.77	0.78
			4	0.58	0.77	0.78
I=20	High	0.3	1	0.84	0.43	0.85
			2	0.84	0.43	0.85
			3	0.84	0.43	0.85
			4	0.84	0.43	0.85
		0.6	1	0.84	0.65	0.86
			2	0.84	0.65	0.86
			3	0.84	0.65	0.86
			4	0.84	0.65	0.86
		0.9	1	0.84	0.88	0.91
			2	0.84	0.88	0.91
			3	0.84	0.88	0.91
			4	0.84	0.88	0.91
	Low	0.3	1	0.73	0.40	0.74
			2	0.73	0.40	0.74
			3	0.73	0.39	0.74
			4	0.73	0.39	0.74

	0.6	1	0.73	0.62	0.78
		2	0.73	0.62	0.77
		3	0.73	0.62	0.77
		4	0.73	0.62	0.78
	0.9	1	0.73	0.84	0.86
		2	0.73	0.84	0.86
		3	0.73	0.84	0.86
		4	0.73	0.84	0.86

Table A14. Adjusted OPI Mean P-value and Variance in Achievement Test

Data condition			Subscale	OPI Mean *p* value	OPI score variance
Subscale length	Subscale consistency	Between-subscales correlation			
I=10	High	0.3	1	0.72	0.20
			2	0.71	0.19
			3	0.74	0.19
			4	0.73	0.20
		0.6	1	0.73	0.21
			2	0.74	0.20
			3	0.73	0.21
			4	0.70	0.22
		0.9	1	0.73	0.23
			2	0.73	0.23
			3	0.73	0.23
			4	0.71	0.23
	Low	0.3	1	0.69	0.15
			2	0.69	0.15
			3	0.68	0.15
			4	0.68	0.15
		0.6	1	0.68	0.17
			2	0.67	0.17
			3	0.68	0.17
			4	0.68	0.17
		0.9	1	0.69	0.19
			2	0.68	0.19
			3	0.69	0.19
			4	0.68	0.19
I=20	High	0.3	1	0.73	0.17
			2	0.72	0.18
			3	0.73	0.18
			4	0.71	0.17
		0.6	1	0.72	0.20
			2	0.73	0.21
			3	0.73	0.20
			4	0.71	0.21
		0.9	1	0.73	0.23
			2	0.72	0.23
			3	0.71	0.24
			4	0.73	0.23
	Low	0.3	1	0.67	0.14
			2	0.68	0.14
			3	0.69	0.14

		4	0.67	0.14
	0.6	1	0.67	0.17
		2	0.67	0.17
		3	0.68	0.17
		4	0.68	0.17
	0.9	1	0.68	0.19
		2	0.67	0.20
		3	0.68	0.19
		4	0.67	0.20

Table A15. Adjusted OPI Mean P-value and Variance in Ability Test

Data condition			Subscale	OPI Mean *p* value	OPI score variance
Subscale length	Subscale consistency	Between-subscales correlation			
I=10	High	0.3	1	0.50	0.22
			2	0.51	0.22
			3	0.51	0.22
			4	0.51	0.22
		0.6	1	0.52	0.24
			2	0.49	0.24
			3	0.49	0.25
			4	0.49	0.24
		0.9	1	0.52	0.27
			2	0.50	0.27
			3	0.51	0.26
			4	0.51	0.26
	Low	0.3	1	0.51	0.17
			2	0.51	0.16
			3	0.51	0.17
			4	0.49	0.16
		0.6	1	0.51	0.19
			2	0.49	0.18
			3	0.51	0.19
			4	0.49	0.19
		0.9	1	0.51	0.21
			2	0.48	0.21
			3	0.51	0.21
			4	0.49	0.21
I=20	High	0.3	1	0.49	0.20
			2	0.51	0.20
			3	0.52	0.21
			4	0.48	0.20
		0.6	1	0.49	0.24
			2	0.50	0.24
			3	0.48	0.24
			4	0.49	0.25
		0.9	1	0.49	0.27
			2	0.49	0.27
			3	0.51	0.27
			4	0.51	0.27
	Low	0.3	1	0.49	0.16
			2	0.48	0.15
			3	0.50	0.16

		4	0.49	0.16
	0.6	1	0.51	0.19
		2	0.47	0.19
		3	0.48	0.19
		4	0.51	0.19
	0.9	1	0.50	0.21
		2	0.49	0.22
		3	0.48	0.22
		4	0.48	0.22

Table A16. Overall Goodness-of-Fit Comparison among IRT Models over 100 Replications in Achievement Tests

Data condition			Models (# of par)	AIC	-2lnL	Δ-2lnL	Comparing models
Subscale length	Subscale consistency	Between-subscales correlation					
I=10	High	0.3	1PL (41)	136,618.6	136,536.6		
			2PL (80)	136,618.3	136,458.3	78.3*	2PLvs.1PL
			3PL (120)	136,649.9	136,409.9	48.4	3PLvs.2PL
			MIRT-2PL (86)	132,087.8	131,915.8	4,542.5*	2Plvs.MIRT
		0.6	1PL (41)	131,831.6	131,749.6		
			2PL (80)	130,771.4	130,611.4	1,138.2*	2PLvs.1PL
			3PL (120)	130,747.5	130,607.5	3.9	3PLvs.2PL
			MIRT-2PL (86)	130,355.9	130,183.9	427.5*	2Plvs.MIRT
		0.9	1PL (41)	127,032.0	126,950.0		
			2PL (80)	126,984.0	126,824.0	126.0*	2PLvs.1PL
			3PL (120)	127,040.7	126,800.7	23.3	3PLvs.2PL
			MIRT-2PL (86)	125,742.9	125,660.9	1,163.1*	2Plvs.MIRT
	Low	0.3	1PL (41)	149,073.9	148,991.9		
			2PL (80)	149,071.5	148,911.5	80.4*	2PLvs.1PL
			3PL (120)	149,142.8	148,902.8	8.7	3PLvs.2PL
			MIRT-2PL (86)	145,618.1	145,446.1	3,465.4*	2Plvs.MIRT
		0.6	1PL (41)	147,799.1	147,717.1		
			2PL (80)	147,777.2	147,617.2	99.9*	2PLvs.1PL
			3PL (120)	147,834.2	147,594.2	23.0	3PLvs.2PL
			MIRT-2PL (86)	147,104.4	146,932.4	684.8*	2Plvs.MIRT
		0.9	1PL (41)	144,094.9	144,012.9		
			2PL (80)	144,060.8	143,900.8	112.1*	2PLvs.1PL
			3PL (120)	144,115.0	143,875.0	25.8	3PLvs.2PL
			MIRT-2PL (86)	143,127.3	142,955.3	945.5*	2Plvs.MIRT

I=20	High	0.3	1PL (81)	270,490.4	270,328.4		
			2PL (160)	270,485.9	270,165.9	162.5*	2PLvs.1PL
			3PL (240)	270,597.0	270,117.0	48.9	3PLvs.2PL
			MIRT-2PL (166)	254,298.7	253,966.7	16,199.2*	2Plvs.MIRT
		0.6	1PL (81)	262,054.2	261,892.2		
			2PL (160)	262,014.7	261,694.7	197.5*	2PLvs.1PL
			3PL (240)	262,113.3	261,633.3	61.4	3PLvs.2PL
			MIRT-2PL (166)	254,856.3	254,524.3	7,170.4*	2Plvs.MIRT
		0.9	1PL (81)	248,924.9	248,762.9		
			2PL (160)	248,811.9	248,491.9	271.0*	2PLvs.1PL
			3PL (240)	248,914.1	248,434.1	57.8	3PLvs.2PL
			MIRT-2PL (166)	245,966.0	245,146.0	3,345.9*	2Plvs.MIRT
	Low	0.3	1PL (81)	296,796.1	296,634.1		
			2PL (160)	296,792.6	296,472.6	161.5*	2PLvs.1PL
			3PL (240)	296,911.8	296,431.8	40.8	3PLvs.2PL
			MIRT-2PL (166)	289,750.9	289,418.9	7,053.7*	2Plvs.MIRT
		0.6	1PL (81)	291,925.9	291,763.9		
			2PL (160)	291,885.7	291,565.7	198.2*	2PLvs.1PL
			3PL (240)	291,988.3	291,508.3	57.4	3PLvs.2PL
			MIRT-2PL (166)	288,141.0	287,875.4	3,690.3*	2Plvs.MIRT
		0.9	1PL (81)	285,995.9	285,833.9		
			2PL (160)	285,883.5	285,563.5	270.4*	2PLvs.1PL
			3PL (240)	285,992.0	285,512.0	51.5	3PLvs.2PL
			MIRT-2PL (166)	284,872.0	284,540.0	1,023.5*	2Plvs.MIRT

Table A17. Overall Goodness-of-Fit Comparison among IRT Models over 100 Replications in Ability Tests

Data condition			Models (# of par)	AIC	-2lnL	Δ-2lnL	Comparing models
Subscale length	Subscale consistency	Between-subscales correlation					
I=10	High	0.3	1PL (41)	153,305.8	153,223.8		
			2PL (80)	153,294.2	153,134.2	89.6*	2PLvs.1PL
			3PL (120)	153,293.9	153,083.9	50.3	3PLvs.2PL
			MIRT-2PL (86)	147,142.8	146,970.8	6,163.4*	2Plvs.MIRT
		0.6	1PL (41)	147,708.4	147,626.4		
			2PL (80)	147,699.4	147,539.4	87.0*	2PLvs.1PL
			3PL (120)	147,706.7	147,496.7	42.7	3PLvs.2PL
			MIRT-2PL (86)	145,469.9	145,297.9	2,241.5*	2Plvs.MIRT
		0.9	1PL (41)	141,768.5	141,686.5		
			2PL (80)	141,712.2	141,552.2	134.3*	2PLvs.1PL
			3PL (120)	141,768.1	141,528.1	24.1	3PLvs.2PL
			MIRT-2PL (86)	141,522.4	141,350.4	201.8*	2Plvs.MIRT
	Low	0.3	1PL (41)	159,875.3	159,793.3		
			2PL (80)	159,878.7	159,718.7	74.6*	2PLvs.1PL
			3PL (120)	159,956.0	159,716.7	2.0	3PLvs.2PL
			MIRT-2PL (86)	157,520.0	157,348.0	2,370.7*	2Plvs.MIRT
		0.6	1PL (41)	157,177.3	157,095.3		
			2PL (80)	157,155.0	156,995.0	100.3*	2PLvs.1PL
			3PL (120)	157,214.70	156,974.7	20.3	3PLvs.2PL
			MIRT-2PL (86)	156,347.6	156,175.6	819.4*	2Plvs.MIRT
		0.9	1PL (41)	153,928.0	153,846.0		
			2PL (80)	153,885.3	153,725.3	120.7*	2PLvs.1PL
			3PL (120)	153,940.90	153,700.9	24.4	3PLvs.2PL
			MIRT-2PL (86)	153,451.3	153,279.3	446.0*	2Plvs.MIRT

I=20	High	0.3	1PL (81)	303,343.6	303,181.4		
			2PL (160)	303,356.4	303,036.4	145.0*	2PLvs.1PL
			3PL (240)	303,466.60	302,986.6	49.8	3PLvs.2PL
			MIRT-2PL (166)	285,540.3	285,208.3	17,828.1*	2Plvs.MIRT
		0.6	1PL (81)	291,025.3	290,863.3		
			2PL (160)	290,959.7	290,639.7	223.6*	2PLvs.1PL
			3PL (240)	291,108.20	290,628.2	11.50	3PLvs.2PL
			MIRT-2PL (166)	282,334.1	282,002.1	8,637.6*	2Plvs.MIRT
		0.9	1PL (81)	278,475.3	278,313.3		
			2PL (160)	278,343.2	278,023.2	290.1*	2PLvs.1PL
			3PL (240)	278,486.0	278,006.2	17.0	3PLvs.2PL
			MIRT-2PL (166)	273,649	273,317.0	4,706.20*	2Plvs.MIRT
	Low	0.3	1PL (81)	315,576.0	316,414.0		
			2PL (160)	316,562.5	316,242.5	171.5*	2PLvs.1PL
			3PL (240)	316,720.60	316,240.6	1.90	3PLvs.2PL
			MIRT-2PL (166)	309,821.1	309,489.1	6,753.4*	2Plvs.MIRT
		0.6	1PL (81)	310,874.0	310,712.0		
			2PL (160)	310,812.8	310,492.8	219.2*	2PLvs.1PL
			3PL (240)	310,944.50	310,464.5	28.30	3PLvs.2PL
			MIRT-2PL (166)	307,816.3	307,484.3	3,008.5*	2Plvs.MIRT
		0.9	1PL (81)	305,182.7	305,020.7		
			2PL (160)	305,063.0	304,743.0	277.7*	2PLvs.1PL
			3PL (240)	305,204.60	304,724.6	18.40	3PLvs.2PL
			MIRT-2PL (166)	304,795.5	304,463.5	279.5*	2Plvs.MIRT

Table A18. Comparisons of Empirical Reliability from Both Unidimemsional and Multidimensional Subscale Scores in Achievement Tests

Subscale length	Subscale consistency	Between-subscales correlation	Subscale	Empirical reliability		
				Total test	U2PL Mean	M2PL Mean
I=10	High	0.3	1	0.79	0.68	0.70
			2		0.67	0.69
			3		0.67	0.69
			4		0.69	0.70
		0.6	1	0.85	0.68	0.74
			2		0.68	0.73
			3		0.68	0.74
			4		0.69	0.74
		0.9	1	0.88	0.68	0.87
			2		0.68	0.87
			3		0.68	0.87
			4		0.68	0.87
	Low	0.3	1	0.69	0.55	0.57
			2		0.55	0.58
			3		0.55	0.58
			4		0.55	0.58
		0.6	1	0.77	0.55	0.64
			2		0.55	0.64
			3		0.55	0.64
			4		0.55	0.64
		0.9	1	0.81	0.55	0.80
			2		0.55	0.80
			3		0.55	0.80
			4		0.55	0.81
I=20	High	0.3	1	0.88	0.80	0.81
			2		0.81	0.81
			3		0.80	0.81
			4		0.80	0.81
		0.6	1	0.92	0.80	0.83
			2		0.80	0.83
			3		0.80	0.83
			4		0.81	0.83
		0.9	1	0.93	0.80	0.94
			2		0.80	0.94
			3		0.80	0.94
			4		0.80	0.94
	Low	0.3	1	0.82	0.71	0.72
			2		0.71	0.72

		3		0.71	0.72
		4		0.71	0.72
	0.6	1	0.87	0.72	0.77
		2		0.71	0.76
		3		0.71	0.76
		4		0.71	0.76
	0.9	1	0.90	0.71	0.89
		2		0.71	0.89
		3		0.71	0.89
		4		0.71	0.90

Table A19. Comparisons of Empirical Reliability from Both Unidimemsional and Multidimensional Subscale Scores in Ability Tests

Subscale length	Subscale consistency	Between-subscales correlation	Subscale	Empirical reliability		
				Total test	U2PL Mean	M2PL Mean
I=10	High	0.3	1	0.81	0.72	0.73
			2		0.72	0.73
			3		0.72	0.73
			4		0.72	0.73
		0.6	1	0.87	0.72	0.76
			2		0.72	0.77
			3		0.72	0.77
			4		0.72	0.77
		0.9	1	0.90	0.74	0.89
			2		0.72	0.90
			3		0.71	0.89
			4		0.71	0.90
	Low	0.3	1	0.70	0.57	0.60
			2		0.57	0.59
			3		0.58	0.60
			4		0.57	0.59
		0.6	1	0.78	0.58	0.66
			2		0.56	0.65
			3		0.59	0.67
			4		0.58	0.66
		0.9	1	0.83	0.58	0.81
			2		0.58	0.80
			3		0.58	0.81
			4		0.59	0.81
I=20	High	0.3	1	0.90	0.83	0.75
			2		0.83	0.74
			3		0.83	0.74
			4		0.83	0.74
		0.6	1	0.93	0.83	0.78
			2		0.83	0.77
			3		0.83	0.78
			4		0.84	0.77
		0.9	1	0.95	0.84	0.90
			2		0.83	0.91
			3		0.84	0.90
			4		0.83	0.90
	Low	0.3	1	0.83	0.74	0.84
			2		0.73	0.84

		3		0.73	0.84
		4		0.73	0.84
	0.6	1	0.88	0.74	0.85
		2		0.73	0.86
		3		0.73	0.86
		4		0.73	0.86
	0.9	1	0.91	0.73	0.94
		2		0.73	0.94
		3		0.73	0.94
		4		0.74	0.94

Table A20. RMSE-MBs for CTT Subscale Scores over 100 Replicated Achievement Data

Data condition			Subscale	$\sigma(e_x)$	$RMSE - MB_{Kelley}$ $\sigma(R(\tau_s \mid S_s)$	$RMSE - MB_{HH}$ $\sigma(R(\tau_s \mid S_T)$	$RMSE - MB_{Haberman}$ $\sigma(R(\tau_s \mid S_s, S_T)$
I=10	High	0.3	1	1.28	1.08	1.57	1.06
			2	1.28	1.08	1.56	1.05
			3	1.28	1.08	1.56	1.05
			4	1.28	1.08	1.58	1.06
		0.6	1	1.28	1.08	1.26	0.99
			2	1.28	1.07	1.26	0.98
			3	1.28	1.07	1.26	0.98
			4	1.28	1.08	1.27	0.99
		0.9	1	1.28	1.08	0.83	0.79
			2	1.28	1.08	0.83	0.79
			3	1.28	1.08	0.83	0.79
			4	1.28	1.08	0.83	0.79
	Low	0.3	1	1.41	1.05	1.29	1.02
			2	1.40	1.05	1.29	1.02
			3	1.40	1.05	1.29	1.02
			4	1.40	1.05	1.29	1.02
		0.6	1	1.40	1.05	1.07	0.94
			2	1.40	1.05	1.08	0.94
			3	1.40	1.05	1.07	0.94
			4	1.40	1.05	1.07	0.94
		0.9	1	1.40	1.05	0.77	0.76
			2	1.40	1.05	0.76	0.75
			3	1.40	1.05	0.77	0.76
			4	1.40	1.05	0.77	0.76
I=20	High	0.3	1	1.81	1.65	3.04	1.63
			2	1.81	1.65	3.04	1.63
			3	1.81	1.65	3.04	1.63
			4	1.81	1.64	3.03	1.62
		0.6	1	1.81	1.65	2.40	1.55
			2	1.81	1.65	2.38	1.55
			3	1.81	1.65	2.40	1.55
			4	1.81	1.65	2.39	1.55
		0.9	1	1.81	1.65	1.43	1.27
			2	1.81	1.65	1.43	1.27
			3	1.81	1.65	1.42	1.26
			4	1.82	1.65	1.43	1.27
	Low	0.3	1	1.98	1.68	2.48	1.65
			2	1.98	1.68	2.48	1.65
			3	1.98	1.68	2.47	1.64
			4	1.99	1.68	2.48	1.65

	0.6	1	1.98	1.68	1.98	1.54
		2	1.98	1.67	1.96	1.53
		3	1.98	1.68	1.98	1.53
		4	1.98	1.68	1.98	1.54
	0.9	1	1.98	1.68	1.29	1.22
		2	1.98	1.68	1.29	1.23
		3	1.99	1.68	1.30	1.23
		4	1.99	1.68	1.29	1.22

Table A21. RMSE-MBs for CTT Subscale Scores over 100 Replicated Ability Data

Subscale length	Subscale consistency	Between-subscales correlation	Subscale	$\sigma(e_x)$	$RMSE - MB_{Kelley}$ $\sigma(R(\tau_s \mid S_s)$	$RMSE - MB_{HH}$ $\sigma(R(\tau_s \mid S_T)$	$RMSE - MB_{Haberman}$ $\sigma(R(\tau_s \mid S_s, S_T)$
I=10	High	0.3	1	1.37	1.17	1.74	1.14
			2	1.37	1.17	1.75	1.14
			3	1.37	1.17	1.75	1.14
			4	1.37	1.17	1.74	1.14
		0.6	1	1.37	1.17	1.40	1.07
			2	1.37	1.17	1.40	1.07
			3	1.37	1.17	1.40	1.07
			4	1.37	1.17	1.40	1.07
		0.9	1	1.37	1.17	0.90	0.85
			2	1.37	1.17	0.90	0.85
			3	1.36	1.16	0.90	0.85
			4	1.37	1.17	0.90	0.85
	Low	0.3	1	1.46	1.11	1.38	1.08
			2	1.46	1.11	1.38	1.08
			3	1.46	1.11	1.38	1.08
			4	1.46	1.11	1.38	1.08
		0.6	1	1.46	1.11	1.15	1.00
			2	1.46	1.11	1.14	1.00
			3	1.46	1.11	1.14	1.00
			4	1.46	1.11	1.15	1.00
		0.9	1	1.46	1.11	0.81	0.80
			2	1.46	1.11	0.82	0.80
			3	1.46	1.11	0.81	0.80
			4	1.46	1.11	0.82	0.80
I=20	High	0.3	1	1.93	1.78	3.40	1.75
			2	1.94	1.78	3.40	1.76
			3	1.94	1.78	3.41	1.76
			4	1.94	1.78	3.40	1.75
		0.6	1	1.93	1.78	2.66	1.67
			2	1.93	1.78	2.65	1.67
			3	1.93	1.78	2.66	1.68
			4	1.93	1.78	2.66	1.68
		0.9	1	1.94	1.78	1.58	1.38
			2	1.94	1.78	1.56	1.38
			3	1.93	1.78	1.56	1.37
			4	1.94	1.78	1.57	1.38
	Low	0.3	1	2.07	1.77	2.65	1.73
			2	2.07	1.77	2.65	1.73
			3	2.07	1.77	2.65	1.73

		4	2.07	1.77	2.65	1.73
	0.6	1	2.07	1.77	2.12	1.62
		2	2.06	1.77	2.12	1.62
		3	2.07	1.77	2.11	1.62
		4	2.07	1.77	2.12	1.62
	0.9	1	2.07	1.77	1.36	1.29
		2	2.07	1.77	1.37	1.29
		3	2.07	1.77	1.37	1.29
		4	2.07	1.77	1.36	1.29

Table A22. Expected True Subscale Scores in Achievement Tests

Subscale length	Subscale consistency	Between-subscales correlation	Subscale	Mean	SD
I=10	High	0.3	1	6.91	2.00
			2	6.94	1.99
			3	6.96	1.99
			4	6.87	2.02
		0.6	1	6.94	2.00
			2	6.91	1.99
			3	6.94	1.99
			4	6.88	2.02
		0.9	1	6.92	2.01
			2	6.88	2.01
			3	6.91	2.00
			4	6.87	2.01
	Low	0.3	1	6.46	1.58
			2	6.50	1.58
			3	6.52	1.58
			4	6.52	1.57
		0.6	1	6.50	1.58
			2	6.47	1.59
			3	6.47	1.59
			4	6.51	1.58
		0.9	1	6.50	1.59
			2	6.47	1.58
			3	6.50	1.58
			4	6.46	1.59
I=20	High	0.3	1	13.80	3.99
			2	13.80	4.00
			3	13.83	4.00
			4	13.86	3.98
		0.6	1	13.84	4.00
			2	13.80	4.00
			3	13.76	4.01
			4	13.77	4.01
		0.9	1	13.77	4.02
			2	13.81	4.01
			3	13.86	3.99
			4	13.77	4.02
	Low	0.3	1	12.96	3.16
			2	12.94	3.17
			3	13.03	3.16
			4	12.92	3.17

	0.6	1	12.93	3.18
		2	13.02	3.15
		3	13.03	3.16
		4	12.97	3.17
	0.9	1	13.01	3.16
		2	12.91	3.17
		3	12.93	3.17
		4	12.93	3.17

Table A23. Expected True Subscale Scores in Ability Tests

Subscale length	Subscale consistency	Between-subscales correlation	Subscale	Mean	SD
I=10	High	0.3	1	5.03	2.24
			2	5.05	2.25
			3	4.97	2.25
			4	5.01	2.24
		0.6	1	5.00	2.25
			2	4.99	2.25
			3	5.05	2.24
			4	5.01	2.25
		0.9	1	5.01	2.25
			2	5.01	2.25
			3	5.00	2.24
			4	5.07	2.24
	Low	0.3	1	5.03	1.71
			2	5.01	1.70
			3	5.04	1.71
			4	5.00	1.71
		0.6	1	5.00	1.71
			2	4.98	1.71
			3	5.04	1.72
			4	4.98	1.71
		0.9	1	5.01	1.71
			2	5.00	1.71
			3	5.01	1.71
			4	4.98	1.72
I=20	High	0.3	1	10.02	4.51
			2	10.08	4.50
			3	10.00	4.50
			4	10.00	4.50
		0.6	1	10.10	4.51
			2	9.94	4.48
			3	10.03	4.50
			4	10.02	4.52
		0.9	1	9.90	4.52
			2	10.03	4.50
			3	10.00	4.49
			4	10.02	4.49
	Low	0.3	1	9.92	3.42
			2	9.94	3.42
			3	9.96	3.41
			4	9.98	3.41

0.6	1	9.99	3.43
	2	9.89	3.41
	3	9.97	3.41
	4	10.01	3.43
0.9	1	10.01	3.42
	2	10.02	3.42
	3	9.99	3.41
	4	9.92	3.41

*Table A24. Correlations of True Subscale **θ**s and CTT Raw Subscale Scores with Expected True Subscale Scores in Achievement Tests*

Subscale length	Subscale consistency	Between-subscales correlation	Expected true subscale scores	True Subscale θs				CTT raw subscale scores			
				1	2	3	4	1	2	3	4
I=10	High	r = 0.3	1	0.97	0.29	0.29	0.29	0.84	0.24	0.24	0.24
			2	0.29	0.97	0.29	0.29	0.24	0.84	0.24	0.24
			3	0.29	0.29	0.97	0.29	0.24	0.24	0.84	0.24
			4	0.29	0.29	0.29	0.97	0.24	0.24	0.24	0.85
		r = 0.6	1	0.97	0.58	0.58	0.58	0.84	0.49	0.49	0.49
			2	0.58	0.97	0.58	0.58	0.49	0.84	0.49	0.49
			3	0.58	0.58	0.97	0.58	0.49	0.49	0.84	0.49
			4	0.58	0.58	0.58	0.97	0.49	0.49	0.49	0.84
		r = 0.9	1	0.97	0.87	0.87	0.87	0.84	0.75	0.75	0.75
			2	0.87	0.97	0.87	0.87	0.75	0.84	0.75	0.75
			3	0.87	0.87	0.97	0.87	0.75	0.75	0.84	0.75
			4	0.87	0.87	0.87	0.97	0.75	0.75	0.75	0.84
	Low	r = 0.3	1	0.97	0.29	0.29	0.29	0.91	0.26	0.26	0.26
			2	0.29	0.97	0.29	0.29	0.26	0.91	0.26	0.26
			3	0.29	0.29	0.97	0.29	0.26	0.26	0.91	0.26
			4	0.29	0.29	0.29	0.97	0.26	0.26	0.26	0.91
		r = 0.6	1	0.97	0.58	0.58	0.58	0.91	0.53	0.53	0.53
			2	0.58	0.97	0.58	0.58	0.53	0.91	0.53	0.54
			3	0.58	0.58	0.97	0.58	0.53	0.53	0.91	0.53
			4	0.58	0.59	0.58	0.97	0.53	0.53	0.53	0.91
		r = 0.9	1	0.97	0.88	0.87	0.88	0.91	0.81	0.82	0.82
			2	0.87	0.97	0.87	0.87	0.82	0.91	0.81	0.81
			3	0.87	0.87	0.97	0.87	0.82	0.81	0.91	0.81
			4	0.88	0.87	0.87	0.97	0.82	0.81	0.81	0.91

I=20	High	r = 0.3	1	0.99	0.30	0.30	0.30	0.75	0.22	0.22	0.22
			2	0.30	0.99	0.30	0.30	0.22	0.75	0.22	0.22
			3	0.30	0.30	0.99	0.29	0.22	0.22	0.75	0.22
			4	0.30	0.30	0.29	0.99	0.22	0.22	0.22	0.75
		r = 0.6	1	0.99	0.59	0.59	0.59	0.75	0.44	0.44	0.44
			2	0.59	0.99	0.59	0.59	0.44	0.75	0.44	0.44
			3	0.59	0.59	0.99	0.59	0.44	0.44	0.75	0.44
			4	0.59	0.59	0.59	0.99	0.44	0.44	0.44	0.75
		r = 0.9	1	0.99	0.89	0.89	0.89	0.75	0.67	0.67	0.67
			2	0.89	0.99	0.89	0.89	0.67	0.75	0.67	0.67
			3	0.89	0.89	0.99	0.89	0.67	0.67	0.75	0.67
			4	0.89	0.89	0.89	0.99	0.67	0.67	0.67	0.75
	Low	r = 0.3	1	0.99	0.29	0.29	0.29	0.85	0.25	0.25	0.25
			2	0.29	0.99	0.29	0.30	0.25	0.85	0.25	0.25
			3	0.29	0.29	0.99	0.30	0.25	0.25	0.85	0.25
			4	0.29	0.30	0.30	0.99	0.25	0.25	0.25	0.85
		r = 0.6	1	0.99	0.59	0.59	0.59	0.85	0.50	0.50	0.50
			2	0.59	0.99	0.60	0.60	0.50	0.85	0.50	0.51
			3	0.59	0.60	0.99	0.59	0.51	0.51	0.85	0.50
			4	0.59	0.60	0.59	0.99	0.50	0.51	0.50	0.85
		r = 0.9	1	0.99	0.89	0.89	0.89	0.85	0.76	0.76	0.76
			2	0.89	0.99	0.89	0.89	0.76	0.85	0.76	0.76
			3	0.89	0.89	0.99	0.89	0.76	0.76	0.85	0.76
			4	0.89	0.89	0.89	0.99	0.76	0.76	0.76	0.85

*Table A25. Correlations of True Subscale **θ**s and CTT Raw Subscale Scores with Expected True Subscale Scores in Ability Tests*

Subscale length	Subscale consistency	Between-subscales correlation	Expected true subscale scores	True Subscale θs				CTT raw subscale scores			
				1	2	3	4	1	2	3	4
I=10	High	r = 0.3	1	0.99	0.30	0.30	0.30	0.85	0.25	0.25	0.25
			2	0.30	0.99	0.30	0.30	0.25	0.85	0.25	0.25
			3	0.30	0.30	0.99	0.30	0.25	0.25	0.86	0.25
			4	0.30	0.30	0.30	0.99	0.25	0.25	0.25	0.85
		r = 0.6	1	0.99	0.59	0.59	0.59	0.86	0.51	0.51	0.51
			2	0.59	0.99	0.59	0.59	0.51	0.86	0.51	0.51
			3	0.59	0.59	0.99	0.59	0.51	0.50	0.85	0.50
			4	0.59	0.59	0.59	0.99	0.51	0.51	0.51	0.86
		r = 0.9	1	0.99	0.89	0.89	0.89	0.86	0.77	0.77	0.77
			2	0.89	0.99	0.89	0.89	0.77	0.86	0.77	0.76
			3	0.89	0.89	0.99	0.89	0.77	0.77	0.86	0.76
			4	0.89	0.89	0.89	0.99	0.77	0.77	0.77	0.85
	Low	r = 0.3	1	0.99	0.30	0.30	0.30	0.92	0.27	0.27	0.27
			2	0.29	0.99	0.29	0.30	0.27	0.92	0.27	0.27
			3	0.30	0.29	0.99	0.30	0.27	0.27	0.92	0.27
			4	0.30	0.30	0.30	0.99	0.27	0.27	0.27	0.92
		r = 0.6	1	0.99	0.59	0.59	0.59	0.92	0.54	0.54	0.54
			2	0.59	0.99	0.59	0.59	0.54	0.92	0.54	0.54
			3	0.59	0.59	0.99	0.59	0.54	0.54	0.92	0.54
			4	0.59	0.59	0.59	0.99	0.55	0.54	0.54	0.92
		r = 0.9	1	0.99	0.89	0.89	0.89	0.92	0.82	0.82	0.82
			2	0.89	0.99	0.89	0.89	0.82	0.92	0.82	0.82
			3	0.89	0.89	0.99	0.89	0.82	0.82	0.92	0.82
			4	0.89	0.89	0.89	0.99	0.82	0.82	0.82	0.92

I=20	High	r = 0.3	1	1.00	0.30	0.30	0.30	0.76	0.23	0.23	0.23
			2	0.30	1.00	0.30	0.30	0.23	0.76	0.22	0.23
			3	0.30	0.30	1.00	0.30	0.23	0.23	0.76	0.23
			4	0.30	0.30	0.30	1.00	0.23	0.23	0.23	0.76
		r = 0.6	1	1.00	0.60	0.60	0.60	0.76	0.45	0.45	0.45
			2	0.60	1.00	0.60	0.60	0.45	0.76	0.45	0.45
			3	0.60	0.60	1.00	0.60	0.45	0.45	0.76	0.45
			4	0.60	0.60	0.60	1.00	0.45	0.45	0.45	0.76
		r = 0.9	1	1.00	0.90	0.90	0.90	0.76	0.68	0.68	0.68
			2	0.90	1.00	0.90	0.90	0.68	0.76	0.68	0.68
			3	0.90	0.90	1.00	0.90	0.68	0.68	0.76	0.68
			4	0.90	0.90	0.90	1.00	0.68	0.68	0.68	0.76
	Low	r = 0.3	1	1.00	0.30	0.30	0.30	0.86	0.25	0.25	0.25
			2	0.30	1.00	0.30	0.30	0.25	0.86	0.25	0.25
			3	0.30	0.30	1.00	0.30	0.25	0.25	0.85	0.25
			4	0.30	0.30	0.30	1.00	0.25	0.25	0.25	0.85
		r = 0.6	1	1.00	0.60	0.60	0.60	0.86	0.51	0.51	0.51
			2	0.60	1.00	0.60	0.60	0.51	0.86	0.51	0.51
			3	0.60	0.60	1.00	0.60	0.51	0.51	0.86	0.51
			4	0.60	0.60	0.60	1.00	0.51	0.51	0.51	0.86
		r = 0.9	1	1.00	0.90	0.90	0.90	0.86	0.77	0.77	0.77
			2	0.90	1.00	0.90	0.90	0.77	0.86	0.77	0.77
			3	0.90	0.90	1.00	0.90	0.77	0.77	0.86	0.77
			4	0.90	0.90	0.90	1.00	0.77	0.77	0.77	0.85

Table 26. RMSE-SBs for CTT Subscale Scores over 100 Replicated Achievement Data

Data condition			Subscale	$RMSE - SB_{Raw}$	$RMSE - SB_{Kelley}$	$RMSE - SB_{HH}$	$RMSE - SB_{Haberman}$
I=10	High	0.3	1	0.56	0.56	0.88	0.55
			2	0.56	0.56	0.88	0.55
			3	0.56	0.56	0.87	0.55
			4	0.56	0.56	0.87	0.54
		0.6	1	0.56	0.56	0.67	0.51
			2	0.56	0.56	0.67	0.51
			3	0.56	0.56	0.67	0.51
			4	0.56	0.56	0.67	0.51
		0.9	1	0.56	0.56	0.42	0.40
			2	0.56	0.56	0.42	0.40
			3	0.56	0.56	0.42	0.40
			4	0.56	0.56	0.42	0.40
	Low	0.3	1	0.71	0.71	0.92	0.69
			2	0.71	0.71	0.92	0.69
			3	0.71	0.71	0.92	0.69
			4	0.71	0.71	0.92	0.69
		0.6	1	0.71	0.71	0.73	0.62
			2	0.71	0.71	0.73	0.63
			3	0.71	0.71	0.73	0.63
			4	0.71	0.71	0.73	0.63
		0.9	1	0.71	0.71	0.50	0.49
			2	0.71	0.71	0.50	0.49
			3	0.71	0.71	0.50	0.49
			4	0.71	0.71	0.50	0.49
I=20	High	0.3	1	0.42	0.42	0.84	0.42
			2	0.42	0.42	0.84	0.41
			3	0.42	0.42	0.84	0.41
			4	0.42	0.42	0.84	0.42
		0.6	1	0.42	0.42	0.63	0.39
			2	0.42	0.42	0.63	0.39
			3	0.42	0.42	0.63	0.39
			4	0.42	0.42	0.63	0.39
		0.9	1	0.42	0.42	0.36	0.32
			2	0.42	0.42	0.36	0.32
			3	0.42	0.42	0.36	0.32
			4	0.42	0.42	0.36	0.32
	Low	0.3	1	0.55	0.55	0.87	0.54
			2	0.55	0.55	0.87	0.54
			3	0.55	0.55	0.87	0.54
			4	0.55	0.55	0.87	0.54
		0.6	1	0.55	0.55	0.66	0.50

		2	0.55	0.55	0.66	0.50
		3	0.55	0.55	0.66	0.50
		4	0.55	0.55	0.66	0.50
	0.9	1	0.55	0.55	0.42	0.39
		2	0.55	0.55	0.42	0.39
		3	0.55	0.55	0.42	0.40
		4	0.55	0.55	0.42	0.39

Table 27. RMSE-SBs for CTT Subscale Scores over 100 Replicated Ability Data

Data condition			Subscale	$RMSE - SB_{Raw}$	$RMSE - SB_{Kelley}$	$RMSE - SB_{HH}$	$RMSE - SB_{Haberman}$
I=10	High	0.3	1	0.54	0.54	0.87	0.53
			2	0.54	0.54	0.86	0.53
			3	0.54	0.54	0.86	0.53
			4	0.54	0.54	0.86	0.53
		0.6	1	0.54	0.54	0.66	0.49
			2	0.54	0.54	0.66	0.49
			3	0.54	0.54	0.66	0.49
			4	0.54	0.54	0.66	0.49
		0.9	1	0.54	0.54	0.41	0.39
			2	0.54	0.54	0.41	0.39
			3	0.54	0.54	0.41	0.39
			4	0.54	0.54	0.41	0.39
	Low	0.3	1	0.69	0.69	0.91	0.67
			2	0.69	0.69	0.91	0.67
			3	0.70	0.70	0.91	0.67
			4	0.69	0.69	0.91	0.67
		0.6	1	0.69	0.69	0.72	0.61
			2	0.69	0.69	0.72	0.61
			3	0.69	0.69	0.71	0.61
			4	0.69	0.69	0.72	0.61
		0.9	1	0.69	0.69	0.49	0.48
			2	0.69	0.69	0.49	0.48
			3	0.69	0.69	0.49	0.48
			4	0.69	0.69	0.49	0.48
I=20	High	0.3	1	0.40	0.40	0.83	0.40
			2	0.40	0.40	0.83	0.40
			3	0.40	0.40	0.83	0.40
			4	0.40	0.40	0.83	0.40
		0.6	1	0.40	0.40	0.62	0.38
			2	0.40	0.40	0.62	0.38
			3	0.40	0.40	0.62	0.38
			4	0.40	0.40	0.62	0.38
		0.9	1	0.40	0.40	0.36	0.31
			2	0.40	0.40	0.36	0.31
			3	0.40	0.40	0.36	0.31
			4	0.40	0.40	0.36	0.31
	Low	0.3	1	0.54	0.54	0.86	0.52
			2	0.54	0.54	0.86	0.53
			3	0.54	0.54	0.86	0.53
			4	0.54	0.54	0.86	0.53
		0.6	1	0.54	0.54	0.65	0.49

		2	0.54	0.54	0.66	0.49
		3	0.54	0.54	0.65	0.49
		4	0.53	0.53	0.65	0.49
	0.9	1	0.54	0.54	0.41	0.39
		2	0.54	0.54	0.41	0.38
		3	0.54	0.54	0.41	0.39
		4	0.54	0.54	0.41	0.39

Table A28. RMSE-SBs for IRT Subscale θs over 100 Replicated Achievement Data

Subscale length	Subscale consistency	Between-subscales correlation	Subscale	$RMSE_{UIRT2PL}$	$RMSE_{MIRT2PL}$
I=10	High	0.3	1	0.57	0.56
			2	0.57	0.56
			3	0.58	0.56
			4	0.57	0.56
		0.6	1	0.57	0.52
			2	0.58	0.52
			3	0.58	0.52
			4	0.57	0.52
		0.9	1	0.58	0.45
			2	0.57	0.45
			3	0.58	0.45
			4	0.57	0.45
	Low	0.3	1	0.68	0.67
			2	0.68	0.67
			3	0.68	0.67
			4	0.68	0.66
		0.6	1	0.68	0.60
			2	0.68	0.60
			3	0.67	0.61
			4	0.67	0.60
		0.9	1	0.68	0.49
			2	0.68	0.49
			3	0.68	0.50
			4	0.67	0.49
I=20	High	0.3	1	0.45	0.44
			2	0.45	0.44
			3	0.45	0.44
			4	0.45	0.44
		0.6	1	0.45	0.42
			2	0.45	0.41
			3	0.45	0.42
			4	0.45	0.41
		0.9	1	0.46	0.39
			2	0.46	0.40
			3	0.45	0.40
			4	0.45	0.40
	Low	0.3	1	0.55	0.54
			2	0.54	0.53
			3	0.54	0.53
			4	0.54	0.53

	0.6	1	0.54	0.49
		2	0.54	0.50
		3	0.55	0.50
		4	0.55	0.50
	0.9	1	0.54	0.44
		2	0.54	0.45
		3	0.54	0.44
		4	0.54	0.44

Table A29. RMSE-SBs for IRT Subscale $\boldsymbol{\theta}$s over 100 Replicated Ability Data

Subscale length	Subscale consistency	Between-subscales correlation	Subscale	$RMSE_{UIRT2PL}$	$RMSE_{MIRT2PL}$
I=10	High	0.3	1	0.54	0.53
			2	0.53	0.52
			3	0.54	0.52
			4	0.54	0.53
		0.6	1	0.54	0.49
			2	0.54	0.48
			3	0.54	0.49
			4	0.54	0.49
		0.9	1	0.52	0.42
			2	0.54	0.42
			3	0.54	0.42
			4	0.54	0.42
	Low	0.3	1	0.66	0.64
			2	0.66	0.64
			3	0.66	0.65
			4	0.67	0.65
		0.6	1	0.66	0.60
			2	0.67	0.60
			3	0.65	0.59
			4	0.66	0.59
		0.9	1	0.66	0.49
			2	0.66	0.48
			3	0.66	0.48
			4	0.66	0.48
I=20	High	0.3	1	0.41	0.41
			2	0.41	0.40
			3	0.41	0.40
			4	0.41	0.40
		0.6	1	0.41	0.39
			2	0.41	0.38
			3	0.41	0.39
			4	0.41	0.38
		0.9	1	0.42	0.36
			2	0.41	0.36
			3	0.41	0.36
			4	0.42	0.36
	Low	0.3	1	0.52	0.51
			2	0.53	0.52
			3	0.53	0.52
			4	0.52	0.51

	0.6	1	0.53	0.48
		2	0.52	0.48
		3	0.53	0.48
		4	0.53	0.48
	0.9	1	0.52	0.41
		2	0.53	0.41
		3	0.52	0.40
		4	0.52	0.41

Table A30. Correlations between True Subscale θs and Estimated Subscale Scores in Achievement Tests (I=10)

Data condition			Measurement framework	Method	Correlation with true subscale θs			
Subscale Length	Subscale consistency	Between-subscales correlation			1	2	3	4
I=10	High	0.3	CTT-based	Raw	0.82	0.82	0.82	0.83
				Kelley	0.82	0.82	0.82	0.83
				HH	0.61	0.61	0.61	0.62
				Haberman	0.83	0.83	0.83	0.83
			IRT-based	U2PL	0.83	0.82	0.82	0.82
				M2PL	0.83	0.83	0.83	0.83
				OPI	0.75	0.74	0.74	0.75
		0.6	CTT-based	Raw	0.82	0.82	0.82	0.82
				Kelley	0.82	0.82	0.82	0.82
				HH	0.77	0.77	0.77	0.77
				Haberman	0.86	0.86	0.86	0.86
			IRT-based	U2PL	0.83	0.82	0.82	0.83
				M2PL	0.86	0.86	0.86	0.86
				OPI	0.80	0.80	0.80	0.81
		0.9	CTT-based	Raw	0.82	0.82	0.82	0.82
				Kelley	0.82	0.82	0.82	0.82
				HH	0.90	0.90	0.90	0.90
				Haberman	0.91	0.91	0.91	0.91
			IRT-based	U2PL	0.82	0.83	0.82	0.82
				M2PL	0.90	0.90	0.90	0.90
				OPI	0.86	0.86	0.86	0.86
	Low	0.3	CTT-based	Raw	0.74	0.74	0.74	0.74
				Kelley	0.74	0.74	0.74	0.74
				HH	0.58	0.58	0.58	0.57
				Haberman	0.76	0.76	0.76	0.76
			IRT-based	U2PL	0.73	0.74	0.74	0.74
				M2PL	0.75	0.76	0.76	0.76
				OPI	0.67	0.68	0.68	0.69
		0.6	CTT-based	Raw	0.74	0.74	0.74	0.74
				Kelley	0.74	0.74	0.74	0.74
				HH	0.74	0.74	0.73	0.73
				Haberman	0.80	0.80	0.80	0.80
			IRT-based	U2PL	0.74	0.74	0.74	0.74
				M2PL	0.80	0.80	0.80	0.80
				OPI	0.77	0.77	0.77	0.77
		0.9	CTT-based	Raw	0.74	0.74	0.74	0.74

	Kelley	0.74	0.74	0.74	0.74
	HH	0.87	0.87	0.86	0.87
	Haberman	0.87	0.87	0.87	0.87
IRT-based	U2PL	0.74	0.73	0.74	0.74
	M2PL	0.87	0.87	0.87	0.87
	OPI	0.85	0.85	0.85	0.85

Table A31. Correlations between True Subscale $\boldsymbol{\theta}$s and Estimated Subscale Scores in Achievement Tests (I=20)

Data condition			Measurement framework	Method	Correlation with True subscale θs			
Subscale Length	Subscale consistency	Between-subscales correlation			1	2	3	4
I=20	High	0.3	CTT-based	Raw	0.88	0.90	0.89	0.89
				Kelley	0.88	0.90	0.89	0.89
				HH	0.65	0.65	0.65	0.64
				Haberman	0.90	0.90	0.90	0.90
			IRT-based	U2PL	0.89	0.90	0.90	0.90
				M2PL	0.90	0.90	0.90	0.90
				OPI	0.74	0.77	0.77	0.76
		0.6	CTT-based	Raw	0.89	0.89	0.90	0.90
				Kelley	0.89	0.89	0.88	0.89
				HH	0.80	0.80	0.80	0.80
				Haberman	0.91	0.91	0.91	0.91
			IRT-based	U2PL	0.90	0.90	0.89	90
				M2PL	0.91	0.91	0.91	0.91
				OPI	0.83	0.84	0.84	0.85
		0.9	CTT-based	Raw	0.89	0.89	0.89	0.89
				Kelley	0.89	0.89	0.89	0.89
				HH	0.93	0.93	0.93	0.93
				Haberman	0.94	0.94	0.94	0.94
			IRT-based	U2PL	0.90	0.89	0.90	0.90
				M2PL	0.92	0.92	0.92	0.92
				OPI	0.90	0.90	0.90	0.90
	Low	0.3	CTT-based	Raw	0.84	0.84	0.84	0.84
				Kelley	0.84	0.84	0.84	0.84
				HH	0.62	0.62	0.62	0.62
				Haberman	0.85	0.85	0.85	0.85
			IRT-based	U2PL	0.84	0.84	0.84	0.85
				M2PL	0.85	0.85	0.85	0.85
				OPI	0.72	0.73	0.74	0.73
		0.6	CTT-based	Raw	0.84	0.84	0.84	0.84
				Kelley	0.84	0.84	0.84	0.84
				HH	0.78	0.78	0.78	0.78
				Haberman	0.87	0.87	0.87	0.87
			IRT-based	U2PL	0.85	0.84	0.84	0.84
				M2PL	0.87	0.87	0.87	0.87
				OPI	0.83	0.82	0.83	0.83
		0.9	CTT-based	Raw	0.84	0.84	0.84	0.84

	Kelley	0.84	0.84	0.84	0.84
	HH	0.91	0.91	0.91	0.91
	Haberman	0.92	0.92	0.92	0.92
IRT-based	U2PL	0.84	0.84	0.84	0.84
	M2PL	0.90	0.90	0.90	0.89
	OPI	0.90	0.90	0.90	0.90

Table A32. Correlations between True Subscale $\boldsymbol{\theta}$s and Estimated Subscale Scores in Ability Tests (I=10)

Data condition			Measurement framework	Method	Correlation with True subscale θs			
Subscale Length	Subscale internal consistency	Between-subscales correlation			1	2	3	4
10	High	0.3	CTT-based	Raw	0.85	0.85	0.85	0.85
				Kelley	0.85	0.85	0.85	0.85
				HH	0.63	0.63	0.63	0.63
				Haberman	0.85	0.85	0.85	0.85
			IRT-based	U2PL	0.85	0.85	0.85	0.85
				M2PL	0.85	0.86	0.85	0.85
				OPI	0.75	0.77	0.77	0.77
		0.6	CTT-based	Raw	0.85	0.85	0.85	0.85
				Kelley	0.85	0.85	0.85	0.85
				HH	0.78	0.78	0.78	0.78
				Haberman	0.87	0.87	0.87	0.88
			IRT-based	U2PL	0.84	0.85	0.85	0.85
				M2PL	0.87	0.87	0.87	0.87
				OPI	0.83	0.83	0.83	0.83
		0.9	CTT-based	Raw	0.85	0.85	0.85	0.85
				Kelley	0.85	0.85	0.85	0.85
				HH	0.91	0.91	0.91	0.91
				Haberman	0.91	0.92	0.92	0.92
			IRT-based	U2PL	0.85	0.84	0.85	0.85
				M2PL	0.90	0.90	0.90	0.90
				OPI	0.90	0.90	0.90	0.90
	Low	0.3	CTT-based	Raw	0.76	0.76	0.76	0.76
				Kelley	0.76	0.76	0.76	0.76
				HH	0.59	0.58	0.58	0.59
				Haberman	0.77	0.77	0.77	0.77
			IRT-based	U2PL	0.75	0.75	0.76	0.75
				M2PL	0.77	0.77	0.77	0.77
				OPI	0.70	0.68	0.70	0.68
		0.6	CTT-based	Raw	0.75	0.75	0.76	0.76
				Kelley	0.75	0.75	0.76	0.76
				HH	0.74	0.74	0.74	0.74
				Haberman	0.81	0.81	0.81	0.81
			IRT-based	U2PL	0.75	0.75	0.76	0.76
				M2PL	0.81	0.81	0.81	0.81
				OPI	0.78	0.78	0.79	0.78
		0.9	CTT-based	Raw	0.76	0.76	0.76	0.76

		Kelley	0.76	0.76	0.76	0.76
		HH	0.88	0.88	0.88	0.88
		Haberman	0.88	0.88	0.88	0.88
	IRT-based	U2PL	0.76	0.76	0.76	0.76
		M2PL	0.79	0.79	0.79	0.79
		OPI	0.87	0.87	0.87	0.87

Table A33. Correlations between True Subscale $\boldsymbol{\theta}$s and Estimated Subscale Scores in Ability Tests (I=20)

Data condition			Measurement framework	Method	Correlation with True subscale θs			
Subscale Length	Subscale internal consistency	Between-subscales correlation			1	2	3	4
I=20	High	0.3	CTT-based	Raw	0.91	0.91	0.91	0.91
				Kelley	0.91	0.91	0.91	0.91
				HH	0.65	0.65	0.66	0.65
				Haberman	0.92	0.92	0.92	0.91
			IRT-based	U2PL	0.91	0.91	0.91	0.91
				M2PL	0.92	0.92	0.92	0.92
				OPI	0.77	0.77	0.78	0.77
		0.6	CTT-based	Raw	0.91	0.91	0.91	0.91
				Kelley	0.91	0.91	0.91	0.91
				HH	0.81	0.81	0.81	0.81
				Haberman	0.92	0.92	0.92	0.92
			IRT-based	U2PL	0.91	0.91	0.91	0.91
				M2PL	0.92	0.93	0.92	0.92
				OPI	0.85	0.85	0.85	0.85
		0.9	CTT-based	Raw	0.91	0.91	0.91	0.91
				Kelley	0.91	0.91	0.91	0.91
				HH	0.94	0.94	0.94	0.94
				Haberman	0.95	0.95	0.95	0.95
			IRT-based	U2PL	0.91	0.91	0.91	0.91
				M2PL	0.92	0.92	0.92	0.92
				OPI	0.93	0.93	0.93	0.93
	Low	0.3	CTT-based	Raw	0.85	0.85	0.85	0.85
				Kelley	0.85	0.85	0.85	0.85
				HH	0.63	0.63	0.63	0.63
				Haberman	0.86	0.86	0.86	0.86
			IRT-based	U2PL	0.84	0.84	0.84	0.85
				M2PL	0.87	0.86	0.86	0.87
				OPI	0.76	0.74	0.76	0.76
		0.6	CTT-based	Raw	0.85	0.85	0.85	0.85
				Kelley	0.85	0.85	0.85	0.85
				HH	0.79	0.79	0.79	0.78
				Haberman	0.88	0.88	0.88	0.88
			IRT-based	U2PL	0.85	0.85	0.85	0.85
				M2PL	0.88	0.88	0.88	0.88
				OPI	0.84	0.83	0.84	0.83
		0.9	CTT-based	Raw	0.85	0.85	0.85	0.85

	Kelley	0.85	0.85	0.85	0.85
	HH	0.92	0.92	0.92	0.92
	Haberman	0.92	0.92	0.92	0.92
	U2PL	0.85	0.85	0.85	0.86
IRT-based	M2PL	0.90	0.90	0.90	0.90
	OPI	0.92	0.92	0.92	0.92

APPENDIX B

DATA SIMULATION AND SUBSCALE SCORING

```
**************************ACHIEVEMENT & ABILITY TEST
DATASETS*************************************/
/** simulated DATA for an ACHIEVEMENT test,and ABILITY test types.
*/
/** where bs ranges between p=[0.6,0.8] for achievement tests and p=[0.4,0.6] for an ability test
*/
/** therefore, b values were determined based on the criteria making p values like above
*/
/** where achievement tests with ~N(0.0, 2.0) and ability tests with ~N(-1.2, 2.0)          */
/** correlations between subscales are one of [0.3, 0.6, 0.9]                              */
/** The possible number of subscale items is 10, and 20.                                   */
/** N = 3000.                                                                              */
/** The number of total items is 40 and 80 and the number of subscales will be fixed to 4.
*/
/*************************************************************************************
************************/

LIBNAME dataGen 'C:\Users\CML Lab\Desktop\HJ\dissertation\simulation\dataGen';
LIBNAME scoring 'C:\Users\CML Lab\Desktop\HJ\dissertation\simulation\scoring';
LIBNAME corrSets 'C:\Users\CML Lab\Desktop\HJ\dissertation\simulation\scoreCorr';


%Macro subscale scores(seed,simulNum,testType,cstcy,nSub,nTotal,N,cor,Ncons);
/* "Ncons" is determined by four types follows: ach/ab, homo/hetero, 10/20 subscale items,
0.3/0.6/0.9 correlations */
/* <ach = 1, ab = 2/ homo = 1, hetero = 2 /10 = 1, 20 = 2/ 0.3 = 3, 0.6 = 6 0.9 = 9>
*/

%DO sim = 1 %TO &simulNum;
%let realSeed=&seed+&sim*3;

/*************************************************************************************
*******/
/* Randomly generate item estimates & true thetas                                      */
/* The as and bs were respectively created in log-normal and normal distributions     */
/* Items are generated with simple structures                                          */
/*************************************************************************************
*******/

data dataGen.itemParams&Ncons&sim;
        call streaminit(&realSeed+1);/*first simulation*/
```

```
	if &testType="Ach" then
		do s = 1 to 4; /* '4' is the number of subscales */
			do i = 1 to &nSub;
				type = &testType;
				totalItems = i + (s-1) * &nSub;
				scale = cats("sub",s);
				subItems = i;
				if &cstcy = "homo" then
					a = exp(rand("Normal",0.179,0.083)); /* ~logN(1.2,0.1)*/
				else if &cstcy = "hetero" then
					a = exp(rand("Normal",-0.227,0.100)); /* ~logN(0.8,0.08)
*/
				b = rand("Normal", -0.9, 0.5); /* most p value between 0.6 ~ 0.8 */
				drop s i;
				output;
			end;
		end;

	else if &testType="Ab" then
		do s = 1 to 4;
			do i = 1 to &nSub;
				type = &testType;
				totalItems = i + (s-1) * &nSub;
				scale = cats("sub",s);
				subItems = i;
				if &cstcy = "homo" then
					a = exp(rand("Normal",0.179,0.083));
				else if &cstcy = "hetero" then
					a = exp(rand("Normal",-0.227,0.100));
				b = rand("Normal",0.0, 0.5);
				drop s i;
				output;
			end;
		end;
run;

/*************************************************************************
**********/
/* Randomly generate true thetas for subscales from N examinees - with specific corrs  */
/*************************************************************************
**********/

proc iml;
	mean = {0, 0, 0, 0};
	corr = {1 &cor &cor &cor,
		&cor 1 &cor &cor,
```

```
                    &cor &cor 1 &cor,
                    &cor &cor &cor 1};
        var = {1 1 1 1};
        cov = corr # sqrt(var`*var);
        numExaminees = &N;
        call streaminit(&realSeed+2); /*second simulation*/
        theta = RandNormal(numExaminees, mean, cov);
        print (theta[:,]);
        sampleMean = mean(theta);
        sampleCov = cov(theta);
        print sampleMean;
        print sampleCov;
        create work.trueThetas from theta[colname={"trueTheta1" "trueTheta2" "trueTheta3"
"trueTheta4"}];
        append from theta;
        close work.trueThetas;

/* Adding examinee ID and reorder variables */
data dataGen.trueThetas&Ncons&sim;
        retain id;
        set work.trueThetas;
                id = _N_;
run;

/*************************************************************************
****/
/*        0. Make a file with true thetas, id, and item Numbers.
/*   1. Sort and then Merge item estimates & true thetas.
/*   2. Compute item-solving probabilities from item estimates & true thetas.
/*   3. Compare them with randomly generated univariate numbers.
/*   4. Get responses.
/*   5. Create two datasets with specific number of items using the given dataset.
*************************************************************************
*****/

/* True thetas with id and item numbers */

data work.idItem;
        do i = 1 to &N;
                do t = 1 to &nTotal;
                        id = i;
                        totalItems = t;
                        drop i t;
                        output;
                end;
        end;
```

```
run;

proc sort data = work.idItem;
	by id;
run;

proc sort data = dataGen.trueThetas&Ncons&sim;
	by id;
run;

data work.vecTrueThetas&Ncons&sim;
	merge dataGen.trueThetas&Ncons&sim work.idItem;
	by id;
run;

/* Sort item & true theta and Merge Item and True params */

proc sort data = work.vecTrueThetas&Ncons&sim;
	by totalItems;
run;

proc sort data = dataGen.itemParams&Ncons&sim;
	by totalItems;
run;

data work.params&Ncons&sim;
	merge work.vecTrueThetas&Ncons&sim dataGen.itemParams&Ncons&sim;
	by totalItems;
run;

/* Compute item-solving probabilities from Item estimates & True thetas */
/* Get responses based on the derived item and person estimates	*/

data dataGen.probs&Ncons&sim;
	set work.params&Ncons&sim;
	call streaminit(&realSeed+3);
	if scale = "sub1" then itemP = exp(a*(trueTheta1-b))/(1+exp(a*(trueTheta1-b)));
	else if scale = "sub2" then itemP = exp(a*(trueTheta2-b))/(1+exp(a*(trueTheta2-b)));
	else if scale = "sub3" then itemP = exp(a*(trueTheta3-b))/(1+exp(a*(trueTheta3-b)));
	else if scale = "sub4" then itemP = exp(a*(trueTheta4-b))/(1+exp(a*(trueTheta4-b)));
	r = rand("Uniform"); /*third simulation*/
	if itemP < r then resp = 0;
	else resp = 1;
	drop r;
run;
```

```
proc sort data = dataGen.probs&Ncons&sim;
	by id totalItems subItems;
run;

data work.tempResp;
	set dataGen.probs&Ncons&sim;
	drop scale totalItems subItems itemP trueTheta1 trueTheta2 trueTheta3 trueTheta4 type a
b;
run;

proc transpose data = work.tempResp
				out = dataGen.resp&Ncons&sim
	  prefix = r;
				var resp;
				by id;
run;

/* summing scores - one step to get subscale scores */
data work.summedScores;
	set dataGen.resp&Ncons&sim;
		sumTotal = sum(of r1-r40);
		sumSub1 = sum(of r1-r10);
		sumSub2 = sum(of r11-r20);
		sumSub3 = sum(of r21-r30);
		sumSub4 = sum(of r31-r40);
run;

/*mean of sumTotal, sumSub1-sumSub4 - another step to get subscale scores */
proc sql;
	create table sumStats as
	select id, mean(sumTotal) as meanTotal, mean(sumSub1) as meanSub1, mean(sumSub2)
as meanSub2,
	mean(sumSub3) as meanSub3, mean(sumSub4) as meanSub4, std(sumTotal) as SDTotal,
std(sumSub1) as SDSub1,
   std(sumSub2) as SDSub2, std(sumSub3) as SDSub3, std(sumSub4) as SDSub4 from
work.summedScores;
quit;

data work.summedScores&Ncons&sim;
	merge work.summedScores sumStats;
run;

/* This part was written to check if b values are properly set*/
proc univariate data=work.summedScores&Ncons&sim;
	var sumTotal;
	output pctlpre=P_ pctlpts=0 to 100 by 25;
```

```
run;

proc print data=data1;
run;

/* Create correlation files */
proc corr data = work.summedScores&Ncons&sim outp=work.corrsub1Total;
	var sumSub1 sumTotal;
run;

proc corr data = work.summedScores&Ncons&sim outp=work.corrsub2Total;
	var sumSub2 sumTotal;
run;

proc corr data = work.summedScores&Ncons&sim outp=work.corrsub3Total;
	var sumSub3 sumTotal;
run;

proc corr data = work.summedScores&Ncons&sim outp=work.corrsub4Total;
	var sumSub4 sumTotal;
run;


/* Retrieve corr values from tables to use in the computation of regression Coefficients */

proc sql;
	create table r1 as
	select _TYPE_, RSub1Total
	from work.Corrsub1total(rename = (sumTotal = RSub1Total))
	where _Type_= "CORR" and RSub1Total lt 1;

	create table r2 as
	select _TYPE_, RSub2Total
	from work.Corrsub2total(rename = (sumTotal = RSub2Total))
	where _Type_= "CORR" and RSub2Total lt 1;

	create table r3 as
	select _TYPE_, RSub3Total
	from work.Corrsub3total(rename = (sumTotal = RSub3Total))
	where _Type_= "CORR" and RSub3Total lt 1;

	create table r4 as
	select _TYPE_, RSub4Total
	from work.Corrsub4total(rename = (sumTotal = RSub4Total))
	where _Type_= "CORR" and RSub4Total lt 1;
```

```
quit;

/* Computing Cronbach alpha values based on responses */

proc corr data = work.summedScores&Ncons&sim alpha nocorr outp=alphaTotal;
	var r1-r40;
run;

proc corr data = work.summedScores&Ncons&sim alpha nocorr outp=alphaSub1;
	var r1-r10;
run;

proc corr data = work.summedScores&Ncons&sim alpha nocorr outp=alphaSub2;
	var r11-r20;
run;

proc corr data = work.summedScores&Ncons&sim alpha nocorr outp=alphaSub3;
	var r21-r30;
run;

proc corr data = work.summedScores&Ncons&sim alpha nocorr outp=alphaSub4;
	var r31-r40;
run;

/* Retrieve the Cronbach alpha value from each table to use in the computation of regression
Coefficients */

proc sql;
	create table alpT as
	select _TYPE_, alphaTotal
	from work.alphaTotal(rename = (r1 = alphaTotal))
	where _Type_= "RAWALPHA";

	create table alp1 as
	select _TYPE_, alpha1
	from work.alphaSub1(rename = (r10 = alpha1))
	where _Type_= "RAWALPHA";

	create table alp2 as
	select _TYPE_, alpha2
	from work.alphaSub2(rename = (r20 = alpha2))
	where _Type_= "RAWALPHA";

	create table alp3 as
	select _TYPE_, alpha3
	from work.alphaSub3(rename = (r30 = alpha3))
```

```
        where _Type_= "RAWALPHA";

        create table alp4 as
        select _TYPE_, alpha4
        from work.alphaSub4(rename = (r40 = alpha4))
        where _Type_= "RAWALPHA";

quit;

data work.allCorrs;
        merge r1 r2 r3 r4 alpT alp1 alp2 alp3 alp4;
run;

data work.reCoeffs&Ncons&sim;
        merge work.summedScores&Ncons&sim work.allCorrs;
run;

proc sql;
        create table dataGen.scores&Ncons&sim as
        select *, sum(RSub1Total) as subr1, sum(RSub2Total) as subr2, sum(RSub3Total) as
subr3, sum(RSub4Total) as subr4,
   sum(alphaTotal) as alpTotal, sum(alpha1) as alp1, sum(alpha2) as alp2, sum(alpha3) as alp3,
sum(alpha4) as alp4
   from work.reCoeffs&Ncons&sim;
quit;

data dataGen.scores&Ncons&sim;
        set dataGen.scores&Ncons&sim;
        drop _Type_ rSub1Total rSub2Total rSub3Total rSub4Total alphaTotal alpha1 alpha2
alpha3 alpha4;
run;

data scoring.finalSubscale scores&Ncons&sim;
        set dataGen.scores&Ncons&sim;
        /*Kelley's subscale scoring */
        kelleySub1 = meanSub1 + alp1*(sumSub1-meanSub1);
        kelleySub2 = meanSub2 + alp2*(sumSub2-meanSub2);
        kelleySub3 = meanSub3 + alp3*(sumSub3-meanSub3);
        kelleySub4 = meanSub4 + alp4*(sumSub4-meanSub4);
        kelleyTotal = kelleySub1 + kelleySub2 + kelleySub3 + kelleySub4;/* Kelley total sum
scores */

        /*Holland-Hosken's subscale scoring */
        /* _nom2 computes the correlation between true subscale scores and observed total
scores */
        HHb1_nom1 = (SDsub1*sqrt(alp1))/SDTotal;
```

```
        HHb1_nom2 = sqrt(alpTotal)*((subr1)/(sqrt(alp1)*sqrt(alpTotal))-(SDsub1*sqrt(1-
alp1))**2/(SDsub1*sqrt(alp1)*SDTotal*sqrt(alpTotal)));
        HHb2_nom1 = (SDsub2*sqrt(alp2))/SDTotal;
        HHb2_nom2 = sqrt(alpTotal)*((subr2)/(sqrt(alp2)*sqrt(alpTotal))-(SDsub2*sqrt(1-
alp2))**2/(SDsub2*sqrt(alp2)*SDTotal*sqrt(alpTotal)));
        HHb3_nom1 = (SDsub3*sqrt(alp3))/SDTotal;
        HHb3_nom2 = sqrt(alpTotal)*((subr3)/(sqrt(alp3)*sqrt(alpTotal))-(SDsub3*sqrt(1-
alp3))**2/(SDsub3*sqrt(alp3)*SDTotal*sqrt(alpTotal)));
        HHb4_nom1 = (SDsub4*sqrt(alp4))/SDTotal;
        HHb4_nom2 = sqrt(alpTotal)*((subr4)/(sqrt(alp4)*sqrt(alpTotal))-(SDsub4*sqrt(1-
alp4))**2/(SDsub4*sqrt(alp4)*SDTotal*sqrt(alpTotal)));
        HHSub1 = meanSub1 + (sumTotal-meanTotal)*(HHb1_nom1*HHb1_nom2);
        HHSub2 = meanSub2 + (sumTotal-meanTotal)*(HHb2_nom1*HHb2_nom2);
        HHSub3 = meanSub3 + (sumTotal-meanTotal)*(HHb3_nom1*HHb3_nom2);
        HHSub4 = meanSub4 + (sumTotal-meanTotal)*(HHb4_nom1*HHb4_nom2);
        HHTotal = HHSub1 + HHSub2 + HHSub3 + HHSub4; /* Holland and Hosken's total
sum scores */

        /* HABERMAN's subscale scoring */
        /* Regression coefficients for subscale scores */
        Haberb11 = (SDsub1*sqrt(alp1))*(sqrt(alp1)-HHb1_nom2*(subr1))/(SDsub1*(1-
subr1**2));
        Haberb21 = (SDsub2*sqrt(alp2))*(sqrt(alp2)-HHb2_nom2*(subr2))/(SDsub2*(1-
subr2**2));
        Haberb31 = (SDsub3*sqrt(alp3))*(sqrt(alp3)-HHb3_nom2*(subr3))/(SDsub3*(1-
subr3**2));
        Haberb41 = (SDsub4*sqrt(alp4))*(sqrt(alp4)-HHb4_nom2*(subr4))/(SDsub4*(1-
subr4**2));
        /* regression coefficients for total scores */
        Haberb12 = (SDsub1*sqrt(alp1))*(HHb1_nom2-sqrt(alp1)* subr1)/(SDTotal*(1-
subr1**2));
        Haberb22 = (SDsub2*sqrt(alp2))*(HHb2_nom2-sqrt(alp2)* subr2)/(SDTotal*(1-
subr2**2));
        Haberb32 = (SDsub3*sqrt(alp3))*(HHb3_nom2-sqrt(alp3)* subr3)/(SDTotal*(1-
subr3**2));
        Haberb42 = (SDsub4*sqrt(alp4))*(HHb4_nom2-sqrt(alp4)* subr4)/(SDTotal*(1-
subr4**2));

        /*Haberman Scores */
        HaberSub1 = meanSub1 + Haberb11*(sumSub1-meanSub1)+ Haberb12*(sumTotal-
meanTotal);
        HaberSub2 = meanSub2 + Haberb21*(sumSub2-meanSub2)+ Haberb22*(sumTotal-
meanTotal);
        HaberSub3 = meanSub3 + Haberb31*(sumSub3-meanSub3)+ Haberb32*(sumTotal-
meanTotal);
```

```
HaberSub4 = meanSub4 + Haberb41*(sumSub4-meanSub4)+ Haberb42*(sumTotal-
meanTotal);
HaberTotal = HaberSub1+HaberSub2+HaberSub3+HaberSub4;

/*Computing RMSE */
RMSE_RawSub1 = SDsub1*sqrt(1-alp1);
RMSE_RawSub2 = SDsub2*sqrt(1-alp2);
RMSE_RawSub3 = SDsub3*sqrt(1-alp3);
RMSE_RawSub4 = SDsub4*sqrt(1-alp4);

RMSE_KelleySub1 = (SDsub1*sqrt(alp1))* sqrt(1-alp1);
RMSE_KelleySub2 = (SDsub2*sqrt(alp2))* sqrt(1-alp2);
RMSE_KelleySub3 = (SDsub3*sqrt(alp3))* sqrt(1-alp3);
RMSE_KelleySub4 = (SDsub4*sqrt(alp4))* sqrt(1-alp4);

RMSE_HHSub1 =(SDsub1*sqrt(alp1))*sqrt(1-HHb1_nom2**2);
RMSE_HHSub2 =(SDsub2*sqrt(alp2))*sqrt(1-HHb2_nom2**2);
RMSE_HHSub3 =(SDsub3*sqrt(alp3))*sqrt(1-HHb3_nom2**2);
RMSE_HHSub4 =(SDsub4*sqrt(alp4))*sqrt(1-HHb4_nom2**2);

RMSE_HaberSub1 = RMSE_KelleySub1*sqrt(1-((HHb1_nom2-
sqrt(alp1)*subr1)/(sqrt(1-alp1)*sqrt(1-subr1**2)))**2);
RMSE_HaberSub2 = RMSE_KelleySub2*sqrt(1-((HHb2_nom2-
sqrt(alp2)*subr2)/(sqrt(1-alp2)*sqrt(1-subr2**2)))**2);
RMSE_HaberSub3 = RMSE_KelleySub3*sqrt(1-((HHb3_nom2-
sqrt(alp3)*subr3)/(sqrt(1-alp3)*sqrt(1-subr3**2)))**2);
RMSE_HaberSub4 = RMSE_KelleySub4*sqrt(1-((HHb4_nom2-
sqrt(alp4)*subr4)/(sqrt(1-alp4)*sqrt(1-subr4**2)))**2);

PRMSE_KelleySub1 = alp1;
PRMSE_KelleySub2 = alp2;
PRMSE_KelleySub3 = alp3;
PRMSE_KelleySub4 = alp4;

PRMSE_HHSub1 = HHb1_nom2**2;
PRMSE_HHSub2 = HHb2_nom2**2;
PRMSE_HHSub3 = HHb3_nom2**2;
PRMSE_HHSub4 = HHb4_nom2**
```

REFERENCES

Adams, R.J., Wilson, M.R., and Wang, W.C. (1997). The multidimensional random coefficients multinomial logit. *Applied Psychological Measurement, 21,* 1-24.

Ackerman, T.A. & Shu, Zhang (April 2009). Using Confirmatory MIRT Modeling to provide Diagnostic Information in Large Scale Assessment. Paper presented at the Annual meeting of the National Council on Measurement in Education, San Diego, CA.

Allen, M. J. & Yen, W. (2002). *Introduction to measurement theory.* Monterey, CA: Brooks-Cole.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) (1999). *Standards for educational and psychological testing.* Washington, DC: AERA.

Andrich, D (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Behrens, J. T., Mislevy, R. J., Bauer, M., Williamson, D. M., & Levy, R. (2004). Introduction to evidence centered design and lessons learned from its application in a global e-learning program. *The International Journal of Testing, 4,* 295-301. Birnbaum, 1957, 1958

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: application of an EM algorithm. *Psychometrika, 46*(4), 443-445.

Cai, L (2013). flexMIRT version 2: Flexible multilevel multidimensional item analysis and test scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302.de la Torre, 2011

de la Torre, J., & Douglas, J. (2008). Model evaluation and multiple strategies in cognitive diagnosis: An analysis of fraction subtraction data. *Psychometrika, 73*(4), 595-624.

DiBello, L. V., Roussos, L. A., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26.* Psychometrics (pp. 979-1027). Amsterdam, The Netherlands: Elsevier

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 Fransisco, CA.

Goodman, D. P., & Huff, K. (2006). *Findings from a national survey of teachers on the demand for and use of diagnostic information from large-scale assessments.* Manuscript in preparation, College Board, New York.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197.

Embretson, S. E. (1984). A general multicomponent latent trait model for response processes. *Psychometrika, 49*(2), 175-186.

Embretson, S. E. (1985). Multicomponent latent trait models for test design. In: Embretson, S.E. (Ed.), *Test Design: Developments in Psychology and Psychometrics.* Academic Press, New York, pp. 195-218.

Embretson, S. E. (1994). Applications of cognitive design systems to test development. In C. Reynolds (Ed.), *Advances in cognitive assessment: An interdisciplinary perspective* (pp. 107-135). New York: Plenum.

Embretson, S. E. (1996). Cognitive design systems and the successful performer: A study on spatial ability. *Journal of Educational Measurement, 33*(1), 29-39.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*(3), 380-396.

Embretson, S. E. (2006). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343-368.

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

Embretson, S. E., & Yang, X. (2013). A multicomponent latent trait model for diagnosis. *Psychometrika, 78*(1), 14-36.

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359-374.

Fu, J. (2005). A polytomous extension of the fusion model and its Bayesian parameter estimation. Unpublished doctoral dissertation, University of Wisconsin-Madison.

Haberman, S. J. (2005). When can subscale scores have value? *ETS Research Report Series, 1,* i-15.

Haberman, S. J. (2008). When can subscale scores have value? *Journal of Educational*

and Behavioral Statistics, 33, 204–229.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscale scores using multidimensional item response theory. *Psychometrika, 75,* 209-227.

Haberman, S. J., Sinharay, S., & Puhan, G. (2006). *Subscale scores for institutions* (ETS RR-06-13). Princeton, NJ: ETS.

Haertel, E. H. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*, 301–323.

Hartz, S. (2002). *A Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practice.* Unpublished doctoral thesis, University of Illinois at Urbana-Champaign.

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

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

Henson, R.A., Roussos, L., Templin, J.L. (2004) Cognitive diagnostic "fit" indices. Unpublished ETS Project Report, Princeton, NJ.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log linear models with latent variables. *Psychometrika, 74*(2), 191-210.

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

Jun, H., Lutz, M., Morrison, K., & Embretson, S. E. (2013). The Incremental Contribution of Cognitive Complexity to Specified Skills in Middle School Mathematics Test Items. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Junker, B.W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*(3), 258-272.

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Messick, S. (1989). Validity, In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education/Macmillan.

Mislevy, R. J. (1993). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19-39). Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*(4), 439-483.

Mislevy, R. J., Almond, R.G., & Steinberg, L. S., (2003). A brief introduction to Evidence-Centered Design. CSE Technical Report 632, The National Center for research on Evaluation, Standard, and Student Testing (CRESST), Center for the Study of Evaluation (SCE), UCLA, Los Angeles, CA.

Monaghan, W. (2006). The facts about subscale scores (ETS R&D Connections No. 4). Princeton, NJ: Educational Testing Service. Retrieved January 29, 2009, from http://www.ets.org/Media/Research/pdf/RD_Connec tions4.pdf

National Research Council (2001). *Knowing what students know: The science and design of educational assessment.* Committee on the Foundations of Assessment. Pellegrino, J., Chudowsky, N., and Glaswer, R., editors. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Lawrence Erlbaum.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361–373.

Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*(4), 293-311.

Rupp, A. A., & Templin, J. (2008). Unique characteristics of cognitive diagnostic models: A comprehensive review of the current state-of-the-art. *Measurement*, *6*(4), 219-262.

Rupp, A. A., Templin, J., & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York, NY: Guilford Press.

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

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscale scores basedon classical test theory: To report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21-28.

Sinharay, S., Puhan, G., & Haberman, S. (2011). An NCME Instructional Module on Subscale scores. *Educational Measurement: Issues and Practice, 30*(3), 29-40.

Skorupski, W., & Carvajal, J. (2010). A comparison of approaches for improving the reliability of objective level scores. *Educational and Psychological Measurement, 70*, 357-375.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 263-331). New York: Macmillan.

Snow, R. E., & Lohman, D. F. (1993). Cognitive psychology, new test design, and new test theory: An introduction. In N. Frederiksen, R. J. Mislevy & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 1-18). Hillsdale, NJ: Erlbaum.

Templin, J. L. (2006). *CDM user's guide.* Unpublished manuscript.

Templin, J. & Bradshaw, L. (2013). Measuring the reliability of diagnostic classification model examinee estimates. *Journal of Classification, 30*(2), 251-275.

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*(3), 287-305.

Wainer, H., Vevea, J.L., Camacho, F., Reeve, B., Rosa, K., Nelson, L., Swygert, K., & 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. 343-387). Mahwah, NJ: Lawrence Erlbaum.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL, *Journal of Educational Measurement, 37*(3), 203-220.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45,* 479-494.

Yao, L. H., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31* (2), 83–105.

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

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05–16). Princeton, NJ: Educational Testing Service.

von Davier M, & Yamamoto K. (2004). A class of models for cognitive diagnosis. Spearman Conference; 2004. Available from www.von-davier.com.

von Davier, M., & Yamamoto, K. (2007). Mixture distribution Rasch models and hybrid Rasch models. in M. Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models*.