A Restricted Bi-factor Model of Subdomain Relative Strengths and Weaknesses

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

Yu-Feng Chang

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

Mark L. Davison

August, 2015

Acknowledgements

In this long journey of obtaining my PhD, there are a lot of people I want to express my appreciation to. Especially, I would like to thank my advisor, Mark Davison, for his mentoring and support. Without him, I would not have been able to finish my degree. I have learned many things from him, and I am still learning more and more with each discussion we have. I thank you more than what words can describe here.

Also, I would like to thank my committee members, Ernest Davenport, Chun Wang, Andy Zieffler, and Michael Rodriguez for their thoughtful feedback during every stage of building my dissertation. I thank you for your knowledge and insight.

Likewise, I would like to thank my friends and cohort members in the QME program. I have learned more from you than any books. I thank you for what you have done to enrich my life.

Finally, I would like to the Chang family and Berry family, because of your support and positive thoughts, I was able to complete my dissertation.

Dedication

To my family: my father, who is supportive; my husband, who is patient; my daughter, who is sweet.

Abstract

There are increasing demands to report subscores in educational and psychological assessments. Subscores provide unique information about examinees (Sinharay, Puhan & Haberman, 2011). However, there has been much debate about reporting subscores because subscores require meeting certain standards and psychometric qualities as a prerequisite to reporting them. Because there is an increasing need for improving the methods of estimating subscores, multidimensional item response theory (MIRT) is one of the methods to estimate subscores.

One MIRT model is the item bi-factor model, which includes a general dimension on which all items load and specific dimensions corresponding to the subdomains from which the items come (Holzinger & Swineford's, 1937; Gibbons & Hedeker, 1992). However, there is a challenge to interpreting the specific dimension scores in the item bi-factor model while the general dimension score is readily interpreted. The specific dimension scores are residuals from the general factor and residuals can be difficult to interpret.

To solve this issue, a restricted bi-factor model was proposed in this paper. This paper contains a real data study and a simulation study to evaluate this model. The results of two studies, interpretation of the model, and practical application of the model were discussed.

Table of Contents

List of Tables

List of Figures

Chapter I: Introduction

Statement of the Problem

There are increasing demands to report subscores in educational and psychological assessments. According to the National Research Council report "Knowing What Students Know" (2001), the target of assessment is to provide particular information about an examinee's knowledge, skill, and abilities. Subscores provide unique information about examinees (Sinharay, Puhan & Haberman, 2011). Various audiences see the usefulness of subscores. For test takers, subscores are desirable because they can provide strengths and weaknesses about examinees' abilities and skills to help with future remedial studies. For educators, subscores can suggest remedies for examinees lacking certain abilities. For state and academic institutions, subscores could be tools to evaluate a curriculum's effectiveness. For college and university admission officers, subscores can provide distinct information for admission purposes when candidates have similar total scores. For policy makers, subscores may be guides to change state curriculums and provide more funding in different content areas (Haladyna & Kramer, 2004; Leading & Monaghan, 2006).

However, there has been much debate about reporting subscores in the field of educational and psychological assessment because subscores require meeting certain standards and psychometric qualities as a prerequisite to reporting them. Standard 5.12 of the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) states: "Scores should not be reported for individuals

unless the validity, comparability, and reliability of such scores have been established." Moreover, Standard 1.12 of the same document states: "If a test provides more than one score, the distinctiveness of the separate scores should be demonstrated." Also, Standard 2.1 states: "For each total score, subscores, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported" (AERA, APA, NCME, 1999). Inaccurate information about subscores or reporting subcores without proper interpretation could lead to inaccurate decisions.

The reasons that subscores may lack psychometric qualities are as follows. First, there may not be enough items within each domain. Tate (2004) stated sufficient validity and reliability for subscores can minimize incorrect individual instructional decisions. Second, subscores in educational measurements often refer to domain areas. If an assessment with several domains is constructed to measure a single construct, little reason exists to expect useful subscores (Sinharay, Puhan & Haberman, 2011). In this case, correlations between domains tend to be relatively high. Correlations between domains may be a sign that the assessment is unidimensional.

Once the psychometric quality and number of dimensions of assessments are assured, several alternatives can be used to compute subscores. Multidimensional item response theory (MIRT) is one method to estimate subscores. Ability estimation using the MIRT method and augmentation methods perform best in estimating the true subscores compared to other methods (Dwyer, Boughton, Yao, Steffen, & Lewis, 2006; Fu & Qu, 2010). Yao and Boughton (2007) showed that MIRT ability estimation

performs better than percentage or number-correct or the objective performance index (OPI), which is a unidimensional IRT subscale scoring approach.

One of the multidimensional item response theory models is the item bi-factor model. One of the advantages for the item bi-factor model is that it accounts for a general dimension on which all items load and specific dimensions corresponding to the subdomains from which the items come (Holzinger & Swineford's, 1937; Gibbons & Hedeker, 1992). However, there is a challenge to interpreting the specific dimension scores in the item bi-factor model while the general dimension score is readily interpreted. The specific dimension scores are residuals from the general factor and residuals can be difficult to interpret (see Chapter 2).

Standard 2.1 states: "For each total score, subscores, or combinations of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported" (AERA, APA, NCME, 1999). Reporting subscores should be interpetable or it is not worth reporting them. In other words, it is a big drawback to apply the item bi-factor model since it is difficult to interpret its subscores.

Significance of the Study

The purpose of this dissertation is to develop a restricted bi-factor model in which the general dimension represents the examinee's overall performance in a domain, and each specific dimension represents a deviation from that overall performance. The specific dimension scores describe the examinee's pattern of strengths and weaknesses relative to the examinee's overall performance.

Moreover, the real data study in this dissertation is to demonstrate the method of the restricted bi-factor model and how to interpret the general dimension score and the specific dimension scores. The simulation study aims to answer the following research questions.

1. How are reliability, sensitivity, specificity, and recovery of person parameters, influenced by the level of correlation between dimensions?
2. How are reliability, sensitivity, specificity, and recovery of person parameters, influenced by the number of items in the test?
3. How are reliability, sensitivity, specificity, and recovery of person parameters, influenced by the number of dimensions in the test?

Chapter 2 reviews the literature, including the unidimensional item response theory (UIRT) and multidimensional item response theory (MIRT) models, the connection between factor analysis and IRT, person parameters estimation, the bi-factor model and application, and the use of subscores. This study presents a new approach to modeling subscores. Chapter 3 describes the method of the restricted bi-factor model that is demonstrated with real and simulated data. The conditions of the simulation study, details of the real data study, and evaluation criteria are described in Chapter 3. Chapter 4 presents the results of the real data study and simulation study. Chapter 5 concludes with a discussion of the results and implications for the field of educational and psychological measurement.

Chapter II: Literature Review

Introduction

There are many advantages to implementing the bi-factor multidimensional IRT model on multidimensional data. However, a majority of the literature on multidimensional IRT has focused on estimation and interpretation of the item parameters whereas interpretation of person parameters has rarely been emphasized. For example, Simms, Grös, Watson, and O'Hara (2008) applied the bi-factor model on the inventory of depression and anxiety symptoms. This study compared the bi-factor model and the one-factor model. This study only covered magnitudes of the general and specific factor loadings and item difficulty parameters. Nevertheless, this study only discussed the estimation method of person parameter and person scores of the general factor. According to Standard 2.1 of the *Standards for Educational and Psychological Testing* each total score, subscores, or combination of scores should be interpreted. Moreover, interpretation of the person parameters can be more meaningful, especially in diagnosing examinees' strengths and weaknesses and providing useful feedback for examinees on the specific domains of the assessments. This study demonstrates a restricted bi-factor IRT model, which can be a good approach to providing meaningful person parameters and feedback on individual differences.

This chapter reviews previous research to provide background for the restricted bi-factor model. The first section below reviews unidimensional item response theory (UIRT) and multidimensional item response theory (MIRT). The second section below illustrates the connection between factor analysis and IRT. The third section below

conceptualizes item parameters estimation and person parameter estimation. Lastly, the fourth section reviews the bi-factor model, including the full-information item bi-factor analysis, the bi-factor model in the IRT framework, and some studies applying the bi-factor model.

Unidimensional Item Response Theory (UIRT)

The critical restriction for traditional IRT (UIRT) is the unidimensional assumption meaning the construct of measurement contains only a single ability. In educational assessment fields, math assessments are a very noticeable example of multidimensionality. Under the traditional IRT assumption, only overall math ability is measured on math assessments although math assessments have a strong connection to reading ability. As the math example shows, the unidimensional assumption does not always hold. In other words, there could be more than one dimension underlying personality and educational assessments.

Item response theory (IRT) uses latent traits of individuals and items as predictors of observed responses. The major advantage of IRT is that person parameters and item parameters in IRT are located on the same scale. IRT has been commonly used in educational and psychological measurement. However, traditional IRT has three strong assumptions. The three assumptions under traditional IRT are unidimensionality, local independence, and functional form. First, the unidimensionality assumption means only one ability or one trait is measured by examinees' performance on a set of items. Nevertheless, in a real application, there will most likely be some degree of violation of the unidimensionality assumption. The degree of violation may or may not be an issue. If the violation is severe, unidimensional IRT may not be useful. Instead,

multidimensional IRT models should be considered. Second, the assumption of local independence is that, conditional on ability level, how a person responds to an item is independent of responses to any other items. Under unidimensionality local independence can be explained as how the examinee responds to an item only depends on the examinee's latent traits, not how the examinee answers any other questions. The third assumption is functional form, which is the function specified by the model relating responses to traits or abilities (De Ayala, 2009).

The assumption of unidimensionality among the three assumptions is more critical since this assumption is always violated to some degree. However, the severe violation of this assumption may produce incorrect estimation as well as incorrect interpretations. In this situation, multidimensional IRT models should be used instead. The simple logistic forms with the one-parameter logistic (1PL), two-parameter logistic (2PL), and three-parameter logistic (3PL) are stated below. The 2-parameter normal ogive model is

$$P(x_{ij}=1|\theta_j,a_i,b_i)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp\left[-\frac{z^2}{2}\right]dz \tag{2.1}$$

where $z=a_i(\theta_j-b_i)$, a_i is the item discrimination parameter, b_i is the difficulty parameter, and $P(x_{ij}=1|\theta_j,a_i,b_i)$ is the probability of person *j* answering correctly item *i*. Using the logistic form, the 2-parameter model can be expressed

$$P(x_{ij}=1|\theta_j,a_i,b_i)=\frac{\exp[1.7a_i(\theta_j-b_i)]}{1+\exp[1.7a_i(\theta_j-b_i)]} \tag{2.2}$$

where 1.7 is a scaling factor.

The one-parameter logistic model is

$$P(x_{ij}=1|\theta_j,b_i,a)=\frac{\exp[1.7a(\theta_j-b_i)]}{1+\exp[1.7a(\theta_j-b_i)]} \quad (2.3)$$

In the 1PL model, items have a constant value for the discrimination parameter (a). One of the 1PL models is the Rasch model for which the discrimination parameter (a) is equal to 1.

The 3-parameter logistic model (Birnbaum, 1968) is

$$P(x_{ij}=1|\theta_j,a_i,b_i,c_i)=c_i+(1-c_i)\frac{\exp[1.7a_i(\theta_j-b_i)]}{1+\exp[1.7a_i(\theta_j-b_i)]} \quad (2.4)$$

where c_i is the lower asymptote parameter (the guessing parameter) for the item.

Multidimensional Item Response Theory (MIRT)

Multidimensional IRT (MIRT) means the test manifests the influence of more than one latent trait. From the model point of view, each examinee has several abilities and each item has several discrimination parameters but only one guessing parameter and one intercept. There are three assumptions in the MIRT model: a functional form assumption, the conditional independence assumption, and a dimensionality assumption. A functional form assumption means the data follow the function specified by the model. The second assumption, the conditional independence, indicates that the conditional distributions of the item responses are all independent of each other (Lord & Novick, 1968). Third, a dimensionality assumption states that the observations on the variable are a function of a set of continuous latent person variable. Essentially, dimensionality assessment is required prior to MIRT analysis (De Ayala, 2009).

There are two types of multidimensional IRT models. One is the compensatory model (Rasch, 1960; Reckase, 1972) and the other is the non-compensatory model (Sympson, 1978; Whitely, 1980). The compensatory model has an assumption that high ability on one dimension can compensate for low ability on another dimension in terms of probability of correct response whereas the non-compensatory model implies that answering the item correctly required the ability on each dimension to have non-zero probability. The compensatory model is based on a linear combination of θ-coordinates which yield the same sum with different combinations of θ-values. Take an item with two dimensions for example. For the non-compensatory model, a person with very high ability on one dimension and very low ability on the other dimension has very low probability of answering correctly on this item. However, for the compensatory model, the same person has some substantial probability of answering the item correctly (Ansley & Forsyth, 1985; Reckase, 2009). The compensatory model is more dominant in the application research literature, and it is connected to factor analysis. In addition, it is difficult to estimate the non-compensatory model parameters (Spray, Davey, Reckase, Ackerman & Carlson, 1990). Because this study focused on compensator MIRT models for dichotomous items, several compensatory MIRT models for dichotomous items are described here.

Multidimensional Extension of the Three-Parameter Logistic Model (M3PL). The most used and updated compensatory multidimensional extension of the three-parameter logistic model (M3PL) modified by Reckase (1985) is

$$P(x_{ij}=1|\boldsymbol{\theta}_j,\boldsymbol{a}_i,c_i,d_i)=c_i+(1-c_i)\frac{\exp[1.7(\boldsymbol{a}_i\boldsymbol{\theta}'_j+d_i)]}{1+\exp[1.7(\boldsymbol{a}_i\boldsymbol{\theta}'_j+d_i)]} \quad (2.5)$$

where $\boldsymbol{a}_i\boldsymbol{\theta}'_j = \sum_{k=1}^{m} a_{ik}\theta_{jk}$, $\boldsymbol{a}_i$ is a $1\times m$ vector of item discrimination parameters which can also be called slope parameters, $\boldsymbol{\theta}_j$ is a $1\times m$ vector of person parameters, m is the number of dimensions, c_i is the lower asymptote parameter (the guessing parameter), and d_i is an intercept parameter. An overall discrimination index is defined as the multidimensional discrimination index (MDISC; Reckase, 1985).

$$\text{MDISC} = \sqrt{\sum_{k=1}^{m} a_{ik}^2} \tag{2.6}$$

MDIFF in MIRT models, which has the same interpretation as the *b*-parameter in UIRT is given by Equation 2.7.

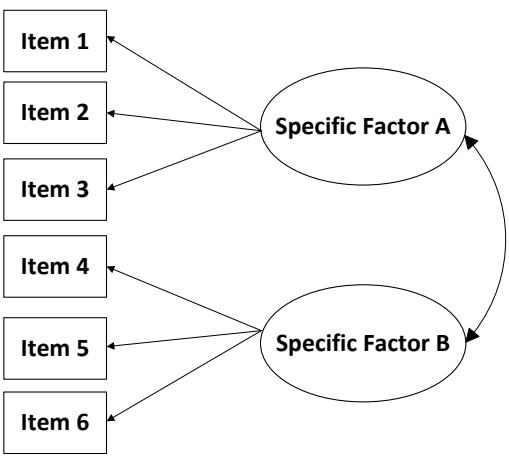
$$\text{MDIFF} = \frac{-d}{\sqrt{\sum_{k=1}^{m} a_{ik}^2}} \tag{2.7}$$

Multidimensional Extension of the Two-Parameter Logistic Model (M2PL). The M2PL is obtained by setting the lower asymptote parameter (c_i) to 0 in Equation 2.5.

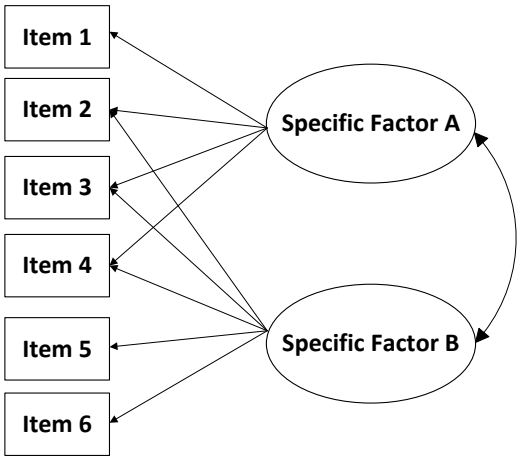
Multidimensional Extension of the Rasch Model. The multidimensional extension of the Rasch Model is obtained by constraining the lower asymptote parameter (c_i) to 0 and the discrimination parameters (a_{ik}) to 1.0 in Equation 2.5. This model becomes a UIRT version of Rasch model when *m* = 1 (Reckase, 2009). Moreover, Adams, Wilson, and Wang (1997) proposed the multidimensional generalization of the Rasch model, the multidimensional random coefficients multinomial logit model (MRCML). The general form of this model can be applied to dichotomous and polytomous items. The MRCML for dichotomous items is identical to theM2PL.

However, the discrimination parameters for the M2PL are estimates from the data, whereas the discrimination parameters for the MRCML are specified by the test developer (Adams, Wilson & Wang, 1997; Reckase, 2009). Adams, Wilson, and Wang (1997) introduced two models in term of test structure, between-item multidimensionality models and within-item multidimensionality models. Between-item multidimensionality models, which are also called simple structure models or multi-unidimensonal models, contain several unidimensional subscales. Within-item multidimensionality models (complex structure models) contain items each of which relates to more than one dimension (Adams, Wilson & Wang, 1997; Wang, Chen, & Cheng, 2004). Figure 2.1 illustrates between-item multidimensionality models and within-item multidimensionality models.

Figure 2.1. Between-item Multidimensionality Models and Within-item Multidimensionality Models



Model A: A Between-item Model



Model B: A Within-item Model

Latent Variable Models. Factor analysis and MIRT have essentially identical mathematical formulas. Four major types of latent variable models for confirmatory

factor analysis are reviewed here (e.g., Rindskopf & Rose, 1988; Chen, West & Sousa, 2006; Reise, Morizot & Hays, 2007; Rijmen, 2010). These models are graphically represented in Figure 2.2. Model A is the multi-unidimensional model in which all the item responses can be accounted for by one common factor. Model A can also be represented as a unidimentional IRT model. Model B is the complex multidimensional model. There are two common factors in the model and the common factors are correlated with each other. The item intercorrelatons can be explained by the multiple correlated traits here. In Figure 2.2, Model C is the bi-factor model. There is one general factor and three specific factors. The general factor and the specific factors are orthogonal in the bi-factor model. After accounting for the general factor, the specific factors explain variance over and above the general factors. The general factor in the bi-factor model explains the item intercorrelations across domains. The specific factors are independent of each other in the bi-factor model because the item intercorrelaitons already are accounted for by the general factor. Model D is the second-order model. The second-order model and the bi-factor model are both used for assessments with multiple highly related domains. They both have similar factor structures. However, the second-order model is potentially applicable when lower-order factors (the specific factors) are substantially correlated with each other and the higher-order factor (the general factor) accounts for the relationship among the lower-order factors and the higher-order factor is linearly dependent on the lower-order factors.

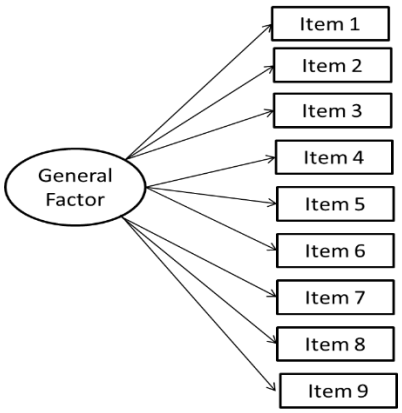
The second-order model is also called a high-order model and the hierarchical model is a special case of the bi-factor model. The relationship between the higher-order

factor model and the hierarchical factor model was studied by Yung, Thissen, & McLeod, (1999).

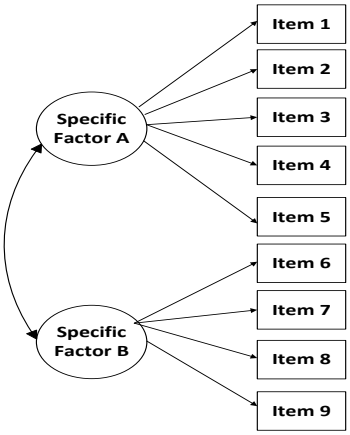
Schmid and Leiman (1957) proposed a Schmid-Leiman transformation for deriving hierarchical factor solutions from higher-order factor solutions with a simple factor clusters structure.

Figure 2.2. Four Latent Variable Models

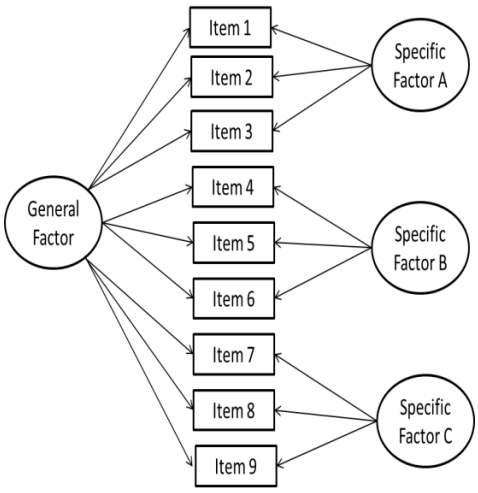
Model A:
The unidimensional model



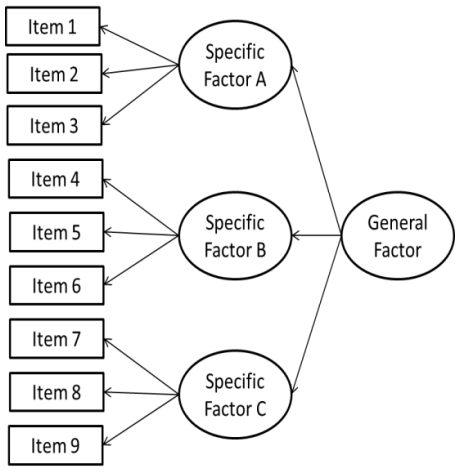
Model B:
The non-hierarchical multidimensional model



Model C:
The bi-factor model

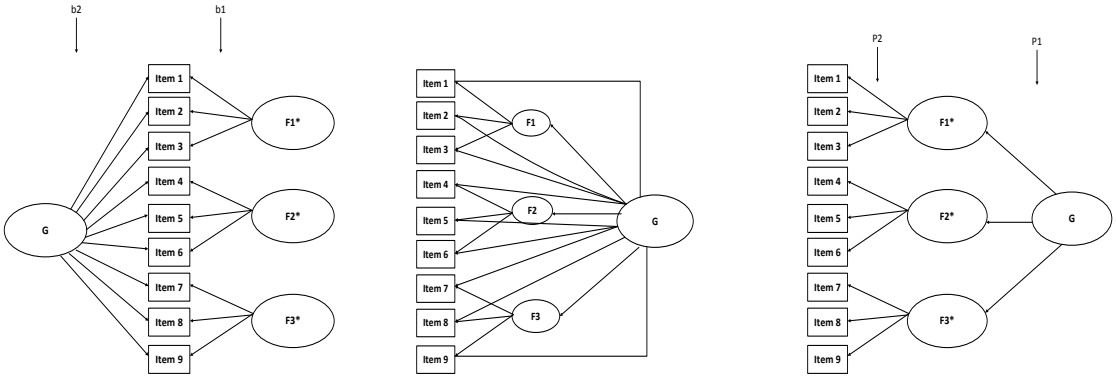


Model D:
The second-order model



The study showed that the Schmid-Leiman transformation produces a- constrained hierarchical factor solution. An unconstrained hierarchical factor model is an equivalent higher-order factor model with direct effects (loadings) on the manifest variables from the higher-order factors. Hence, the class of higher-order factor models (without direct effects of higher-order factors) is nested within the class of unconstrained hierarchical factor models. These models are graphically represented in Figure 2.3.

Figure 2.3. Higher-order Factor Models and Hierarchical Factor Models



Model A: A General Hierarchical Factor Model

Model B: A Higher-order Factor Model

Model C: A Higher-order Factor Model with Direct Effects

Connection between the MIRT Models and Full-information Item Factor Analysis

The MIRT models can be re-parameterized as an item factor model. Item factor analysis does not require calculating inter-item correlation coefficients and is not strongly restricted by the number of items. Essentially, item factor analysis is the classical linear factor model adapted for binary items (Bock, & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988).

Full-Information Item Factor Analysis (FIIFA). Full-information item factor analysis (FIIFA) was developed by Bock, Gibbons and Muraki (1988). Full-information

item factor analysis uses the frequencies of all distinct item response vectors (Bartholomew, 1980). In contrast, the limited-information method is based on low-order joint occurrence frequencies of the item scores (Cristoffersson, 1975).

The full-information item factor analysis is adopted from Thurstone's multiple–factor model, which assumes the *M*-factor model. Thurstone's multiple–factor model is as follows:

$$y_{ij} = \alpha_{i1}\theta_{1j} + \alpha_{i2}\theta_{2j} + ... + \alpha_{im}\theta_{mj} + \varepsilon_{ij} \tag{2.8}$$

$$\varepsilon_{ij} \sim N(0, \sigma_i^2)$$

where y_{ij} is response of person *j* to item *i* and ε_{ij} is an error term . The method posits a correct response of person *j* to item *i* when y_{ij} equals or exceeds γ_i as a threshold and yields an incorrect response otherwise. The full-information item factor model is the probability of an item score, $x_{ij} = 1$, which is a correct response of person *j* to item *i* with abilities $\boldsymbol{\theta}_j = (\theta_{1j}, \theta_{2j}, ..., \theta_{mj})$.

$$P(x_{ij} = 1 | \boldsymbol{\theta}_j) = \tag{2.9}$$

$$\frac{1}{\sqrt{2\pi}\sigma_i} \int_{\gamma_i}^{\infty} \exp\left[-\frac{1}{2} \left(\frac{y_i - \sum_{k=1}^{m} \alpha_{ik}\theta_{kj}}{\sigma_i} \right)^2 \right] dy_i = \Phi\left(-\left(\frac{\gamma_i - \sum_{k=1}^{m} \alpha_{ik}\theta_{kj}}{\sigma_i} \right) \right) = \Phi_i(\boldsymbol{\theta}_j)$$

$$y_i \sim N(0,1)$$

where γ_i is a threshold of item *i*, α_{ik} is the factor loading of item *i* for θ_{kj}, σ_i^2 is the error variance, and y_i is the latent variable. The sample of examinees is drawn from a population with abilities following the multivariate distribution, which

is $\boldsymbol{\theta}_j \sim \text{MVN}(0, \boldsymbol{I})$, but this assumption might be relaxed to have correlated factors and a non-normal distribution. If the incorrect response is $x_{ij} = 0$, the conditional probability can be written as $1 - \Phi_i(\boldsymbol{\theta}_j)$ (Bock, & Aitkin, 1981; Bock, Gibbons, & Muraki, 1988).

Estimation of the Item Thresholds and Factor Loadings. Takane and De Leeuw (1987) proved the equivalence of the marginal likelihood of the two-parameter normal ogive model of multidimensional item response theory and the factor analysis of dichotomized variables. Therefore, estimation of the item thresholds and factor loadings in item factor analysis can be obtained based on the two-parameter normal ogive model of MIRT. The 2PL version of MIRT is shown in Equation 2.10.

$$P(x_{ij} = 1 | \boldsymbol{\theta_j}, \boldsymbol{a_i}, d_i) = \frac{1}{\sqrt{2\pi}} \int_{-(\boldsymbol{a_i}\boldsymbol{\theta}_j' + d_i)}^{\infty} \exp\left[-\frac{t^2}{2}\right] dt = \Phi\left(\boldsymbol{a_i}\boldsymbol{\theta}_j' + d_i\right) \tag{2.10}$$

The parameters a_{ik} and d_i in IRT (Equation 2.10) can be parameterized in terms of the parameters σ_i, γ_i and α_{ik} in the item factor analysis as Equation 2.11.

$$\frac{\gamma_i - \sum_{k=1}^{m} \alpha_{ik}\theta_{jk}}{\sigma_i} = -\sum_{k=1}^{m} a_{ik}\theta_{jk} - d_i \tag{2.11}$$

The parameters a_{ik} and d_i in MIRT are related to the item factor parameters as follows:

$a_{ik} = \frac{\alpha_{ik}}{\sigma_i}$ and $d_i = -\frac{\gamma_i}{\sigma_i}$.

There are several advantages of using the MIRT approach rather than factor analysis. First, there are less parameters being estimated in the MIRT framework. Only two parameters are estimated in the MIRT framework versus the three parameters

estimated in item factor analysis. Second, factor analysis is primarily a technique for data reduction while MIRT is a technique for modeling the interaction between person and test items. Reckase and Hirsch (1991) found that having less dimensions could degrade information of item and person parameters interaction, but more dimensions doesn't cause severe problems. Consequently, MIRT might be a better tool than factor analysis because factor analysis is viewed as a data reduction tool. Also, MIRT analysis uses the same latent space across tests and samples. All analyses can be on a common coordinate system so item parameters for all items are on common metrics (Takane & De Leeuw, 1987; Reckase & Hirsch, 1991; Reckase 2009).

Item Parameter Estimation

Marginal Maximum Likelihood. Item parameters can be estimated by the marginal maximum likelihood method for UIRT (Bock & Aiteken, 1981; Bock, Gibbons, & Muraki , 1988). Under the multidimensional extension of the conditional independence assumption, the marginal probability of person *j* responding to number of items *n* with pattern $\boldsymbol{x}_j = [x_{j1}, x_{j2}, ..., x_{jn}]$ conditional on ability $\boldsymbol{\theta}_j$ can be expressed as

$$P(\boldsymbol{x} = \boldsymbol{x}_j \mid \boldsymbol{\theta}_j) = \prod_{i=1}^{n} [\Phi_i(\boldsymbol{\theta}_j)]^{x_{ij}} [1 - \Phi_i(\boldsymbol{\theta}_j)]^{1-x_{ij}} = L(\boldsymbol{x}_j \mid \boldsymbol{\theta}_j) \tag{2.12}$$

Equation 2.12 is the likelihood function conditional on the trait vector $\boldsymbol{\theta}_j$. If people are randomly sampled from a population with continuous ability distribution $g(\boldsymbol{\theta})$, the unconditional likelihood function of response pattern $\boldsymbol{x}_j$ for a *k*-dimensional latent trait is

$$\tilde{P}_\ell = P(\boldsymbol{x} = \boldsymbol{x}_j) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} L_i(\boldsymbol{\theta})g(\boldsymbol{\theta})d\theta_1\ldots d\theta_k = \tag{2.13}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} L_i(\boldsymbol{\theta})g(\theta_1)\ldots g(\theta_k)d\theta_1\ldots d\theta_k$$

Applying the *k*-dimensional Gauss-Hermite quadrature to Equation 2.13, the numerical approximation of this integral can be expressed as

$$\tilde{P}_\ell = P(\boldsymbol{x} = \boldsymbol{x}_j) = \tag{2.14}$$

$$\sum_{qk}^{Q}\ldots\sum_{q2}^{Q}\sum_{q1}^{Q} P(x = x_j \mid X_{q1}, X_{q2},\ldots, X_{qk})A(X_{q1})A(X_{q2})\ldots A(X_{qk})$$

where X_k is a quadrature point in *k*-dimensional space and $A(X_{qk})$ is the weight for the quadrature point in the separate dimensions. That is, MML does not estimate person parameters but it estimates item parameters using the quadrature procedure in Equation 2.14. The quadrature procedure divides a continuous distribution into a discrete approximation by means of grouping the scores into a small number of groups. Item parameters are estimated by a set of means over the quadrature points. The biggest disadvantage of the numerical approximation of Equation 2.14 is that when the number of dimensions increases, the computation becomes much slower.

Equation 2.14 is an estimate of the probability of observing an item score string in the population of examinees represented by the multivariate density function $g(\boldsymbol{\theta})$. Furthermore, the probability of a set of item score strings, U, can be represented as

$$L(\boldsymbol{U}) = \frac{N!}{r_1!,\ r_2!,\ \ldots,\ r_s!} P(\boldsymbol{x} = \boldsymbol{x}_1)^{r_1} P(\boldsymbol{x} = \boldsymbol{x}_2)^{r_2}\ldots P(\boldsymbol{x} = \boldsymbol{x}_s)^{r_s} \tag{2.15}$$

where N is the number of examinees in a sample, s is the number of item score strings which is $s \leq \min(N, \mathbf{2}^n)$, and r_s is the frequency of occurrence for the item score string $\boldsymbol{x}_s$ for n items. In order to estimate the item parameters, the unconditional likelihood function in Equation 2.15 should be transformed to the log-likelihood form in Equation 2.16. The item parameters can be obtained by maximizing Equation 2.16.

$$\log(L(\boldsymbol{U})) = r_1 \log P(\boldsymbol{x} = \boldsymbol{x}_1) + r_2 \log P(\boldsymbol{x} = \boldsymbol{x}_2) + \ldots + r_s \log P(\boldsymbol{x} = \boldsymbol{x}_s) \qquad (2.16)$$

Expectation/Maximization (EM) algorithm. The estimation of item parameters can be obtained by implementing the expectation-maximization (EM) algorithm to maximize Equation 2.16. Marginal maximum likelihood estimation of item parameters with the EM algorithm was developed by Bock and Aitkin (1981). The numerical procedure involves two steps for an item per cycle. The E step is the expectation step, and the M step is the maximization step. The E step is not iterative. In the E step, using the provisional item parameters, one computes the expected number of examinees at each quadrature point and the expected proportion of examinees at each quadrature point correctly answering this item. The M step is iterated. In the M step, using the known θ improves the estimation of item parameters. Using Newton-Raphson iterations in the final estimation not only can speed-up the nearly converged EM solution but also can obtain standard errors of item parameters which are not provided by the EM algorithm. The EM cycles are continued until the criterion function becomes stable.

Person Parameter Estimation

Individual abilities, which are also called person parameters, can be obtained using two major methods: maximum likelihood (MLE) (Birnbaum, 1968) or Bayesian methods. Maximum a posteriori (MAP) (Samejima, 1969) and expected a posteriori (EAP) (Bock & Aitken, 1981; Bock & Mislevy, 1982) are included in Bayesian methods. The person parameter estimation here focuses on the estimation of person parameters with item parameters known. Reckase (2009), de Ayala (2009), and Yao (2013) illustrate the general ideas about the person parameters using MLE, MAP, and EAP estimation in the MIRT framework. The following sections are MLE, MAP for the person parameters under the multidimensional IRT framework from Reckase's (2009) textbook, and EAP estimations from de Ayala's (2009) work.

Maximum Likelihood Estimation (MLE). Maximum likelihood estimation estimates person parameters (θ-vector) using the observed string of item scores. Under the local independence assumption, the likelihood of the responses parameters is the product of all probabilities of response to items (products of all item response functions). It can be expressed as Equation 2.12. Estimation programs use the maximum of the log of the likelihood rather than the likelihood. The person location of maximum log likelihood conditional on the trait vector θ in *p*-dimensions in Equation 2.12 can be obtained by setting to 0 the first derivative of the log likelihood function in Equation 2.17.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_2} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}) \end{bmatrix} = \mathbf{0} \tag{2.17}$$

There are two ways to find the maximum of the log-likelihood function. One is the empirical MLE and the other is the Newton-Raphson procedure. However, the empirical MLE can't provide the standard error of estimate while the Newton-Raphson procedure can. The iterative procedure by the Newton-Raphson is shown in Equation 2.18.

$$\hat{\boldsymbol{\theta}}^{t+1} = \hat{\boldsymbol{\theta}}^{t} - \frac{f(\hat{\boldsymbol{\theta}}^{t})}{f'(\hat{\boldsymbol{\theta}}^{t})} \tag{2.18}$$

where the $\hat{\boldsymbol{\theta}}^{t}$ is the estimate of $\boldsymbol{\theta}$ for the t^{th} iteration and $\hat{\boldsymbol{\theta}}^{t+1}$ is the updated estimate of $\boldsymbol{\theta}$. $\hat{\boldsymbol{\theta}}^{t+1} = \hat{\boldsymbol{\theta}}^{t}$ means the maximum value is obtained so the iteration can be stopped. $f(\hat{\boldsymbol{\theta}}^{t})$ is the first derivative of the log-likelihood function in Equation 2.19.

$$f(\hat{\boldsymbol{\theta}}^{t}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}^{t}) \tag{2.19}$$

The standard error of measurement (SEM) can be obtained by setting to 0 the second derivative of the log-likelihood function (the Hessian matrix) in Equation 2.20.

$$f'(\hat{\boldsymbol{\theta}}^t) = \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}^t) =$$

$$\begin{bmatrix} \frac{\partial^2}{\partial^2 \theta_1} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}^t) & \frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}^t) & \cdots & \frac{\partial^2}{\partial \theta_1 \partial \theta_p} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}^t) \\ & \frac{\partial^2}{\partial^2 \theta_2} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}^t) & \cdots & \frac{\partial^2}{\partial \theta_2 \partial \theta_p} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}^t) \\ & & \ddots & \vdots \\ & & & \frac{\partial^2}{\partial^2 \theta_p} \ln L(\boldsymbol{x} \mid \boldsymbol{\theta}^t) \end{bmatrix} \tag{2.20}$$

The disadvantages of MLE for the UIRT framework can be expected under the MIRT framework. The MLE for estimating person parameters produces finite estimates under the condition of typical test lengths with small numbers of dimensions. The MLE yields infinite estimates for person parameters when the items are too easy or difficult. Additionally, when there are not enough items, it is hard to differentiate the locations of person abilities. The solution for infinite estimations in MLE is to assign a θ score to individuals with extreme raw scores in practice. To avoid infinite θ estimates, Bayesian methods can be employed. Comparing the MLE to the Bayesian method, the Bayesian method not only depends on the likelihood function, like MLE, but also depends on a prior probability distribution. This is because of Bayes principle, where the posterior probability distribution is equal to the product of the likelihood function and the prior probability distribution. The prior probability distribution can be from previous research or assumptions. If there is no empirical information about $\boldsymbol{\theta}$, the prior probability distribution is often set to the standard multivariate normal distribution with 0 for means and an identity matrix for the variance and covariance matrix, MVN $(0, \Phi)$.

Bayesian Methods. The Bayesian estimate originally is the probabilities of discrete parameters from Thomas Bayes' paper (read 1763, published 1764). However, because θ parameters in item response theory are continuous random variables, the formula using Bayes' theorem for estimating θ parameters can be expressed as

$$f(\theta \mid \boldsymbol{x}) = \frac{L(\boldsymbol{x} \mid \theta) f(\theta)}{f(\boldsymbol{x})} = \frac{L(\boldsymbol{x} \mid \theta) f(\theta)}{\int_{\theta} L(\boldsymbol{x} \mid \theta) f(\theta) d\theta} \tag{2.21}$$

where $f(\theta)$ is the prior probability function for θ, $\boldsymbol{x}$ is an observed response pattern for examinee j, $L(\boldsymbol{x} \mid \theta)$ is the likelihood function in Equation 2.12, and $f(\theta \mid \boldsymbol{x})$ is the posterior probability function. Because the denominator for Equation 2.21 is the same for all values of θ, Equation 2.21 can be rewritten as Equation 2.22.

$$f(\theta \mid \boldsymbol{x}) \propto L(\boldsymbol{x} \mid \theta) f(\theta) \tag{2.22}$$

The mean or mode of the posterior probability function is often used to estimate the person parameters. The maximum a posteriori (MAP) uses the mode of the posterior probability function and the expected a posteriori (EAP) uses the mean of the posterior probability function. There are advantages and disadvantages to MAP or EAP.

The following statement describes the differences between MAP and EAP which are the same for UIRT and MIRT. First, MAP is an iterative method like MLE whereas EAP is non-iterative and based on numerical quadratic methods like MMLE. Essentially, the item parameters are used to compute EAP. EAP has the advantage of being computationally faster over MAP or MLE because EAP method doesn't need to take the first and second partial derivatives of the likelihood function. Second, MAP uses a continuous prior distribution while EAP uses a discrete prior distribution. Third, the

MAP estimation is more regressed toward the mean of the prior than EAP. Fourth, EAP has the lowest mean square error as compared to MLE or MAP when the distribution of the ability is as the prior distribution (Bock & Mislevy, 1982; Embretson & Reise, 2000; De Ayala, 2009).

Maximum a posteriori (MAP). In the Bayesian methods, the prior population ability distribution multiplied by the likelihood function is the posterior distribution. In other words, MLE method is to maximize the likelihood function while MAP method is to maximize the posterior distribution, which is the likelihood function multiplied by the prior distribution. According to Bayes theorem in Equation 2.22, the posterior density function can be represented as $f(\boldsymbol{\theta} \mid \boldsymbol{x})$. To obtain the person parameter, MAP maximizes the log of the posterior density function. That is, it can be obtained by setting to 0 f the first derivative of the log posterior density function in Equation 2.23. The standard error of measurement (SEM) can be obtained by setting to 0 the second derivative of the log posterior density function.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f(\boldsymbol{\theta} \mid \boldsymbol{x}) = \begin{bmatrix} \frac{\partial}{\partial \theta_1} \ln f(\boldsymbol{\theta} \mid \boldsymbol{x}) \\ \frac{\partial}{\partial \theta_2} \ln f(\boldsymbol{\theta} \mid \boldsymbol{x}) \\ \vdots \\ \frac{\partial}{\partial \theta_p} \ln f(\boldsymbol{\theta} \mid \boldsymbol{x}) \end{bmatrix} = \mathbf{0} \tag{2.23}$$

Expected a posteriori (EAP). De Ayala's (2009) textbook describes in detail the formal mathematical equations for the EAP statistical estimations under the MIRT framework. The EAP estimation for person *j* on dimension 1 (θ_1), which is the person score for dimension 1, can be expressed as follows.

$$\hat{\theta}_{j1} = \sum_{r_1=1}^{R_1} X_{r_1} \left[\sum_{r_2=1}^{R_2} L_i(X_{r_1}, X_{r_2}) A(X_{r_2}) \right] \frac{A(X_{r_1})}{\tilde{p}_j} \tag{2.24}$$

Person *j*'s location on dimension 2 can be shown as below.

$$\hat{\theta}_{j2} = \sum_{r_2=1}^{R_2} X_{r_2} \left[\sum_{r_1=1}^{R_1} L_i(X_{r_1}, X_{r_2}) A(X_{r_1}) \right] \frac{A(X_{r_2})}{\tilde{p}_j} \tag{2.25}$$

where R_1 and R_2 are the number of quadrature points on the first dimension and the second dimension, X_{r_1} is the *rth* quadrature point on dimension 1, the weight of X_{r_1} can be expressed as $A(X_{r_1})$, L_i is the likelihood function for person j in Equation 2.12, and $\tilde{p}_j$ is the unconditional probability of person i's response vector. $\tilde{p}_j$ also can be shown as below using a two-dimensional Guess-Hermite quadrature.

$$\tilde{p}_j = \sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} L_i(X_{r_1}, X_{r_2}) A(X_{r_1}) A(X_{r_2}) \tag{2.26}$$

The posterior standard error of $\hat{\theta}_{j1}$ can be computed as follows.

$$\text{PSD}(\hat{\theta}_{j1}) = \sqrt{\frac{\sum_{r_1=1}^{R_1} X_{r_1} \left[\sum_{r_2=1}^{R_2} L_i(X_{r_1}, X_{r_2}) A(X_{r_2}) \right] A(X_{r_1})}{\sum_{r_2=1}^{R_2} \sum_{r_1=1}^{R_1} L_i(X_{r_1}, X_{r_2}) A(X_{r_1}) A(X_{r_2})}} \tag{2.27}$$

The Bi-factor model

The most restrictive assumption for traditional IRT is the unidimensional assumption. Models that relax the unidimensionality assumption, multidimensional IRT (MIRT) models, have been developed to handle multidimensional data. The bi-factor IRT model is a special case in MIRT since it posits that each test reflects just two factors.

Gibbons and Hedeker (1992) adapted Holzinger and Swineford's (1937) f bi-factor analysis to re-parameterize for dichotomous items in the MIRT literature. Soon after, Gibbons, Bock, Hedeker, Weiss, Segawa and Bhaumik (2007) adapted the bi-factor model for polytomous items. Cai, Yang and Hansen (2011) adapted the bi-factor model for multiple-groups and an arbitrary mixing of dichotomous, ordinal, and nominal items.

Holzinger and Swineford (1937) introduced the bi-factor model, which is a classic factor analytic technique. They described the bi-factor pattern as a general factor and a specific factor, which also can be called a group factor. The bi-factor model allows a general factor loaded upon by all variables and group factors loaded upon by some variables.

There are two assumptions for the bi-factor model. First, one general factor and one specific factor underlie each item. Second, group factors should be orthogonal to the other group factors and to the general factor. For example, if there are four items and two group factors

$$\alpha = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \mathbf{0} \\ \alpha_{21} & \alpha_{22} & \mathbf{0} \\ \alpha_{31} & \mathbf{0} & \alpha_{33} \\ \alpha_{41} & \mathbf{0} & \alpha_{43} \end{bmatrix} \tag{2.28}$$

where α_{ik} is the loading of item i (i=1, 2, 3, 4) on latent factor k (k=1, 2, 3). The first column is the general factor loading (Gibbons & Hedeker, 1992).

Full-Information Item Bi-factor Analysis (FIIBFA). The full-information item bi-factor analysis was developed by Gibbons and Hedeker (1992). Gibbons and Hedeker

(1992) noted that in the FIIBFA model each item loads on one general factor and one of *k* group factors. The FIIBFA model is the confirmatory approach to IRT modeling since the FIIBFA model has adopted Holzinger and Swineford's (1937) "bi-factor" model, which is considered a confirmatory factor analysis model by Joreskog (1969). The following FIBFA illustration is summarized from Gibbons and Hedeker (1992) and Seo (2011).

The bi-factor model implies that the *k* dimensional integral is a two-dimensional integral: the general factor is θ_1 and the group factors θ_2, ..., θ_k. To derive the conditional probability of correct response in the full-information bi-factor item factor analysis, the multidimensional version in Equation 2.9 should be revised to obtain the two-dimensional form. The conditional probability of correct response in the full-information bi-factor item factor analysis can be expressed as

$$P(x_{ij}=1\,|\,\theta_{j1},\theta_{jk},\alpha_{j1},\alpha_{jk},r_i)=\frac{1}{\sqrt{2\pi}\sigma_i}\int_{\gamma_i}^{\infty}\exp\left[-\frac{1}{2}\left(\frac{y_i-\alpha_{j1}\theta_{j1}-\alpha_{jk}\theta_{jk}}{\sigma_i}\right)^2\right]dy_i \qquad (2.29)$$

where $\sigma_i=\sqrt{1-\sigma_{j1}^2-\sigma_{jk}^2}$. Equation 2.29 can also be simplified as Equation 2.30.

$$P(x_{ij}=1\,|\,\theta_{j1},\theta_{jk},\alpha_{j1},\alpha_{jk},r_i)=\Phi\left(\frac{y_i-\alpha_{j1}\theta_{j1}-\alpha_{jk}\theta_{jk}}{\sigma_i}\right)=\Phi_i(\theta_{j1},\theta_{jk}) \qquad (2.30)$$

The unconditional likelihood function of a set of response patterns, $\boldsymbol{x}_j$, can be described as

$$P(\boldsymbol{x}_j=1)=\int_{-\infty}^{\infty}\left\{\prod_{k=2}^{k}\int_{-\infty}^{\infty}\left[\prod_{i=1}^{n}\left(\Phi\left[\frac{y_i-\alpha_{j1}\theta_1-\alpha_{jk}\theta_k}{\sigma_i}\right]\right)\right]g(\theta_k)d\theta_k\right\}g(\theta_1)d\theta_1 \qquad (2.31)$$

Since the bi-factor model was adapted from the unrestricted multiple factor model, the general factor and group factor are assumed to be distributed independently. In other

words, there are no inter-factor correlations needed to be estimated. Therefore, Equation 2.31 can be re-expressed as

$$P(\boldsymbol{x}_j = 1) = \int_{-\infty}^{\infty} \left\{ \prod_{k=2}^{k} \int_{-\infty}^{\infty} \left[\prod_{j=1}^{n} \left([\Phi_{jk}(\theta_1, \theta_k)]^{x_{ij}} [1 - \Phi_{jk}(\theta_1, \theta_k)]^{1-x_{ij}} \right) \right] g(\theta_k) d\theta_k \right\} g(\theta_1) d\theta_1 \quad (2.32)$$

After applying Gauss-Hermite quadrature, the marginal likelihood function of Equation 2.32 is approximated by

$$\hat{P}(\boldsymbol{x}_j = 1) \cong \sum_{q1}^{Q} \left\{ \prod_{k=2}^{k} \sum_{qk}^{Q} \left[\prod_{j=1}^{n} \left([\Phi_{jk}(\theta_1, \theta_k)]^{x_{ij}} [1 - \Phi_{jk}(\theta_1, \theta_k)]^{1-x_{ij}} \right) \right] A(X_{qk}) \right\} A(X_{q1}) \quad (2.33)$$

where $\boldsymbol{X}_{\boldsymbol{q}}$ is the quadrature point and $A(\boldsymbol{X}_{\boldsymbol{q}})$ is the quadrature weight.

The method for estimation of item parameters is similar to the FIIFA model. That is, the estimation of item parameters in the FIIBFA model can be obtained by MML using the EM algorithm.

The bi-factor model in the IRT framework (BIRT). A bi-factor extension of the classical three-parameter logistic model which can be revised into MIRT form (Equation 2.5) represents the probability of a correct response for an item *i*

$$P(x_{ij} = 1 | \theta_{io}, \theta_{is}, \mathbf{a_i}, d_i, c_i) = c_i + (1 - c_i) \frac{\exp[1.7(a_{io}\theta_{io} + a_{is}\theta_{is} + d_i)]}{1 + \exp[1.7(a_{io}\theta_{io} + a_{is}\theta_{is} + d_i)]} \quad (2.34)$$

where θ_{io} is the general factor or ability on item *i*, θ_{is} is one of the group latent factors or ability parameters, a_{io} and a_{is} are the discrimination parameters for the general factor and the group factors (*s*=1,…, *k*), and d_i is a scalar parameter related to an overall multidimensional item difficulty. (Cai, Yang & Hansen, 2011).

Several MIRT methods have been proposed to overcome the unidimensional assumption under traditional IRT. However, the bi-factor model has some advantages compared to other MIRT methods since the bi-factor model simplifies the likelihood function.

The following statement summarizes the advantages of applying the bi-factor model rather than other MIRT models. First, there is the relatively simple computation for parameter estimation. No matter how many factors in the data, each item only loads on two factors under the bi-factor IRT model. That is, only a series of two-dimensional integrals are involved in the estimation. BIRT also can be used for dimension reduction so that this model is much more efficient for estimation of parameters. Besides, the bi-factor model doesn't need to estimate inter-factor correlations. Second, BIRT allows a large number of group factors. Third, BIRT allows conditional dependence among identified subsets of items so BIRT is a useful model for assessment with subtests. Finally, the bi-factor model provides parsimonious estimation for item factor solutions compared to the unrestricted full-information item factor analysis (e.g. Gibbons & Hedeker, 1992; Li & Lissitz, 2012; Cai, Yang & Hansen, 2011).

Studies of Application. The following paragraphs describe studies that applied the bi-factor model, especially for health outcomes assessments.

Gibbons and Hedeker (1992) compared the bi-factor model, the simple structure model and the unrestricted factor model. The first study showed that the bi-factor model fitted better than the simple structure model ($x^2_{20} = 336$, $p<0.0001$) which means there is a primary ability dimension with the ACT natural science test. The second study used the

Hamilton Depression Rating Scale (HDRS) data to compare the bi-factor model, the unrestricted factor model, and the simple structure model. That the bi-factor fitted better than the simple structure model ($x^2_{17} = \mathbf{75}$, p<0.0001) means there was a general depressive dimension needed in the model. The unrestricted factor model had a better fit than the bi-factor model ($x^2_{41} = \mathbf{111}$, p<0.0001). It indicated that the specific factors are not independent and an item did not load on just one specific factor as the bi-factor model assumes.

Gustafsson and Balke (1993) applied the bi-factor model, which was called the nested factor model in this paper, to a battery of sixteen aptitude tests in the 6th grade and three standardized achievement tests in the 9th grade (N=866). Nine specific factors were identified in the model. The factor loadings of the general factor and the specific factors and decompositions of variance for the general factor and specific factors in each observed variable were discussed.

Chen, West, and Sousa (2006) conducted a comparison study between the bi-factor and second- order models using a health-related quality of life measurement dataset. This paper mainly focused on the comparison of the two models whereas it didn't address the issues of interpretation and estimation of item or person parameters. However, there are three main findings suggesting that the bi-factor model has some advantages over a second- order model. First, the bi-factor model identified three group factors, rather than the hypothesized four. Second, the bi-factor model fitted better than the second-order model. Third, when the specific factors predicted an external criterion over and above the general factor, it was easy to interpret the result from the bi-factor model.

DeMars (2006) compared the bi-factor model, the testlet-effects model, and the independent-items model using the *Programme for International Student Assessment* 2000 (PISA 2000) of math and reading tests. The local independence assumption is violated because the data has a testlet structure. In addition, the results of DIMTEST showed the data are not unidimensional. Comparing models by the likelihood-ratio test statistic, the results showed the bi-factor model fitted better than other models in both math and reading tests. The average difference and RMSD among the three models for the difficulty and discrimination parameters of the general factor was discussed. However, this study only focused on the general factor of person parameters. The person parameters estimated by EAP were used to compute reliability, correlation, and RMSD. The results showed that there is no difference between trait estimates because the scales of the estimates have posterior distributions with mean of zero and SD of one.

Reise, Morizot, and Hay (2007) studied whether the bi-factor model can provide information about dimensionality assessment, handle violations of local independence due to item clustering, and provide a method for scaling individual differences. This paper fitted the unidimensional IRT model, multidimensional IRT model, and bi-factor model to compare the factor loadings among the three IRT models on a healthcare systems survey. The result showed the bi-factor model fitted best among the three models. Also, when the multidimensional data are forced to fit the uni-dimensional IRT model, the factor loadings could be biased since the assumption of local dependencies is grossly violated by the group factors. And, the other critical finding is that the multidimensional IRT and bi-factor models have similar fit to the data but the bi-factor

model provides information about dimensional assessment, which is item variance partitioning of the general factor and specific factors.

Simms, Grös, Watson and O'Hara (2008) applied the bi-factor model on the inventory of depression and anxiety symptoms for three populations: community adults, psychiatric patients and undergraduates. This study compared the bi-factor model and one-factor model. The results showed the bi-factor model fitted better. First, the relative sizes using the variance accounted for by the general and specific factors and magnitudes of the general and specific factor loadings were discussed. Second, the intercept parameters of the general factor were computed to represent "levels of symptom severity". Third, they also discussed that the factor scores or person parameters were computed for the general factor using the EAP method. The results of the general factor scores showed that the patient samples have relatively lower scores than the other samples. However, this paper stated the limitation that commercial software at that time was not available to compute the person parameters for the specific factors. Since the data suggested significant variation of the specific factors remaining after accounting for the general factor, it is critical to explore the utility of specific factor scores.

Immekus and Imbrie (2008) is about dimensionality assessment using the full-information item bi-factor model for graded response data. This study compared the full-information item bi-factor model and Samejima's unidimensional graded response model to test dimensionality of the State Metacognitive Inventory. Two separate cohorts (Cohort 1 and Cohort 2) were used. Although the chi-square test showed the bi-factor model fitted better than the unidimensional model, item factor loadings were not much different than the estimates using the unidimensional model. Also, only a few items had

substantial specific factor loadings in the bi-factor model. Estimates of person scores using EAP methods for the unidimensional model and the general factor of the bi-factor model yielded scores with correlations of 0.99 for the two sets of data. Their results showed this inventory is unidimensional so the total scores should be reported, not the subscores.

Gibbons, Rush and Immekus (2009) compared the unidimensional IRT model, the simple structure model with 15 uncorrelated latent traits, the bi-factor model, and the models with 6, 10, and 15 sub-domain alternative conceptualization of the scale on the psychiatric diagnostic screening questionnaire. The sample included 3791 individuals with major depressive disorder. The chi-square test showed the bi-factor model fitted better than the unidimensional IRT model and indicated a multidimensional structure for this dataset. The bi-factor model also resulted in statistically significantly better fit over simple structure and indicated that the structure of this data had one general latent trait and some specific latent traits. In the bi-factor model, the majority of the specific factor loadings are higher than the general factor loading. However, only the thresholds of item parameters were interpreted as the relationship between the levels of mental illness with the person parameters wasn't discussed in this study.

Rijmen (2010) showed the restricted testlet model and the second-order model are equivalent and both models are constrained bi-factor models. This paper also compared the unidimensional 2PL model, the bi-factor, and the second-order model. The data used in this paper were the testlet-based international English assessment test with five reading comprehension items for each of four testlets. The result showed the bi-factor model was the best model for this data according to the AIC, BIC, and the likelihood-ratio test

statistic. Only the precision of the item parameters for the unidimensional 2PL model and bi-factor model were discussed in this study whereas this study didn't mention the person parameters.

Brandt (2008) proposed a Rasch subdimension model, which is a special case of the multidimensional random coefficients multinomial logit model (MRCML). A Rasch subdimension model extends the standard Rasch model (Rasch, 1980) and adds parameters for subdimensions. In this model, only the difficulty parameters for the main dimension and subdimensions are estimated. That is, this model contains the general ability estimated from the main dimension and specific abilities estimated from subdimensions. There are three assumptions in this model. First, the person parameters for subdimensions sum to 0 for each person. Second, covariances between all specific dimenisons and the main dimension equal 0, which assumption is the same as the bi-factor model. Third, the item parameters have mean zero. However, the second assumption about the covariances should be adjusted. Due to the first assumption, th eaverage covariances between all specific dimenisons should be negative, not equal to 0.

Studies of Estimation Algorithms. Cai (2010a) applied a Metropolis–Hastings Robbins–Monro (MH-RM) algorithm in exploratory item factor analysis. Soon after, Cai (2010b) applied the MH-RM algorithm in confirmatory item factor analysis. The MH-RM algorithm can be beneficial for large-scale, multidimensional analysis with many items, factors and examinees based on assessments with mixed item formats and missing responses as in Computerized Adaptive Testing (CAT). The most important thing is that the MH-RM algorithm is stable and efficient in practical applications compared to the Monte Carlo expectation-maximization (MCEM) and Markov chain Monte Carlo

(MCMC) algorithma (Cai, 2010b). Cai (2010c) developed a two-tier full-information item factor analysis model with an EM algorithm for full-information maximum marginal likelihood estimation. Multidimensional item response theory models, bifactor models, and testlet response models are special cases of the two-tier full-information item factor analysis model.

Simulation Studies. DeMars (2006) discussed how the testlet data structure fitted the bi-factor model, testlet-effects model, the polytomous model, and the independent-items model. Six tests with test lengths of 25 and 50 items were generated by the bi-factor model, testlet-effects model, and unidimensional model. Five items formed a testlet with five magnitudes of testlet effects. For the three models, the item slopes of the general dimension ranged from 0.6 to 1.4 and item difficulties ranged from -1.5 to 1.5. The guessing parameter was 0.2. However, the slopes of specific dimensions are different. In the bi-factor model, the slopes of specific dimensions, which were independent of the general dimension slope, were set to 0, 0.3, 0.6, 0.9, and 1.2 for each testlet. In the testlet model, the slopes of specific dimensions, which were proportional to the general dimension slope, were set to slopes of the general dimension times 0, 0.3, 0.6, 0.9, and 1.2. The sample size was 2,000 with 100 replications (50 replications for the testlets model). The general factor and 10 testlet traits were independent and drawn from standard normal distributions. This paper only focused on the general dimension score. The person parameters were estimated by the EAP scoring method for the bi-factor and independent-item model in TESTFACT, for the testlet-effects model in WinBUGS, and for the polytomous model in PARSCALE. RMSE, bias, and reliability were reported to check accuracy of trait estimates. The results showed that reliability was overestimated,

root mean square error (RMSE) for item difficulty was higher, and underestimated the item slopes when items were not independent within testlets fitting the independent-items model. Also, as the items within testlets were generated to be independent, the bi-factor model yielded higher RMSE in difficulty and slope parameters.

Li and Rupp (2011) examined the type I error rate and power of the multivariate extension of $s-x^2$ statistic using the bi-factor model. The $s-x^2$ statistic is the item fit statistic proposed by Orlando and Thissen (2000). Data were generated using either simple-structure MIRT or full-information bifactor models, and then UIRT, MIRT, and full-information bifactor models were fit to the generated data. The simulation conditions for the bi-factor model were test length (40, 80), sample size (1000, 4000), difficulty parameters generated from $N(0,1)$, discrimination parameters or factor slopes (3 levels), and 3 levels of latent trait correlations (0, 0.4, 08) with 100 replications. The main result was that nominal Type I error rates of $s-x^2$ statistics for full-information bifactor models were near the nominal rates for most conditions and were not influenced by test length, sample size, or loading structures.

Fukuhara and Kamata (2011) proposed a 2PL DIF model which is an adaptive bi-factor MIRT for testlet-based data. A simulation study was conducted to examine the proposed model and the traditional IRT DIF model. Four simulation factors were magnitude of testlet effect (0.5, 1.0, 2.0), magnitude of DIF (0.5, 0.7 in log-odds), magnitude of item discrimination (0.8, 2), and the proportion of a focal group to all examinees (0.25, 0.5). The estimation method was a fully Bayesian model using Winbugs software. The results indicated that the proposed model has lower DIF

magnitude of bias and higher average DIF detection rates than the traditional IRT DIF model.

Jeon, Rijmen and Rabe-Hesketh (2012) developed an extended multiple-group bi-factor model for DIF. This model relaxed the traditional assumption that all the dimensions are independent to the assumption that the specific dimensions are conditionally independent given the general dimension. A simulation study was conducted to examine the performance of the proposed model. Two simulation factors were specification of the latent variable distributions (5 levels) and DIF sizes (0.2 and 0.5). The result showed that ignoring the correlation structure of the latent traits can bias item parameter estimates and result in poor DIF estimation.

Summary

In the IRT framework, the ideal situation is to have data with unidimensional structures. However, some assessments are multidimensional. For example, intelligence assessments can be measured as a undimensional construct or hierarchical construct that contains a common component of general intelligence and specific cognitive abilities of verbal ability, reasoning, and quantitative ability (Weiss & Gibbons, 2007). For the past several years, many researchers have developed different multidimensional IRT techniques because the strong assumption of unidimensionality in UIRT does not always hold.

In order to handle data with the multidimensional structures, previous researchers (e.g., Gibbons & Hedeker, 1992) have developed bi-factor models which provide a general score for examinees' overall abilities in a domain, as well as specific factor

scores corresponding to subdomains. There are several benefits to implementing bi-factor models compared to other MIRT models, such as simple computation and ease of interpretation.

For the past ten years, many studies have applied bi-factor models in different fields. However, a majority of the research applying the bi-factor model has focused on model comparison (e.g., Gibbons & Hedeker, 1992; Gustafsson & Balke, 1993; DeMars, 2006; Rijmen, 2010), advantages of applying the bi-factor model (e.g., Chen, West, & Sousa, 2006; Reise, Ventura, et al, 2011), the assessment of dimensionality (e.g., Reise, Morizot & Hay, 2007; Immekus & Imbrie, 2008;), estimation and interpretation of the item parameters (e.g., Gustafsson & Balke, 1993; DeMars, 2006; Simms, Grös, Watson & O'Hara, 2008; Gibbons, Rush & Immekus, 2009;), or estimation algorithm of the bi-factor model (Cai, 2010a; Cai, 2010b) whereas person parameters have rarely been emphasized, especially the interpretation of the subscores.

There are two major reasons to develop the restricted bi-factor model. First, in measurement, scores corresponding to specific factors are rarely, if ever, used. Nevertheless, there is an increasing interest in subscores because subscores could provide more detailed and diagnostic information than a total score. For instance, for the purpose of placement or admission decisions, academic institutions often want a profile of performance for their students to know their strengths and weaknesses in different content areas to better evaluate their training and focus on areas that need instructional improvement (Haladyna & Kramer, 2004; Haberman, 2008). Second, the bi-factor scores which were produced after fitting IRT models are usually hard to interpret and understand, especially for test-takers. Since bi-factor models provide a way to estimate

the subscores under the latent trait framework, restricted bi-factor models providing meaningful subscores should be developed.

Chapter III: Method

The purpose of this study was to demonstrate the restricted bi-factor model for estimating and interpreting scores including overall score and subscores and to evaluate the accuracy of the restricted bi-factor score estimates. First, real data were used to demonstrate this model. The real data analyses illustrated estimation of scores and the conditional standard errors for the overall score and subscores, as well as the interpretation. Second, a simulation study was designed to check how accurately the person parameters can be estimated.

Simulated item responses were used to evaluate the restricted bi-factor model recovery of the person parameters. This simulation study addressed three main research questions:

1. How are reliability, sensitivity, specificity, and recovery of person parameters, influenced by the level of correlation between dimensions?
2. How are reliability, sensitivity, specificity, and recovery of person parameters, influenced by the number of items in the test?
3. How are reliability, sensitivity, specificity, and recovery of person parameters, influenced by the number of dimensions in the test?

The next section describes the assumptions, the restricted bi-factor model, and interpretation for the restricted bi-factor model. Real data, simulated data, simulation designs, and evaluation criteria are presented in the remaining parts.

Method

The person ability in the restricted bi-factor model consists of an overall performance and performance in subdomains (a profile of scores). First, it assumes that the overall domain consists of J (j=1,…, J) subdomains underlying the items, and that each subdomain corresponds to a dimension in a multidimensional space, $\theta_1, \theta_2, ..., \theta_J$. Person p (p=1, 2,…, P) can be represented by a profile of scores (subscores), $\boldsymbol{\theta}_p = \{\theta_{p1}, \theta_{p2}, ..., \theta_{pJ}\}$. Second, the model assumes that each item loads on one and only one dimension so that items satisfy a simple structure model where each item has a nonzero discrimination parameter along one and only one dimension.

If items are assumed to satisfy the simple structure model, they also satisfy the restricted bi-factor model. The purpose of the restricted bi-factor model is to represent subscores in a meaningful way. That is, subscores can be represented as relative strengths and weaknesses. This model applies the definition of scores from Davison and Davenport (2002) and Davison, Chang, and Davenport (2014). First, an overall performance level for person p is defined as person p's average performance over the J subdomains corresponding to the general dimension:

$$\bar{\theta}_{\boldsymbol{p}} = \frac{1}{J}\sum_{j=1}^{J}\theta_{pj} \tag{3.1}$$

Second, subscores for person p corresponding to the subdomains (specific dimensions) can be represented as deviation scores:

$$\boldsymbol{\theta}_p^* = \left\{\theta_{pj}^* = \theta_{pj} - \bar{\theta}_p\right\} \tag{3.2}$$

The elements of $\boldsymbol{\theta}_p^*$ sum to zero ($\sum_{j=1}^{J}\theta_{pj}^* = 0$). Taking $J = 3$ for an example, the general dimension score can be expressed as follows:

$$\bar{\theta}_p = \frac{1}{J}\theta_{p1} + \frac{1}{J}\theta_{p2} + \frac{1}{J}\theta_{p3};\ J = 3 \tag{3.3}$$

From Equation 3.2 and Equation 3.3, each specific dimension score can be expressed as follows:

$$\theta_{p1}^* = \theta_{p1} - \bar{\theta}_p = \theta_{p1} - \frac{1}{J}\theta_{p1} - \frac{1}{J}\theta_{p2} - \frac{1}{J}\theta_{p3} = \frac{J-1}{J}\theta_{p1} - \frac{1}{J}\theta_{p2} - \frac{1}{J}\theta_{p3};\ J = 3 \tag{3.4}$$

$$\theta_{p2}^* = \theta_{p2} - \bar{\theta}_p = -\frac{1}{J}\theta_{p1} + \frac{J-1}{J}\theta_{p2} - \frac{1}{J}\theta_{p3};\ J = 3$$

$$\theta_{p3}^* = \theta_{p3} - \bar{\theta}_p = -\frac{1}{J}\theta_{p1} - \frac{1}{J}\theta_{p2} + \frac{J-1}{J}\theta_{p3};\ J = 3$$

Combining Equation 3.3 and Equation 3.4, a weight matrix can be constructed:

$$\boldsymbol{W} = \begin{bmatrix} \frac{1}{J} & \frac{1}{J} & \frac{1}{J} \\ \frac{J-1}{J} & -\frac{1}{J} & -\frac{1}{J} \\ -\frac{1}{J} & \frac{J-1}{J} & -\frac{1}{J} \\ -\frac{1}{J} & -\frac{1}{J} & \frac{J-1}{J} \end{bmatrix};\ J = 3 \tag{3.5}$$

The general form of the weight matrix with *J* dimensions can be expressed as in Equation 3.6.

$$\boldsymbol{W} = \begin{bmatrix} \frac{1}{J} & \frac{1}{J} & \cdots & \frac{1}{J} \\ \frac{J-1}{J} & \frac{-1}{J} & \cdots & \frac{-1}{J} \\ \frac{-1}{J} & \frac{J-1}{J} & \cdots & \frac{-1}{J} \\ \vdots & \vdots & & \vdots \\ \frac{-1}{J} & \frac{-1}{J} & \cdots & \frac{J-1}{J} \end{bmatrix} \tag{3.6}$$

A Restricted bi-factor model. This section describes the two-parameter logistic (2PL) restricted bi-factor model for dichotomous items but the lower asymptote parameter can be added to create a three-parameter logistic (3PL) model. The probability of person *p* answering correctly item *i* from domain *j* based on the ability of θ_j can be expressed as in Equation 3.7.

$$P(x_{ip} = 1 | \bar{\theta}_p, \theta^*_{pj}) = \frac{1}{1 + \exp[-a_{i(j)}\bar{\theta}_p - a_{i(j)}\theta^*_{pj} - c_i]} \tag{3.7}$$

where $a_{i(j)}$ is the discrimination parameter for item *i* from subdomain *j*, c_i is the intercept parameter for item *i*, $\bar{\theta}_p$ is person *p*'s overall ability along the general dimension and θ^*_{pj} is person *p*'s ability on specific dimension *j*. In Equation 3.7, the model assumes that item responses depend on a general dimension ($\bar{\theta}_p$) and the specific dimension (θ^*_{pj}) and that the discrimination parameters for an item ($a_{i(j)}$) is equal for the general dimension and the specific dimension. Since the discrimination parameter is equal for the general dimension and specific dimension, Equation 3.7 can be rewritten in a simple structure

form in which each item has a nonzero discrimination parameter along one and only one dimension, ($a_{i(j)}$) on item *i* and subdomain j :

$$P(x_{ip}=1|\theta_{pj})=\frac{1}{1+\exp[-a_{i(j)}\theta_{pj}-c_i]} \tag{3.8}$$

where the difficulty parameter is $b_{i(j)}=-\frac{c_i}{a_{i(j)}}$ and $\theta_{pj}=\bar{\theta}_p+\theta^*_{pj}$. Items are assumed to satisfy the simple structure model in order to fit the restricted bi-factor model. To obtain the general dimension score and the specific dimension scores of the restricted bi-factor model, one can first compute the person parameter (EAP or MAP) using the simple structure model. Then, using the scoring method described in the person parameter section below to obtain the general dimension score and the specific dimension scores of the restricted bi-factor model along with the conditional standard errors.

Estimating the General and Specific Dimensions Scores. The $(J+1)\times P$ matrix of general and specific dimensions scores for sample size= *P* can be obtained by multiplying the $(J+1)\times J$ weight matrix in Equation 3.6 and the $J\times P$ matrix of person parameter from the simple structure model.

Estimating the Variances and Covariances among the Dimensions. The estimated variances and covariances among the dimensions can be obtained from the person parameters computed for the simple structure model, which is $J\times J$ matrix of the error variances and covariances among dimensions for person *p* ($\mathbf{cov}_P$). Standard multidimensional IRT software (e.g. IRTPRO and FLEXMIRT) provides person

parameter estimation with error variances and covariances using methods, such as maximum likelihood (ML), expected a posterior (EAP), or maximum a posterior (MAP).

Estimating the Conditional Error Variances and Covariances among Dimensions. The matrix theorem in Equation 3.9 can be used to compute the conditional error variances and covariances among dimensions.

$$\text{cov}(\boldsymbol{AX}+a,\boldsymbol{BY}+b)=\boldsymbol{A}\,\text{cov}(\boldsymbol{X},\boldsymbol{Y})\boldsymbol{B}' \tag{3.9}$$

According to Equation 3.9, the conditional error variances and covariances among dimensions for person *p* can be obtained:

$$\mathbf{cov}_P^*=\boldsymbol{W}\mathbf{cov}_P\boldsymbol{W}' \tag{3.10}$$

where $\boldsymbol{W}$ is the $(J+1)\times J$ weight matrix in Equation 3.6. The diagonal of $\mathbf{cov}_P^*$ contains the conditional error variances on general and specific dimension scores. To obtain the conditional standard error of $\bar{\theta}_p$ and $\hat{\theta}_{pj}^*$, one computes the square root of the diagonal matrix of $\mathbf{cov}_P^*$.

The restricted bi-factor model has more restrictive assumptions than the regular bi-factor model. First, items are assumed to satisfy the simple structure model. Second, it is assumed that the discrimination parameter is equal for the general dimension and specific dimensions. However, the restricted bi-factor model is more practical than the regular bi-factor because it allows correlations between the specific dimensions.

Interpretation of the General and Specific Dimension Scores. The restricted bi-factor model provides person scores with conditional standard errors on the overall

performance $\bar{\theta}_p$ and subscores $\hat{\theta}_p^*$. Positive values on the specific dimension score ($\hat{\theta}_p^*$) represents relative strengths, performance in subdomain *j* higher than the overall performance. Negative values on the specific dimension score suggest a relative weakness, performance in subdomain *j* lower than the overall performance.

The Interpretation of Profile Pattern. The profile pattern can be defined as a vector of specific dimension scores for person *p*, θ_p^*. Depending on the application of assessments in various fields, users may interpret profile patterns differently. First, some fields focus on the relatively high scores of the profile pattern. For interest inventories, the highest score in a person's profile pattern can be called a dominant interest and could suggest further career exploration in the specific career cluster. For clinical assessment, the highest score in the profile pattern can be called a dominant behavioral tendency and could suggest the need of intervention. Second, some fields focus on the relatively low scores of the profile pattern. Take an English language assessment for example; a relative weakness can provide teaching guidance for teachers or parents. Third, the relatively high and low scores of the profile pattern can be emphasized. If an achievement assessment consists of math, science and reading, the relatively high scores could suggest career interests and the relatively low scores could suggest a lack of preparation or require a remedy for that domain.

The Interpretation of Overall Performance Score and Profile Pattern in Combination. The overall performance and profile pattern may also be used in combination. Achievement assessments commonly use both scores. Take the English language development assessment for example; the overall performance can be used to

set a standard for ELS students. If ELS students have an overall score lower than the proficiency level, they may need to take some ELS classes to improve their English ability. In this case, the relatively low scores in the student's profile pattern can be used as teaching guides to identify the domains (reading, writing, listening or speaking) in which the teacher can provide more instruction. For clinical assessment, the overall score may be used to identify if examinees need intervention. The relatively high scores, in the profile pattern can be used as guides for treatments in the intervention.

Real Data Study

Participants. The sample for this study comes from 1678 3rd graders (797 females and 876 males) enrolled in English-as-second-language (ESL) classes in a southern state.

Measures. The assessment used in this study is The *English Language Development Assessment* (*ELDA*), which is a criterion-referenced English language proficiency assessment with language domains of Listening, Speaking, Reading, and Writing. It is administered to students identified as limited English proficient (LEP) to measure the language skills and proficiency for both written and spoken English (*Interpretive Guide*, 2013). A proficiency level is scaled from 1 to 5 for each domain and an overall composite score. In the speaking domain, twelve items have a point value of 0-2. Because the restricted bi-factor model described above is a model for dichotomous items and less than ten percent of examinees scored 0 in the speaking items, those items were rescored with a point value of 0-1 (recode 0 as 0, 1 as 0, and 2 as 1). In the writing domain, there are eleven multiple-choice items and four constructed response items, three

items with a point value of 0-3 and one item with a point value of 0-4. In this study, the four constructed response items were dropped. Table 3.1 shows the number of items for each domain.

Table 3.1: Number of Items for Each Domain

Domains	Number of Items
Reading	35
Listening	35
Speaking	12
Writing	11

Simulation Study

A simulation study was conducted to evaluate how accurately person parameters and their conditional standard errors can be estimated under several conditions with common characteristics of test features.

Simulation Conditions. Brandt (2008) proposed a Rasch subdimension modelin which each person has a general ability measured by the general dimension and strengths and weaknesses measured by the specific dimensions. However, the major difference between a Rasch subdimension model and the restricted bi-factor model is the assumption about the covariance between the specific dimensions. A Rasch subdimension model sets the covariance between the specific dimensions to 0. Due to the Rasch subdimension model, this study includes the pseudo Rasch generating model as one of the conditions.

There are five simulation conditions in this study: generating model, test length within each subdimension, number of subdimensions, correlations between

subdimensions, and ability estimation methods. Two levels for the condition of generating model are pseudo Rasch and 2PL simple structure model.

For the generating model of the pseudo Rasch simple structure model, the simulation conditions are test length within subdimensions and ability estimation methods. Each simulation condition consists of two levels. First, number of subdimensions was fixed as 3 and correlation between subdimensions was fixed as .5. Second, the test length within each subdimension assumes that number of items is the same within each subdimension. Two levels in this condition are 15 items within each subdimension and 30 items within each subdimension. Third, the two levels of ability estimation methods are maximum a posteriori (MAP) method and expected a posteriori (EAP) method.

For the generating model of 2PL simple structure model, two levels within four simulation conditions are described below. First, the test length within each subdimension assumes that the number of items is the same within each subdimension. The two levels in this factor are 15 items within each subdimension and 30 items within each subdimension. Second, two levels of number of subdimensions are 3 dimensions and 4 dimensions. Third, two levels of ability estimation methods are maximum a posteriori (MAP) method and expected a posteriori (EAP) method. Forth, the correlation between subdimensions assumes that correlations between subdimensions are the same within a condition and two levels of correlations between subdimensions are .3 (representing small to medium correlation) and .6 (representing high correlation).

The result of increasing correlation between subdimensions can be expected as follows. As correlation between subdimensions increases with fixing other conditions, the conditional error variances of the subdimensions decreases and the conditional error variance of the general dimension increases. The proof of this is in Appendix A. Table 3.2 shows the design, which has a total of 20 crossed conditions.

Table 3.2: The Simulation Conditions

Generating Model	Test length	Number of Dimensions	Correlations	Ability Estimation	
				EAP	MAP
Rasch Simple Structure Model	15	3	.5	Cell 1	Cell 11
	30			Cell 2	Cell 12
2-PL Simple Structure Model	15	3	.3	Cell 3	Cell 13
			.6	Cell 4	Cell 14
		4	.3	Cell 5	Cell 15
			.6	Cell 6	Cell 16
	30	3	.3	Cell 7	Cell 17
			.6	Cell 8	Cell 18
		4	.3	Cell 9	Cell 19
			.6	Cell 10	Cell 20

Simulation Procedures. De la Torre and Song (2009) stated that sample size had no impact on the quality of the multidimensional IRT overall ability estimates. The conclusion was based on a simulation study with sample sizes of 1000, 2000, and 4000. Therefore, in this study, the sample was fixed at N=1000 for all of simulation conditions.

Data Generation. One assumption for the restricted bi-factor model is that items should satisfy a simple structure model. That is, item response patterns were simulated according to the simple structure model and dichotomous items using the pseudo Rasch and two-parameter, multidimensional, compensatory logistic model. The pseudo Rasch model means each discrimination parameter is different due to randomly drawing from a small range of a uniform distribution. The pseudo Rasch model was used because it is

more practical to generate the discrimination parameters randomly from a small range of a uniform distribution than to set all items with the same discrimination parameter as the Rasch model.

Each item only has one non-zero discrimination parameter and one difficulty parameter for the pseudo Rasch and the two-parameter simple structure model. Data generation for the item parameters needs to be specified in advance. The discrimination (slope) parameter of each item was generated randomly from a uniform distribution ranging from 0.2 to 2.2, $a_{i(j)} \sim U[0.2, 2.2]$ for the 2PL model and from a uniform distribution ranging from 0.7 to 1.2, $a_{i(j)} \sim U[0.7, 1.2]$ for the pseudo Rasch model. The difficulty parameter generated followed the standard normal distribution, $b_{i(j)} \sim N(0,1)$ (Rupp & Li, 2011; Zheng, 2013). The MIRT scalar parameter c_i can be computed according to Equation 3.8, $c_i = -a_{i(j)} b_{i(j)}$. The ability parameter followed a multivariate normal distribution, $\boldsymbol{\theta} \sim MVN(0, \sum)$. Take cell 3 for example, in which the correlations between dimensions were 0.3, there were three dimensions, and the mean vector, variance vector, and correlation matrix were $\boldsymbol{u} = \{0,0,0\}; \boldsymbol{\sigma} = \{1,1,1\}; \sum = \begin{bmatrix} 1 & .3 & .3 \\ .3 & 1 & .3 \\ .3 & .3 & 1 \end{bmatrix}$.

This study conducted 100 replications for each cell. The item parameters were fixed for each replication while the person parameters were randomly drawn for each replication.

To ensure the result can be comparable across the conditions, item parameters for the condition of fifteen items within a subdimension were duplicated twice to obtain item parameters for the condition of thirty items.

Simulation Phase. The data simulation for Cell 1 to Cell 10 can be briefly described below. Figure 3.1 summarizes the simulation stages.

1. *Generate item parameters*: generate item parameters in R (R Development Core Team,2010) according to simulation conditions.

2. *Generate response patterns*: fix the item parameters obtained from Step 1, generate response patterns and true person parameters in flexMIRT (Cai, 2012) using the 2 PL

simple structure model with sample size=1,000 for 100 replications using random seeds.

3. *Estimate the person parameters*: import the generated (true) item parameters and estimate the person parameters in flexMIRT using the EAP or MAP method for 100 replications.

4. *Compute the restricted bi-factor model scores*: compute the restricted bi-factor model scores from the estimated person scores (Step 3) in R using the method described in the person parameter section for 100 replications.

5. *Evaluation*: analyze the restricted bi-factor model score for the true and estimated person parameters using evaluation criteria in R.

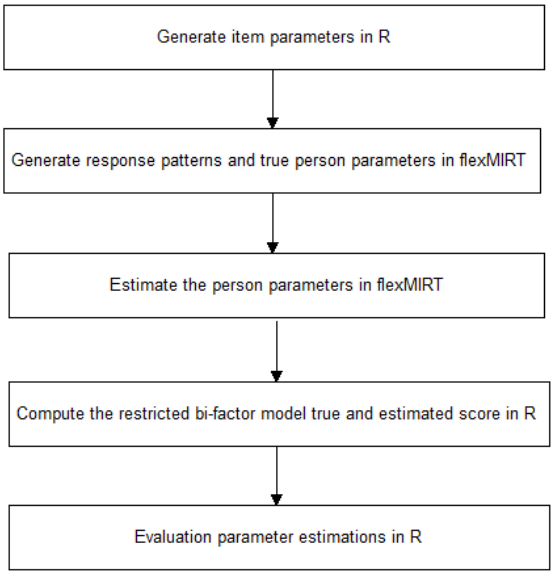


Figure 3.1. Simulation Phase of the Study

Evaluation Criteria. Common criteria to evaluate parameter recovery are bias, root mean squared error (RMSE), standard error (SE), and the average conditional error variance. Bias, RMSE, and SE can be computed from the true parameters and estimated parameters across replications. To show whether the restricted bi-factor model can correctly identify persons' significant strength and weakness, some indices were computed, such as the average reliability index for each dimension, and average sensitivity and specificity. The evaluation criteria were computed for 20 simulation cells.

Bias, RMSE, SE, and the Average Conditional Error Variance. The true theta continuum along each dimension was broken up into 14 intervals to report bias, RMSE, SE, and the average conditional error variance for each interval. In the equations below, θ refers to true (generated) restricted bi-factor person parameters, $\hat{\theta}$ refers to estimated restricted bi-factor person parameters, *R* is the replication, and *N* is the sample size within each interval.

Bias is the deviation of estimated parameters from the true parameters so smaller absolute bias reflects more accurate estimated parameters. Bias and absolute bias were computed for each θ interval. Bias was estimated as:

$$\text{bias} = \frac{\sum_{n=1}^{N}(\hat{\theta}_n - \theta_n)}{N} \tag{3.11}$$

A smaller RMSE indicates more accurate parameter estimates. RMSE is computed as:

$$\text{RMSE} = \sqrt{\frac{\sum_{n=1}^{N}(\hat{\theta}_n - \theta_n)^2}{N}} \tag{3.12}$$

A smaller SE indicates more stable parameter estimates. SE is estimated as:

$$\text{SE} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_n - \frac{\sum_{n=1}^{N}\hat{\theta}_n}{N})^2} \tag{3.13}$$

The average conditional error variance can be computed as:

$$\text{The average of MSE} = \frac{\text{MSE}}{R} \tag{3.14}$$

where MSE is the mean of the conditional error variance for all respondents within one of the *R* replications and the average MSE is taken over all replications in a cell..

Reliability and the z_{pj}. Based on test reliability in CTT, the reliability for the simulation study can be estimated as

$$\text{reliability} = [\text{cor}(\theta, \hat{\theta})]^2 \tag{3.15}$$

The reliability estimation for the real data can be estimated as

$$\text{reliability} = \frac{S^2_{\hat{\theta}}}{S^2_{\hat{\theta}} + MSE} \tag{3.16}$$

where $S^2_{\hat{\theta}}$ is the variance of the MAP or EAP estimates for a dimension.

The average reliability across replications ($\frac{\text{reliability}}{R}$) was computed for each dimension and each cell.

The score of z_{pj} can be estimated as:

$$z_{pj} = \frac{\hat{\theta}_{pj}}{s(\hat{\theta}_{pj} \mid \theta_{pj})} \tag{3.17}$$

where $s(\hat{\theta}_{pj} \mid \theta_{pj})$ is the conditional standard error of $\hat{\theta}_{pj}$.

A significant strength can be defined as z_{pj} above 1 whereas a significant weakness can be defined as z_{pj} below -1. A true strength can be defined as the true specific dimensions scores minus the true general dimension score is above 1 while a true weakness can be deined as the true specific dimension score minus the true general dimension score is below -1. After the significant strength and weakness and true strength and weakness were defined, sensitivity and specificity can used to measure how the restricted bi-factor model identifies the examinees' significant strength or weakness.

Sensitivity is the proportion of people with a significant strength or weakness who are correctly identified as such (Equation 3.18 and 3.19). Specificity is the proportion who do not have a specific strength or weakness along the dimension and who are correctly identified as such (Equation 3.20). The average sensitivity of the strength and weakness and average specificity across the replications for each dimension were computed for 20 cells. Table 3.3 shows the cut-off point for a true strength or weakness and a significant strength or weakness.

$$\text{Sensitivity for the strength} = \frac{\text{\# of people correctly identifying a stength}}{\text{\# of people having a true stength}} \quad (3.18)$$

$$\text{Sensitivity for the weakness} = \frac{\text{\# of people correctly identifying a weakness}}{\text{\# of people having a true weakness}} \quad (3.19)$$

$$\text{Specificity} = \quad (3.20)$$

$$\frac{\text{Number of people without a true strength or weakness, who won't identidy as such}}{\text{Number of people without a true strength or a true weakness}}$$

Table 3.3: Cut-off Points for a True Strength or Weakness and a Significant Strength or Weakness

Label	A True parameter (the True Specific Dimension Score Minus the True General Dimension Score)	An Estimated parameter (z_{pj})
Strength	Above 1	Above 1
Not Either	Between 1 and -1	Between 1 and -1
Weakness	Below -1	Below -1

Chapter IV: Results

This chapter presents the results of the real data study and the simulation study described in Chapter 3. Results present the real data study using the English language development assessment data and second present the simulation study. The discussions of the results for the real data study and the simulation study are presented in Chapter 5.

Results of Real Data Study

The results below are based on the two-parameter and the Rasch, multidimensional, compensatory logistic model for dichotomous items. Each model used the MAP and EAP ability estimation method. In total, there are four person parameters files.

Table 4.1 shows the correlation of EAP and MAP scores for the 2PL and the Rasch restricted bi-factor model. Most of the correlations among the specific dimensions are negative, as expected given a model satisfying the constraint that the sum of the specific dimension scores equals 0. For the 2PL restricted bi-factor model (EAP and MAP), the specific score correlations range from -.74 (Reading and Speaking) to .4 (Reading and Writing). For the Rasch restricted bi-factor model (EAP and MAP), the specific score correlations range from -.62 (Reading and Speaking) to .2 (Reading and Writing).

Table 4.1. Within Model Correlation Matrices: MAP and EAP Person Scores for the Rasch Restricted Bi-factor Model and the 2PL Restricted Bi-factor Model

	EAP for the 2PL Restricted Bi-factor Model				
	General	Reading	Writing	Listening	Speaking
General	1	0.22	0	0.18	-0.21
Reading	0.22	1	0.4	-0.03	-0.74
Writing	0	0.4	1	-0.17	-0.71
Listening	0.18	-0.03	-0.17	1	-0.38
Speaking	-0.21	-0.74	-0.71	-0.38	1
	MAP for the 2PL Restricted Bi-factor Model				
	General	Reading	Writing	Listening	Speaking
General	1	0.25	0.02	0.22	-0.26
Reading	0.25	1	0.38	-0.04	-0.73
Writing	0.02	0.38	1	-0.19	-0.7
Listening	0.22	-0.04	-0.19	1	-0.38
Speaking	-0.26	-0.73	-0.7	-0.38	1
	EAP for the Rasch Restricted Bi-factor Model				
	General	Reading	Writing	Listening	Speaking
General	1	0.3	-0.31	0.08	0
Reading	0.3	1	0.19	-0.12	-0.62
Writing	-0.31	0.19	1	-0.16	-0.7
Listening	0.08	-0.12	-0.16	1	-0.35
Speaking	0	-0.62	-0.7	-0.35	1
	MAP for the Rasch Restricted Bi-factor Model				
	General	Reading	Writing	Listening	Speaking
General	1	0.28	-0.34	0.07	0.04
Reading	0.28	1	0.2	-0.12	-0.62
Writing	-0.34	0.2	1	-0.15	-0.71
Listening	0.07	-0.12	-0.15	1	-0.34
Speaking	0.04	-0.62	-0.71	-0.34	1

Table 4.2 contains two panels for the correlations of the 2PL and the Rasch restricted bi-factor model across a pair of person parameter estimation methods. The first panel is the EAP and the MAP of the 2PL restricted bi-factor model, and the second panel is the EAP and the MAP of the Rasch restricted bi-factor model. Of particular interest in these matrices are the diagonal elements, the correlations between corresponding 2PL and the Rasch restricted bi-factor model person parameter estimation

methods. For both of the 2PL and the Rasch models, all of the diagonal elements are 1. This result indicated that the ability estimation method of the EAP and MAP produced near identical estimated scores.

Table 4.2. Cross Model Correlation Matrices: MAP and EAP Person Scores for the Rasch Restricted Bi-factor Model and the 2PL Restricted Bi-factor Model

	Pair of MAP and EAP for the 2PL Restricted Bi-factor Model					
		MAP				
		General	Reading	Writing	Listening	Speaking
EAP	General	1	0.25	0.01	0.21	-0.25
	Reading	0.23	1	0.39	-0.04	-0.74
	Writing	0.01	0.39	1	-0.18	-0.7
	Listening	0.19	-0.04	-0.18	1	-0.39
	Speaking	-0.22	-0.73	-0.7	-0.37	1
	Pair of MAP and EAP for the Rasch Restricted Bi-factor Model					
		MAP				
		General	Reading	Writing	Listening	Speaking
EAP	General	1	0.28	-0.33	0.07	0.03
	Reading	0.29	1	0.18	-0.12	-0.61
	Writing	-0.31	0.2	1	-0.15	-0.71
	Listening	0.08	-0.13	-0.16	1	-0.34
	Speaking	0.01	-0.63	-0.69	-0.35	1

Table 4.3 shows the reliability estimate for each subscore based on the two models and two estimation methods. Also shown are the average conditional error variances (mean squared errors, MSE) for each dimension and model. Reliabilities were estimated using Equation 3.16 in Chapter 3. Examination of Table 4.3 shows two interesting trends. First, the MAP and the EAP produce the same reliabilities for both models. This result can be expected because Table 4.2 shows the correlations between the MAP and the EAP estimation methods are 1.00 for the 2PL model and the Rasch model. Second, reliabilities in majority of subdimensions for the 2PL model are relatively higher than the Rasch model whereas the MSE for the 2PL model is relatively

lower than for the Rasch model. This result also can be expected because the 2PL model estimated one more item parameter (the discrimination parameter) than the Rasch model.

Table 4.3. The Reliability of MAP and EAP Person Scores for the Rasch Restricted Bi-factor Model and the 2PL Restricted Bi-factor Model

	2PL		The Rasch Model	
Dimension	EAP (MSE)	MAP (MSE)	EAP (MSE)	MAP (MSE)
General	0.93 (0.05)	0.93 (0.05)	0.93 (0.07)	0.93 (0.07)
Reading	0.63 (0.05)	0.62 (0.05)	0.57 (0.08)	0.57 (0.08)
Writing	0.51 (0.11)	0.50 (0.11)	0.55 (0.13)	0.55 (0.13)
Listening	0.51 (0.07)	0.51 (0.07)	0.51 (0.09)	0.5 (0.09)
Speaking	0.73 (0.11)	0.73 (0.11)	0.64 (0.2)	0.63 (0.2)

Table 4.4 shows the number of people identified with a significant strength or weakness in their MAP and EAP specific dimension scores from the 2PL and the Rasch restricted bi-factor model. The score z_{pj} for each person was computed using Equation 3.17. A significant strength can be defined as a z_{pj} score above 1 and a significant weakness can be defined as a z_{pj} score below -1. A significant weakness is a restricted bi-factor score 1 conditional standard error below the person's overall level whereas a significant strength is a restricted bi-factor score 1 conditional standard error above the person's overall level. Take the reading dimension score in Table 4.4 for example. 326 people had a z_{pj} score above 1 for their reading score indicating that they had a significant strength in reading ability relative to their overall English language performance. However, 354 people had a z_{pj} score below -1 for their reading score. This means that 354 people had a significant weakness in reading ability relative to their overall English language performance. 863 people were not identified as having a

significant strength or weakness in the reading dimension. The other three specific dimension scores can be interpreted in the like fashion.

Table 4.4. Number of People Identified with a Significant Strength or Weakness Based on the Ratio of Their Specific Dimension Score to Its Conditional Standard Error for the four Models

EAP for the 2PL Restricted Bi-factor Model				
	Reading	Writing	Listening	Speaking
Weakness	354	280	270	452
None	863	1008	1035	651
Strength	326	255	238	440
MAP for the 2PL Restricted Bi-factor Model				
	Reading	Writing	Listening	Speaking
Weakness	338	261	272	456
None	876	1025	1042	667
Strength	329	257	229	420
EAP for the Rasch Restricted Bi-factor Model				
	Reading	Writing	Listening	Speaking
Weakness	312	284	250	361
None	942	985	1070	821
Strength	289	274	223	361
MAP for the Rasch Restricted Bi-factor Model				
	Reading	Writing	Listening	Speaking
Weakness	307	271	246	367
None	945	986	1076	829
Strength	291	286	221	347

Results of the Simulation Study

Results of the simulation study are presented below. First, correlations between two different numbers of quadrature points are discussed. Second, correlations between two ability estimation methods of MAP and EAP are discussed. Third, comparison between variances of MAP and variance of EAP is discussed. Fourth, variance of the true restricted bi-factor scores are computed and discussed. Fifth, five evaluation criteria of person parameter recovery, including bias, absolute bias, RMSE, SE, and average

conditional error variance (MSE) are presented. Finally, reliability, sensitivity, and specificity of the score, z_{pj}, are discussed.

Correlation between two Numbers of Quadrature Points. The EAP estimation method is based on numerical quadrature methods whereas the MAP estimation is an iterative method. The number of quadrature points was set to 49 points spread from -6 to 6 for all cells, except Cell 9 and Cell 10. However, for Cell 9 and Cell 10, which are the simulation conditions with 4 dimensions, 30 items within each dimension, EAP estimation method, and 0.3 or 0.6 correlation, the number of quadrature points was reduced to 20 points spread from -3 to 3 to reduce computation time. Table 4.5 shows correlations for each dimension and for the SE of each dimension between two numbers of quadrature points for Cell 9. Due to the computation time, there are 4 replications for Cell 9 using 49 quadrature points from -6 to 6 to demonstrate the correction between two sets of quadrature points. For each dimension, the correlations are all 1. For the SE of each dimension, the lowest correlation (r=0.94) is for the SE of the specific dimension 3 (S3). It is concluded that reducing the number of quadrature points does not materially affect the result.

Table 4.5. Correlation for Each Dimension and for the SE of Each Dimension between Two Sets of Quadrature Points

	Correlation between Two Sets of Quadrature Points	
Dimension	For Each Dimension	For the SE of Each Dimension
General (G)	1	0.96
Specific dimension 1 (S1)	1	0.98
Specific dimension 2 (S2)	1	0.93
Specific dimension 3 (S3)	1	0.94
Specific dimension 4 (S4)	1	0.98

Correlation between the MAP and the EAP. To investigate the relationship between the EAP and the MAP, the correlations between the two estimation methods

were computed as the average correlation over 100 replications. Table 4.6 contains correlations for each dimension between the MAP and EAP estimation methods, and Table 4.7 contains correlations for the SE of each dimension. Table 4.6 shows that the correlations are all 1.00 for each dimension between the MAP and EAP ability estimation method. Table 4.7 shows that the lowest correlation for the SE of each dimension is 0.90, which is for the SE of dimension S1 between Cell 3 and Cell 13 and the SE of the S2 dimension between Cell 10 and Cell 20. The result suggests that the two ability estimation methods produce similar estimated scores.

Table 4.6. Correlations for Each Dimension between the EAP and MAP Estimation methods

Cells	G	S1	S2	S3	S4
Cell 1 and Cell 11	1.00	1.00	1.00	1.00	--
Cell 2 and Cell 12	1.00	1.00	1.00	1.00	--
Cell 3 and Cell 13	1.00	1.00	1.00	1.00	--
Cell 4 and Cell 14	1.00	1.00	1.00	1.00	--
Cell 5 and Cell 15	1.00	1.00	1.00	1.00	1.00
Cell 6 and Cell 16	1.00	1.00	1.00	1.00	1.00
Cell 7 and Cell 17	1.00	1.00	1.00	1.00	--
Cell 8 and Cell 18	1.00	1.00	1.00	1.00	--
Cell 9 and Cell 19	1.00	1.00	1.00	1.00	1.00
Cell 10 and Cell 20	1.00	1.00	1.00	1.00	1.00

Table 4.7. Correlations for the SE of Each Dimension for the EAP and MAP Estimation Methods

Cells	SE of G	SE of S1	SE of S2	SE of S3	SE of S4
Cell 1 and Cell 11	1.00	1.00	1.00	1.00	--
Cell 2 and Cell 12	1.00	1.00	1.00	1.00	--
Cell 3 and Cell 13	0.97	0.90	0.98	0.99	--
Cell 4 and Cell 14	1.00	0.99	1.00	0.99	--
Cell 5 and Cell 15	1.00	1.00	0.99	0.99	1.00
Cell 6 and Cell 16	1.00	1.00	0.99	1.00	0.99
Cell 7 and Cell 17	0.99	0.97	0.99	1.00	--
Cell 8 and Cell 18	1.00	1.00	1.00	1.00	--
Cell 9 and Cell 19	0.97	0.98	0.94	0.95	0.97
Cell 10 and Cell 20	0.96	0.97	0.90	0.99	0.94

Comparison between variances of the MAP and the EAP Scores. Another method to compare the EAP and the MAP is to look at the variances of the MAP and the EAP. Variances of the MAP and variance of the EAP were computed as the average variance over 100 replications. Table 4.8 shows the variance of the MAP and variance of the EAP for each dimension. It shows that the MAP has slightly lower variance than the EAP for all the dimensions and all simulation conditions but the difference between the variance of the MAP and the EAP is not large. Therefore, the following paragraphs about the evaluation criteria, reliability, sensitivity, and specificity were summarized only for results of the EAP estimation method.

Table 4.8. Variance of the MAP and Variance of the EAP.

	Ability Estimation				
	EAP				
Cell	G	S1	S2	S3	S4
Cell 1	0.56	0.19	0.19	0.19	--
Cell 2	0.61	0.24	0.24	0.24	--
Cell 3	0.45	0.3	0.33	0.34	--
Cell 4	0.66	0.16	0.17	0.16	--
Cell 5	0.41	0.37	0.38	0.36	0.38
Cell 6	0.64	0.19	0.17	0.19	0.18
Cell 7	0.49	0.36	0.39	0.39	--
Cell 8	0.69	0.2	0.21	0.2	--
Cell 9	0.43	0.43	0.43	0.42	0.44
Cell 10	0.66	0.23	0.22	0.23	0.23
	MAP				
Cell	G	S1	S2	S3	S4
Cell 11	0.53	0.18	0.18	0.18	--
Cell 12	0.58	0.23	0.23	0.23	--
Cell 13	0.42	0.29	0.31	0.31	--
Cell 14	0.6	0.15	0.16	0.16	--
Cell 15	0.38	0.34	0.36	0.34	0.35
Cell 16	0.59	0.18	0.16	0.17	0.18
Cell 17	0.46	0.35	0.37	0.37	--
Cell 18	0.65	0.19	0.19	0.19	--
Cell 19	0.41	0.41	0.42	0.41	0.41
Cell 20	0.63	0.22	0.21	0.21	0.22

Variance of the True Restricted Bi-factor Scores. In order to discuss the correlations between simple structure dimensions and reliability, the variance of the true restricted bi-factor scores were computed as the average variance over 100 replications. Table 4.9 shows the variance of the true restricted bi-factor score for each dimension and 10 cells. It is shown that higher correlations between simple structure dimensions results in lower variance of the true specific dimension scores and higher variance of the true general dimension score. The variance of the true score has direct effects on the reliability. It can be expected that lower variance of the specific true scores results in lower reliability. Therefore, it is predicted that conditions with lower correlations between simple structure dimensions will yield specific factors with higher reliability and a general factor with lower reliability. The proof in Appendix A shows that, when the correlation matrix displays compound symmetry as in the simulation below, the true score variance of the specific dimensions will decline as the dimension correlations increase.

Table 4.9. Variance of the True Restricted Bi-factor Score.

Condition				Dimension				
Test Length	Number of Dimensions	Correlations	Cell	G	S1	S2	S3	S4
		0.3	Cell 3 & Cell 13	0.54	0.46	0.46	0.47	--
	3	0.6	Cell 4 & Cell 14	0.73	0.27	0.27	0.27	--
		0.3	Cell 5 & Cell 15	0.48	0.52	0.53	0.53	0.52
15	4	0.6	Cell 6 & Cell 16	0.7	0.3	0.3	0.3	0.3
		0.3	Cell 7 & Cell 17	0.54	0.46	0.46	0.46	--
	3	0.6	Cell 8 & Cell 18	0.73	0.27	0.27	0.27	--
		0.3	Cell 9 & Cell 19	0.47	0.53	0.53	0.53	0.53
30	4	0.6	Cell 10& Cell 20	0.7	0.3	0.3	0.3	0.3

Bias and Absolute Bias. Conditional bias as a function of θ is shown in Figure 4.1, and absolute conditional bias in Figure 4.4. For test lengths of 15 and 30 when data were generated by the Rasch model. The conditional bias values for the 2PL model are presented in Figure 4.2 and Figure 4.3 and the absolute bias values are in Figure 4.5 and Figure 4.6. Table B.1 in Appendix B contains results of bias values and Table B.2 in Appendix B contains results of absolute bias values for cell 1 to cell 20 with 14 intervals of the true theta continuum. Regarding signed bias, closer bias values to zero mean higher estimation accuracy. In term of absolute bias, lower values indicate higher accuracy. Figures 4.1, 4.2, 4.3, 4.4, 4.5, and 4.6 all illustrate similar conclusions. Because the specific dimension scores illustrate similar results, the plot only shows the specific dimension 1. First, longer length of test for both the general dimension score and the specific dimension scores results in lower absolute bias, which means higher estimation accuracy. Second, , lower dimension correlation for the simple structure dimension scores leads to slightly lower absolute bias for specific dimensions while there is no effect on absolute bias value for the general dimension score when varying the correlation. Third, varying the number of dimensions doesn't have much influence on the absolute bias for the specific dimension scores whereas higher dimensionality results in lower absolute bias for the general dimension score because higher dimensionality means that the general score is based on more items. . Finally, all figures indicate that when the true theta is closer to zero, bias or absolute bias are closer to zero.

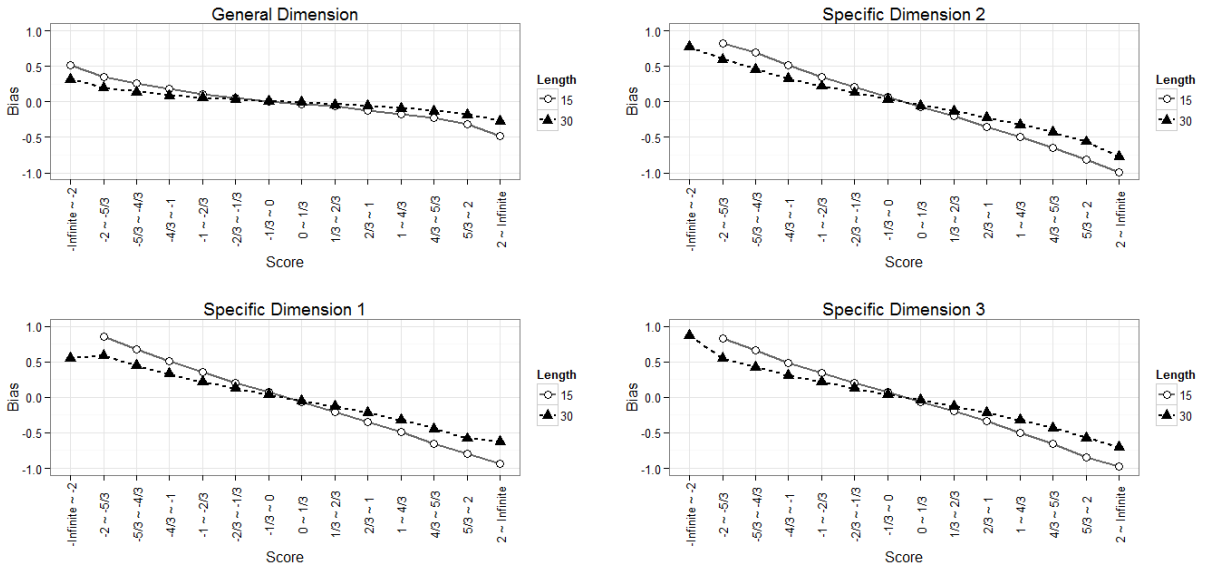


Figure 4.1. Conditional Bias under Two Levels of Length for the Rasch Model and EAP

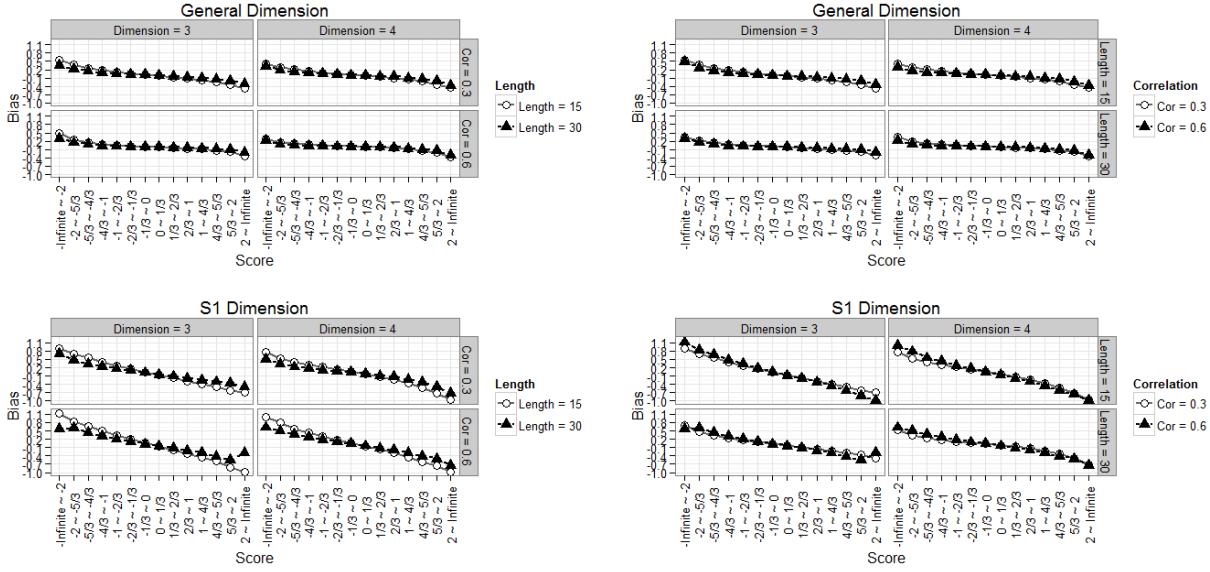


Figure 4.2. Conditional Bias under Various Conditions of Length and Correlation for the 2PL Model and EAP

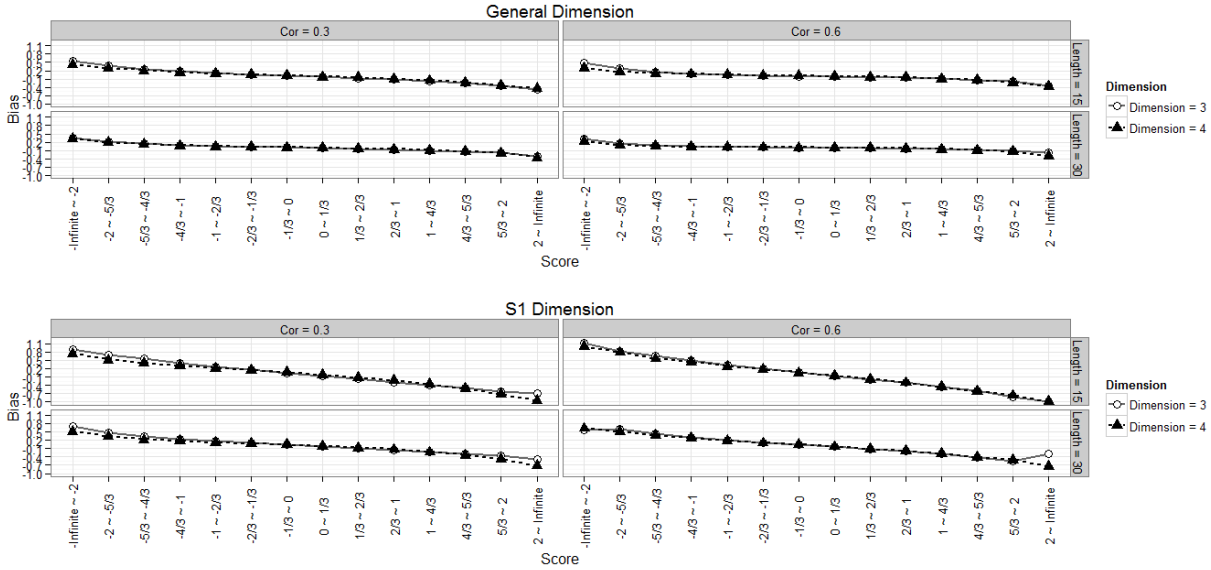


Figure 4.3. Conditional Bias under Varying Conditions of Dimensionality for the 2PL Model and EAP

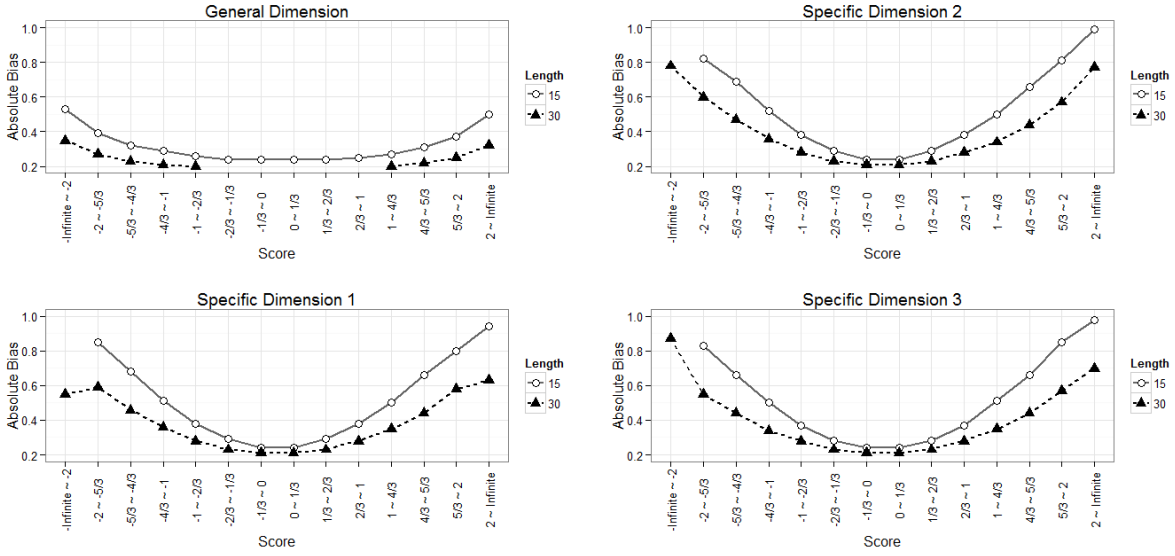


Figure 4.4. Absolute Conditional Bias under two Levels of Length for the Rasch Model and EAP

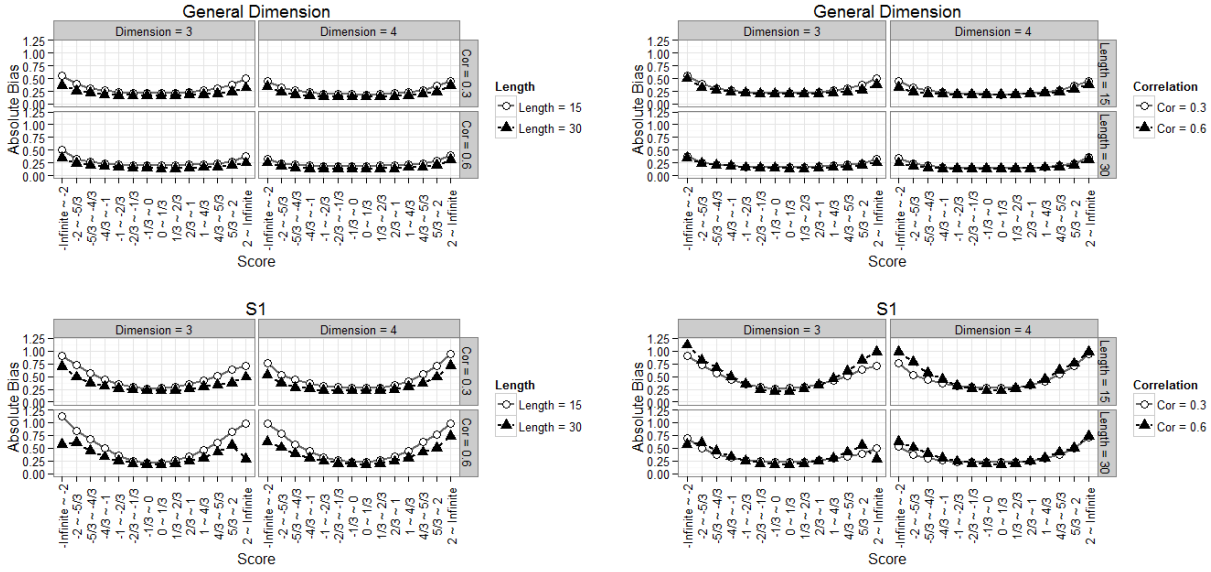


Figure 4.5. Absolute Conditional Bias under Varying Conditions of Length and Correlation for the 2PL Model and EAP

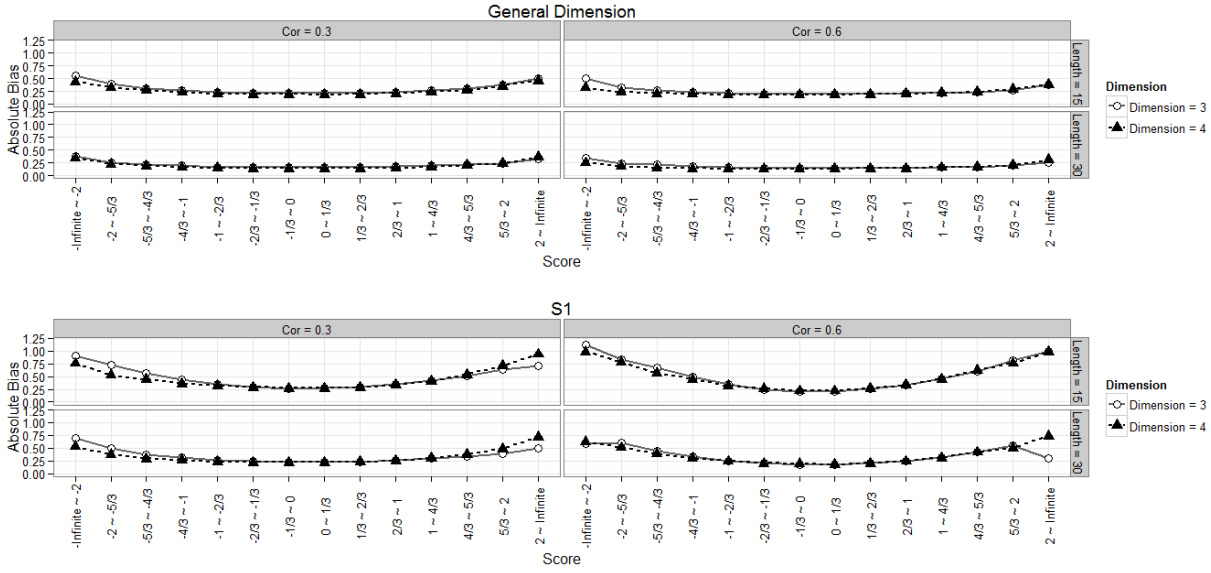


Figure 4.6. Absolute Conditional Bias under Varying Conditions of Dimensionality for the 2PL Model and EAP

The RMSE and the SE. Figure 4.7 presents the conditional RMSE and Figure 4.10 shows the SE for the EAP general dimension score and the specific dimension

scores when the test length increases from 15 to 30 and data were generated by the Rasch model. The RMSE values for data generated by the 2PL model are presented in Figure 4.8 and Figure 4.9, and the SE values are in Figure 4.11 and Figure 4.12. Table B.3 in Appendix B contains RMSE results and Table B.4 in Appendix B contains SE results for cell 1 to cell 20 for 14 intervals of the true theta continuum. The results of RMSE and SE corresponded closely to bias and absolute bias. That is, for the general dimensions, longer test lengths or higher dimensionality decrease RMSE or SE whereas varying correlations between the simple structure dimensions doesn't affect RMSE or SE. For the specific dimensions, longer test lengths result in lower RMSE or SE while increasing dimensions for the specific dimension scores doesn't affect RMSE or SE. When varying the correlations between simple structure dimensions, the RMSE or SE does not show obvious change for the specific dimensions.

One interesting fact here is that the figures relating the RMSE show the same results as bias or absolute bias that when the true theta is closer to zero, RMSE is closer to zero. However, the values of SE are pretty stable for the entire true theta continuum, especially for specific dimension scores, although the extreme values on the true theta continuum are not as stable.

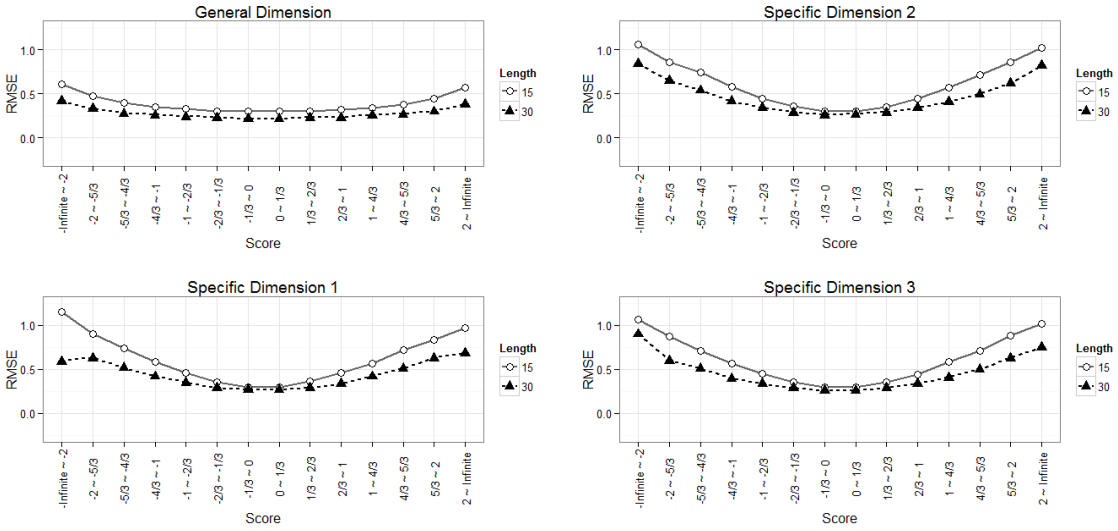


Figure 4.7. Conditional RMSE Under two Levels of Length for the Rasch Model and EAP

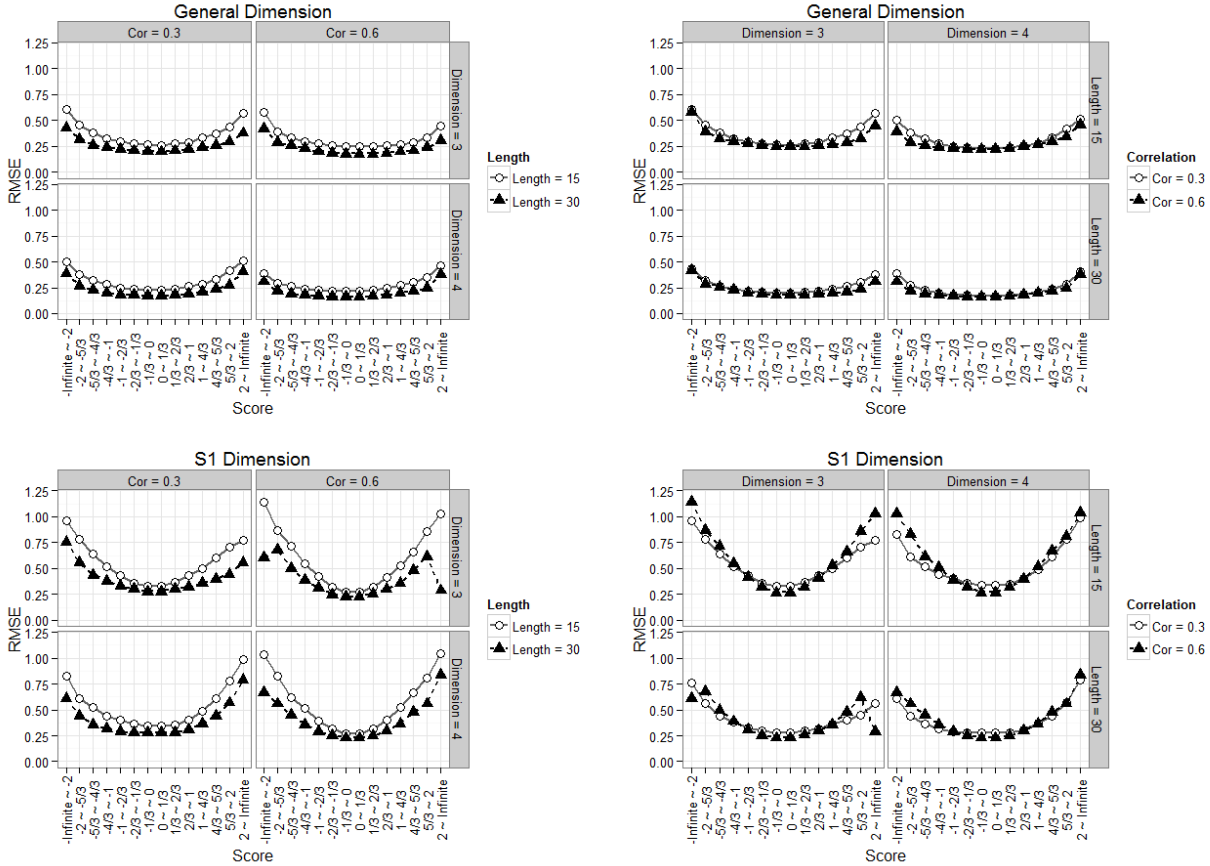


Figure 4.8. Conditional RMSE under Varying Conditions of Length and Correlation for the 2PL Model and EAP

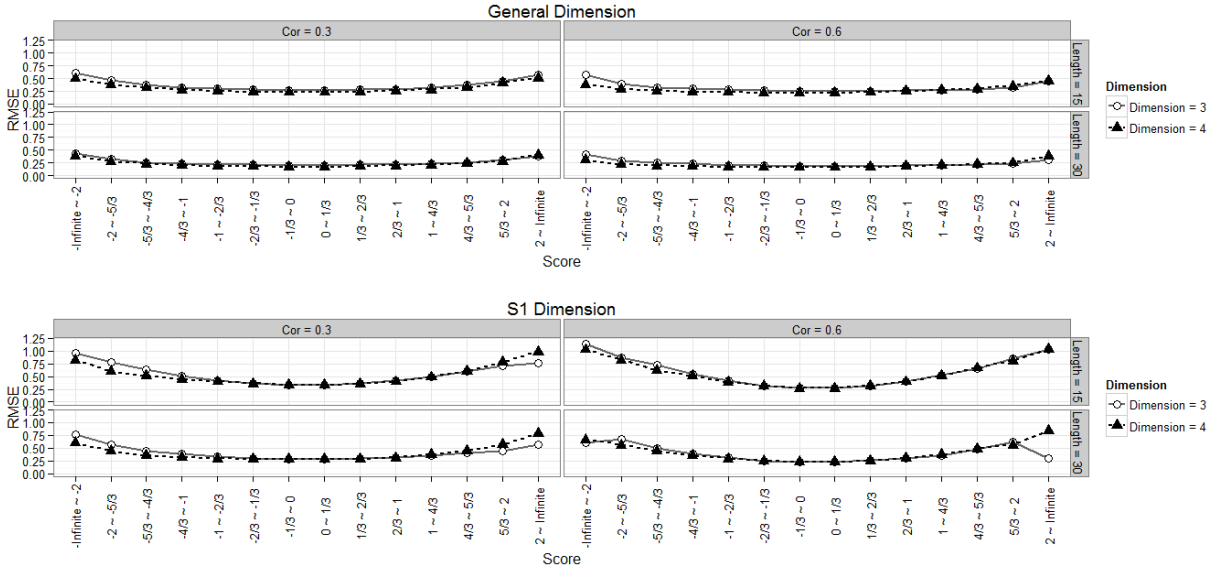


Figure 4.9. Conditional RMSE under Varying Conditions of Dimensionality for 2PL Model and EAP

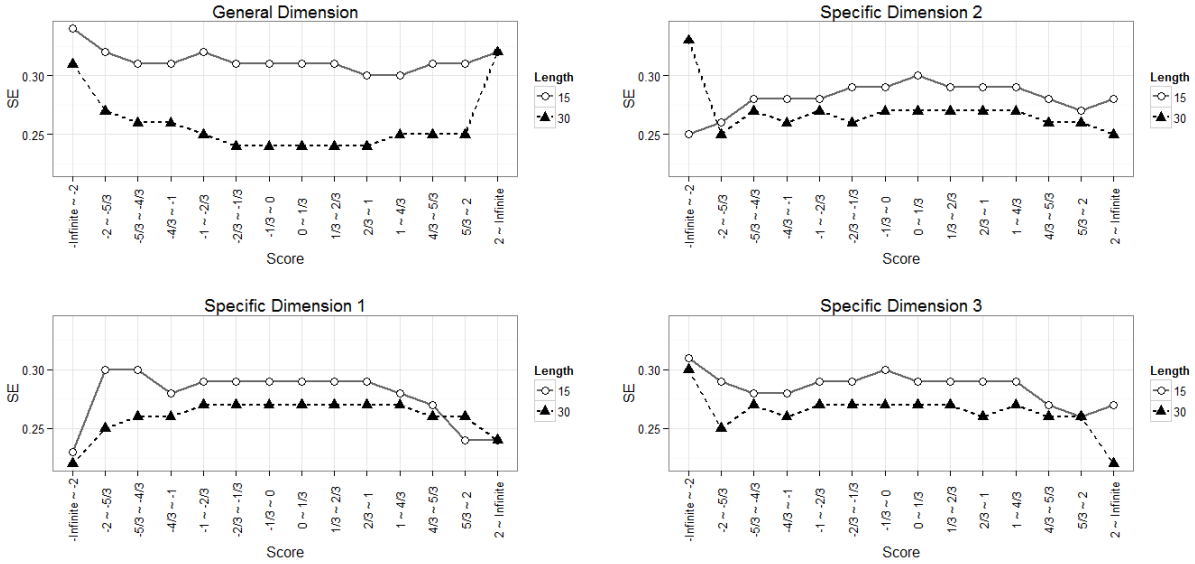


Figure 4.10. Conditional SE Under two Levels of Length for the Rasch Model and EAP

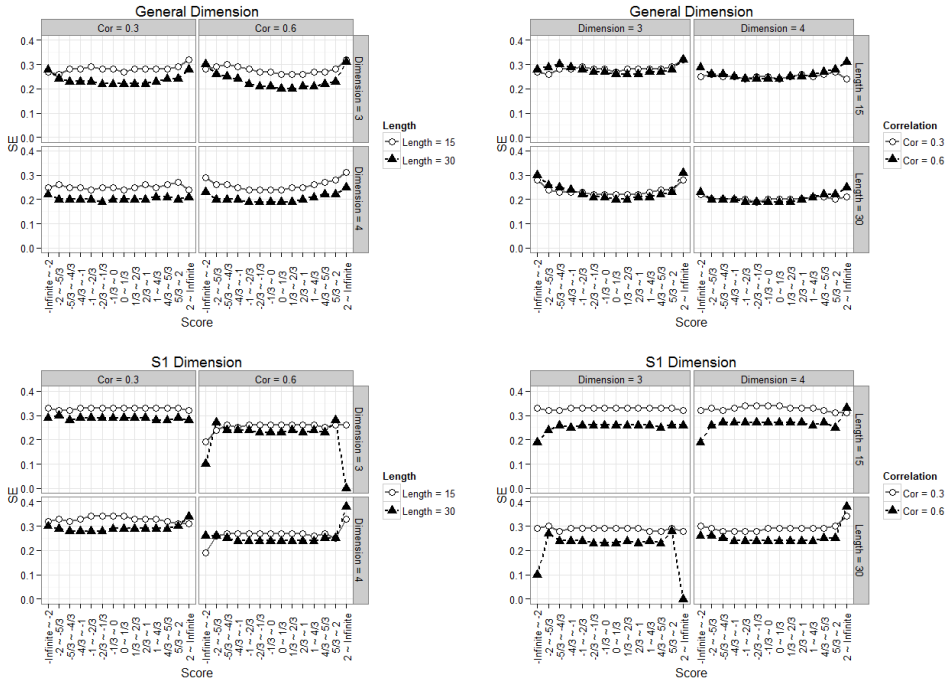


Figure 4.11. Conditional SE under Varying Conditions of Length and Correlation for the 2PL Model and EAP

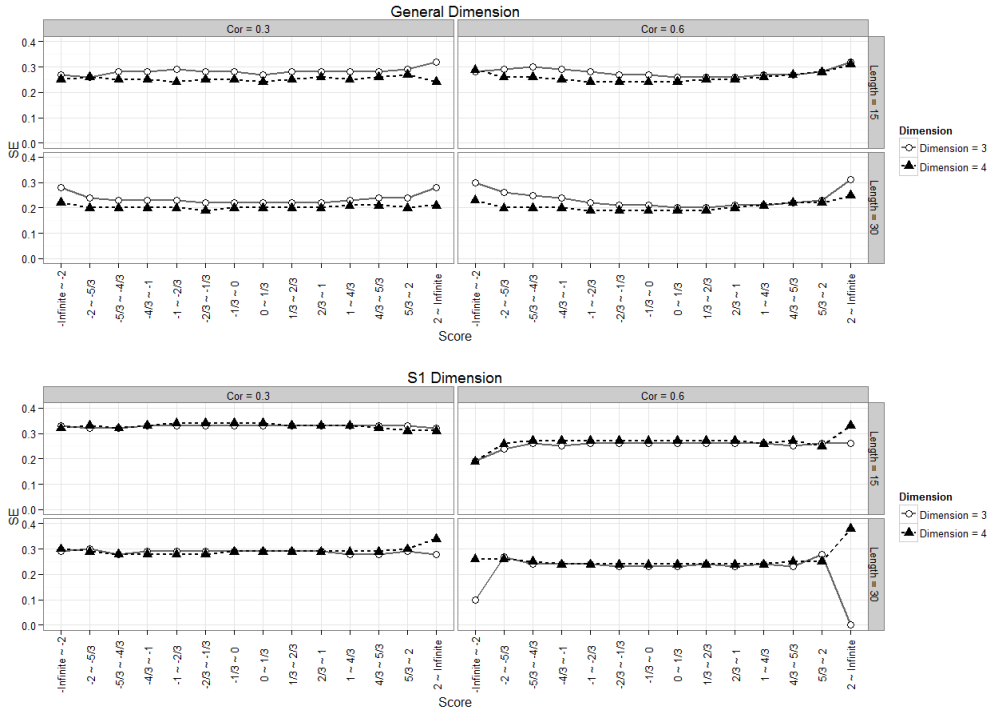


Figure 4.12. Conditional SE under Varying Conditions of Dimensionality for the2PL Model and EAP

Average Conditional Error Variance (Average MSE). Figure 4.13 presents the average conditional MSE for the EAP general dimension score and the specific dimension scores when the test length increased from 15 to 30 and data were generated by the Rasch model. The average MSE values for more simulation conditions and data generated by the 2PL model are presented in Figure 4.14 and Figure 4.15. Table B.5 in Appendix B contains average MSE results for cell 1 to cell 20 with 14 intervals of the true theta continuum. The average MSE results correspond closely to the RMSE, and SE results. Also, it also shows a similar pattern of results, specifically as the true theta approaches zero, average MSE is closer to zero.

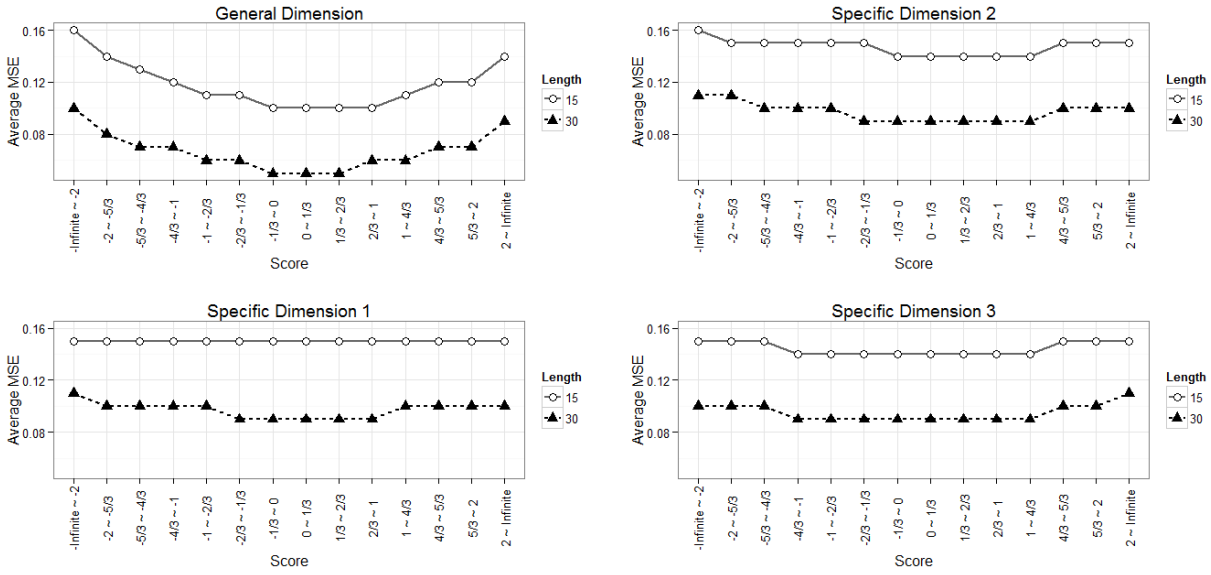


Figure 4.13. Average Conditional MSE Under Two Levels of Length for the Rasch Model and EAP

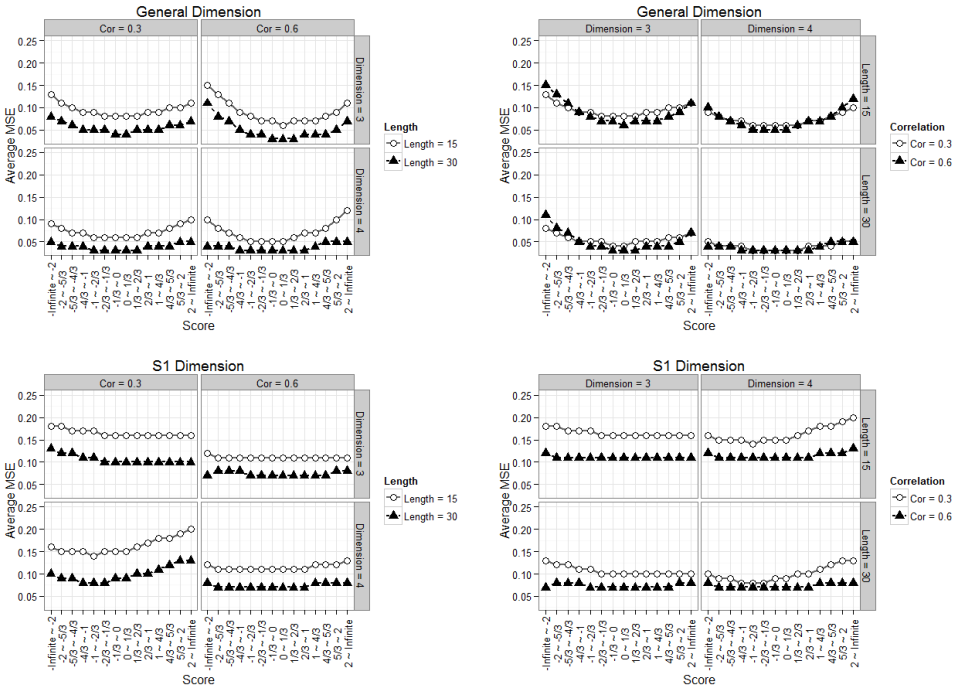


Figure 4.14. Average Conditional MSE under Varying Conditions of Length and Correlation for the Condition of the 2PL Model and EAP

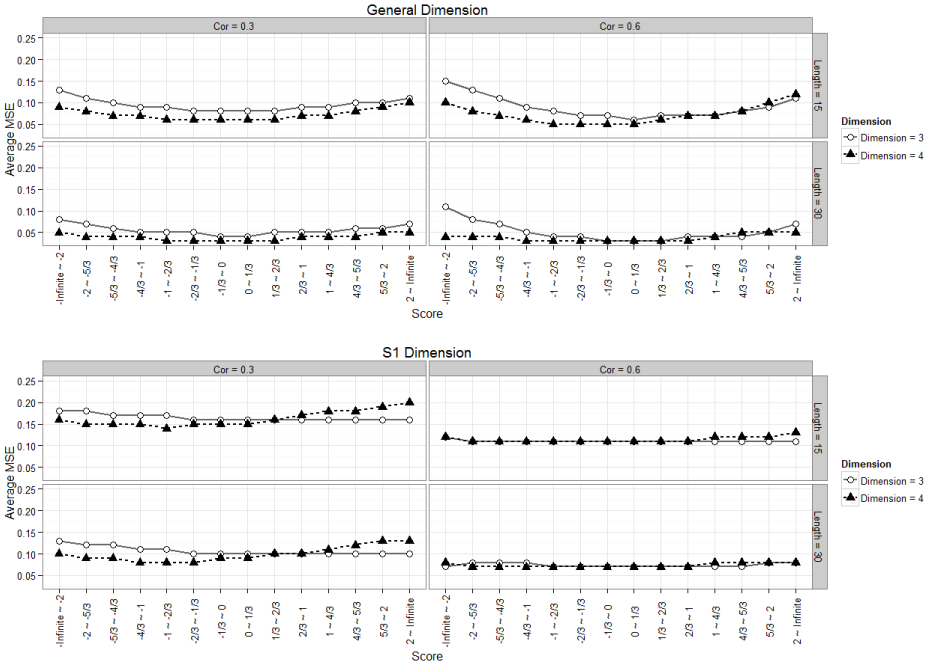


Figure 4.15. Average Conditional MSE for the Condition of Dimensionality and the Condition of the 2PL Model and EAP

Reliability. As mentioned in Chapter 3, the average reliability across replications using Equation 3.15 was computed for each dimension and each cell. Table B.6 in Appendix B contains reliability results for each dimension score for cell 1 to cell 20. There are some trends worth mentioning. First, in the Rasch model and EAP estimation, the varying condition is the length within each subdimension. Figure 4.16 indicates that longer length corresponds to higher reliability for both the general dimension and specific dimensions scores. Second, in the 2PL condition and EAP estimation, three varying conditions are length within each subdimension, correlation between simple structure dimensions, and number of subdimensions. The left plot in Figure 4.17 shows how the reliability varies under the condition of 3 and 4 subdimensions. Reliability is not materially affected for the specific dimensions scores when number of subdimensions increases whereas reliability increases for the general dimension score from three subdimensions to four subdimensions. This is because increasing the number of dimensions means more items contributing to the general dimension. As for the condition of length (the right plot in Figure 4.17), it shows the same trend as for the Rasch data: longer length contributes to higher reliability for both the general dimension and specific dimensions scores. The middle plot in Figure 4.17 shows how the reliability varies as a function of correlations. Reliability increases for the specific dimensions scores of the restricted bi-factor model when the correlation decreases whereas reliability varies little for the general dimension scores when the correlation changes. As is mentioned in the Chapter 3, the reliability increases for specific dimension scores as

correlation decreases because of the increase in true score variance. The proof in e Appendix A shows the relation between correlation and true score variance.

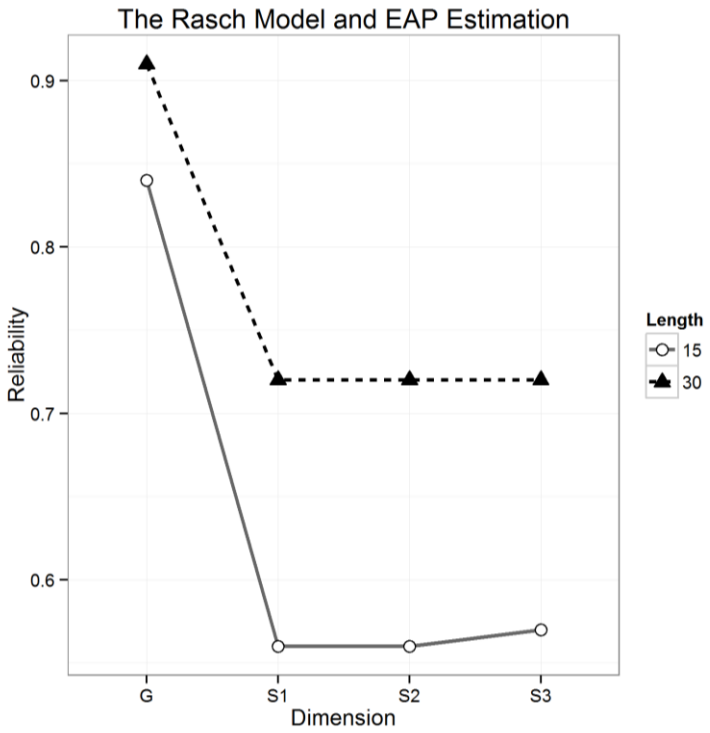


Figure 4.16. Reliability Under two Levels of Length for the Rasch Model and EAP

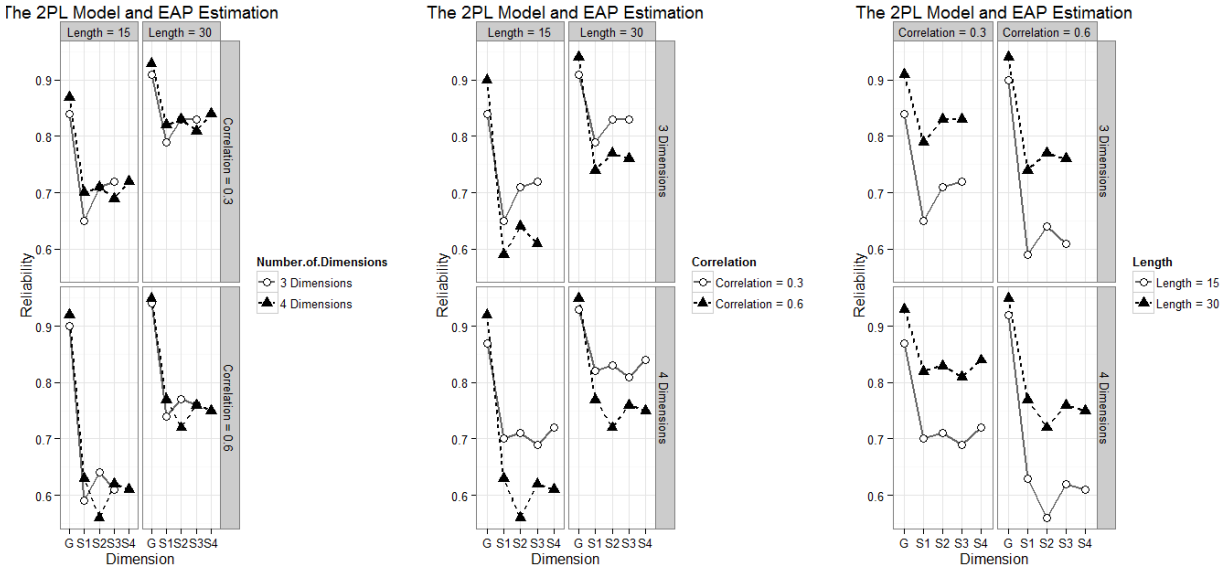


Figure 4.17. Reliability under Varying Conditions for the 2PL Model and EAP

Sensitivity and Specificity of the Score, z_{pj}. Sensitivity and specificity for true strengths and weaknesses were computed for each specific dimension score. Table B.7 in Appendix B contains results of sensitivity and specificity for the strengths and weaknesses along each dimension score for cell 1 to cell 20. Figure 4.18 shows that sensitivity and specificity for the strengths and weaknesses under two levels of Length for the Rasch model and EAP estimation method. It shows that sensitivity for the strengths and weaknesses increases from the smaller to the larger length but specificity decreases from the smaller to the larger test length for each specific dimension score. As test length increases, the person is more likely to be labeled as having a true strength or weakness irrespective of whether or not they have one.

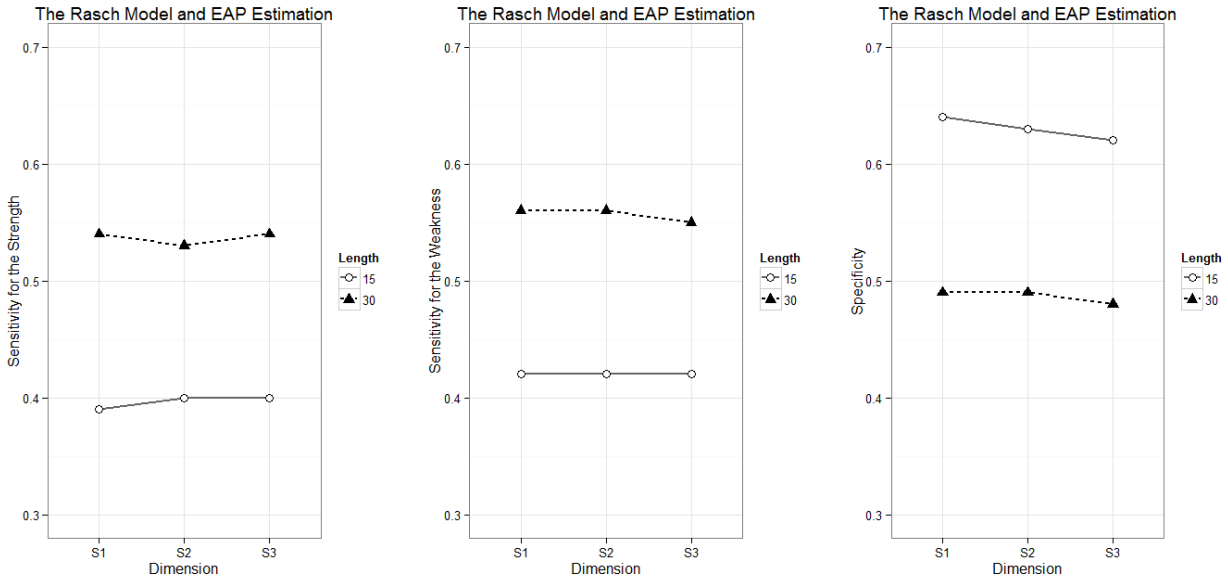


Figure 4.18. Sensitivity and Specificity for True Strengths and Weaknesses Under two Levels of Length for the Rasch Model and EAP

Figure 4.19 shows sensitivity for strengths and Figure 4.20 shows sensitivity for weakness under varying conditions of correlation, number of dimensions and length for

the 2PL Model and EAP. The two figures show similar trends. First, lower correlation results in higher sensitivity for the strengths and weaknesses (the middle plot in Figure 4.19 and Figure 4.20) whereas lower correlation results in lower specificity (the middle plot in Figure 4.21). Second, longer length results in higher sensitivity for strengths and weaknesses (the right plot in Figure 4.19 and Figure 4.20) whereas shorter length results in higher specificity (the left plot in Figure 4.21). Third, as for number of dimensions, it has little effect on sensitivity or specificity as the number of dimensions increases.

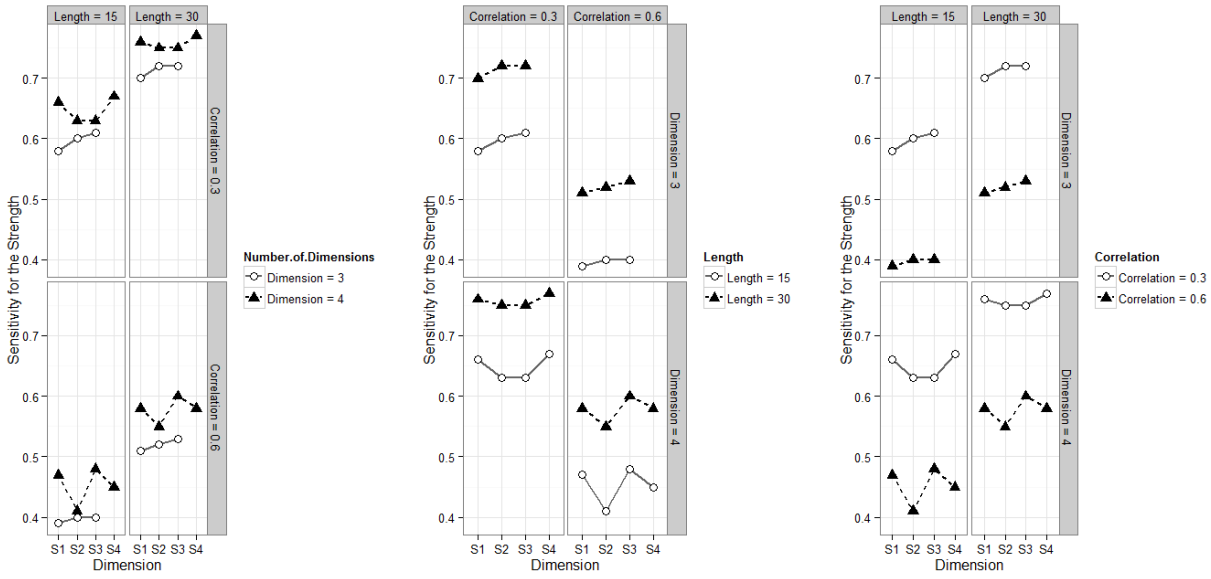


Figure 4.19. Sensitivity to Strengths under Varying Conditions for the 2PL Model and EAP

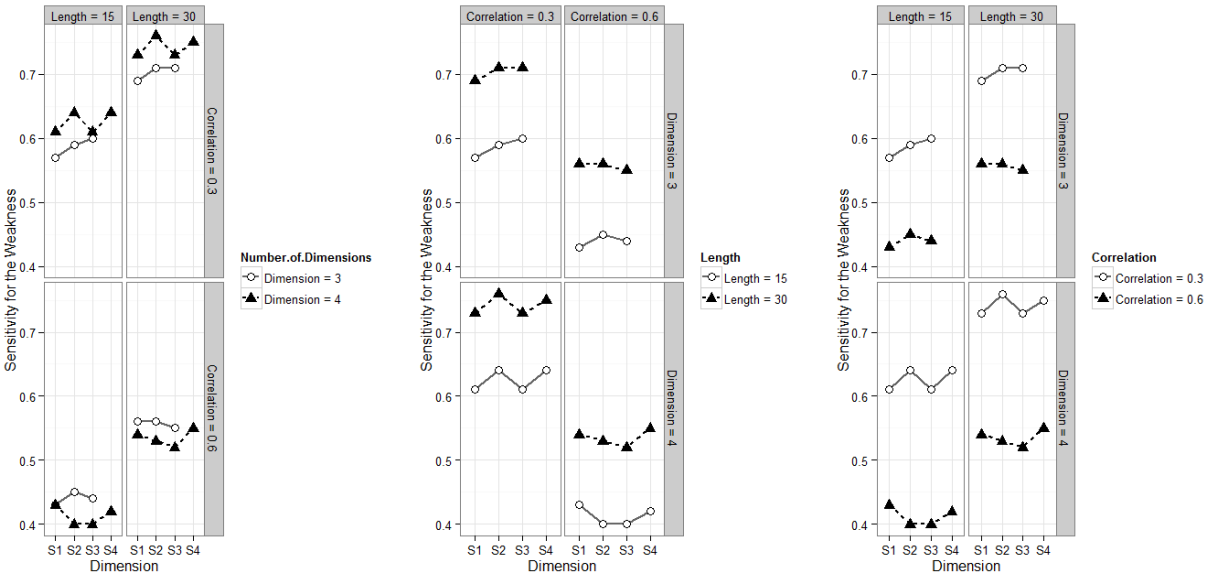


Figure 4.20. Sensitivity to Weaknesses under Varying Conditions for the 2PL Model and EAP

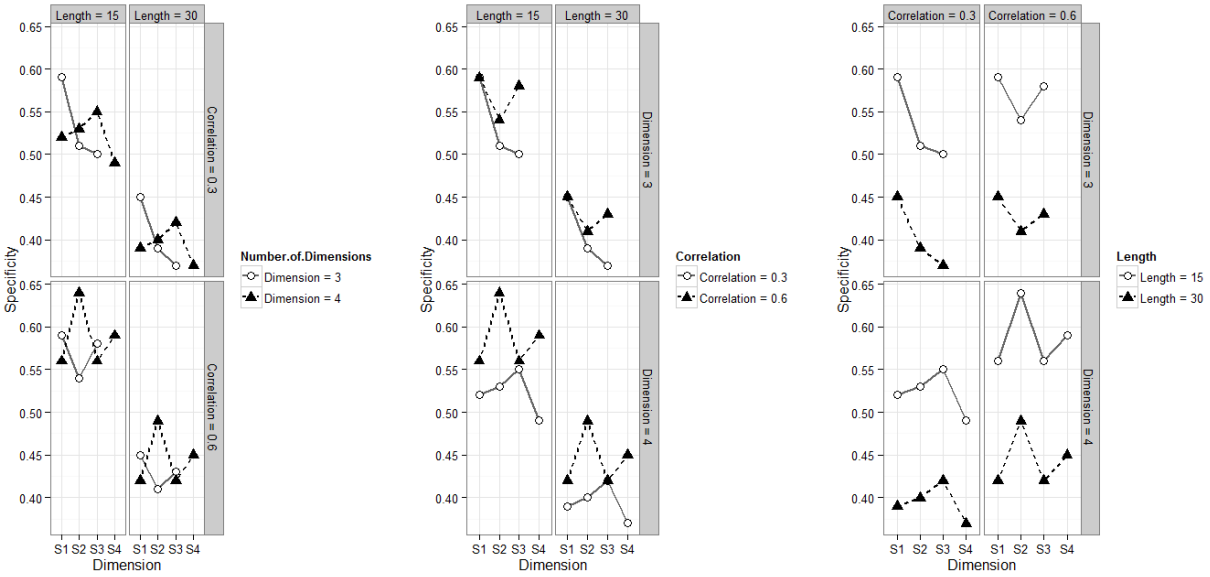


Figure 4.21. Specificity under Varying Conditions for the 2PL Model and EAP

Chapter V: Conclusion

Chapters 1 and 2 stated that there are increasing demands to report subscores in educational and psychological assessments. According to the Standards for Educational and Psychological Testing, reporting subscores requires satisfying standards of psychometric quality, such as validity, comparability, reliability, and interpretability. MIRT ability estimation is one method for estimating subscores. Also, the bi-factor model in MIRT can be especially beneficial, where the model accounts for a general dimension on which all items load and specific dimensions corresponding to the subdomains from which the items come. However, interpretation of subscores is a major obstacle to implementing the bi-factor model in applied testing.

The critical restriction for traditional IRT (UIRT) is the unidimensionality assumption. Substantial research has developed MIRT models to address this issue. Chapter 2 reviewed a variety of MIRT models, including the item and person parameter estimation methods. This study focuses on the MIRT bi-factor model due to the advantage of estimating both the overall score and susbcores. A wide range of studies applying the bi-factor model were reviewed in Chapter 2. In many studies, they have focused on model comparison, advantages for applying the bi-factor model, assessment of dimensionality, estimation and interpretation of the item parameters, or estimation algorithms for the bi-factor model. However, person parameters have rarely been emphasized, especially the interpretation of the subscores in those studies applying the bi-factor model. Although reporting subscores requires interpretable scores, previous

work has not presented approaches to interpret subscores estimated from the bi-factor model.

Brandt (2008) proposed a Rasch subdimension model, which is a special case of the multidimensional random coefficients multinomial logit model (MRCML). This model provides an approach to interpretation of subscores. The limitations of this model are that covariances between all specific dimensions and the main dimension equal 0 and constraining the item discrimination parameters equal to 1 for each item. The main purpose of this study was to build upon the previous studies and to develop a restricted bi-factor model in which the general dimension represents the examinee's overall performance in a domain, and each specific dimension represents a deviation from that overall performance. The specific dimension scores describe the examinee's pattern of strengths and weaknesses relative to the examinee's overall performance.

As discussed in Chapter 3, the restricted bi-factor model assumes data satisfy the simple structure model and the discrimination parameter is equal for the general dimension and specific dimensions. The first assumption allows estimating covariances between all specific dimensions. The main purpose of the restricted bi-factor model is to represent subscores in a meaningful way. There are two main characteristics for interpreting subscores estimated from the restricted bi-factor model. First, positive values on the specific dimension score represent a relative strength, performance in subdomain *j* higher than the overall performance, whereas negative values on the specific dimension score suggests a relative weakness. Second, because the standard errors for

each specific dimension are provided, the statistical significance of a relative strength or weakness can be computed. In the real data study, the 2PL and Rasch restricted bi-factor model with the EAP and MAP estimation methods were illustrated. In the simulation study, the 2PL and Rasch restricted bi-factor model with the EAP and MAP estimation method, different levels of correlations, number of dimensions, and test lengths were used.

Conclusion of Real Data Study

The main purpose of the real data study was to demonstrate the restricted bi-factor model for estimating and interpreting scores including overall score and subscores. The real data study reveals that the EAP and MAP estimation method yields similar estimates for both the 2PL and Rasch restricted bi-factor model shown in Table 4.2 in Chapter 4. Because the EAP and MAP estimation methods yield similar estimates, the reliabilities are almost identical for the EAP and MAP estimation method (see Table 4.3 in Chapter 4). Moreover, comparing the 2PL and Rasch restricted bi-factor model, the 2PL restricted bi-factor model yielded more accurate estimation. This trend is shown in Table 4.3 and Table 4.4 in Chapter 4. Table 4.3 shows that the 2PL restricted bi-factor model yields higher reliabilities than the Rasch model for most of the specific dimensions whereas the MSE for the 2PL model is relatively lower than the Rasch model.

The restricted bi-factor model serves as a tool to identify examinees' relative strengths and weaknesses with specific dimension scores that can be interpreted relative to the overall performance (the general dimension score). Since the restricted bi-factor model provides the specific dimension scores as well as their standard errors, combining

the specific dimension scores and standard errors provides a way to identify subdomains in which the student's performance is significantly above or below their overall performance.

Conclusion of Simulation Study

The purpose of this simulation study is to evaluate the restricted bi-factor model recovery of the person parameters. The results for the simulation conditions of generating model using the pseudo Rasch simple structure model are similar to those using the 2PL simple structure model so the following paragraph discusses the general results of the two models. The results of the MAP and the EAP are discussed first. Moreover, the results of the simulation study for the condition of correlation between dimensions of the simple structure model, the number of items in the test, and the number of dimensions in the test are discussed below for the restricted bi-factor model, with an emphasis on the evaluation criteria for each simulation condition including reliability, sensitivity, specificity, and recovery of person parameters, including bias, root mean squared error (RMSE), standard error (SE), and the average conditional error variance.

Ability Estimation of the MAP and the EAP. The two ability estimation methods produce similar estimated scores but the MAP has slightly lower error variance than the EAP.

Recovery of Person Parameters. First, longer length of test for the general dimension score and the specific dimension scores will result in higher estimation accuracy and more stable parameter estimates, like lower absolute bias, lower RMSE and lower average MSE**.** Second, there is little effect on estimation accuracy for the general

and specific dimension score when varying the correlation. Third, varying the number of dimensions doesn't have much influence on accuracy and stability of parameter estimation for the specific dimension scores whereas higher dimensionality results in more accurate and stable parameter estimation for the general dimension score because higher dimensionality means longer test length.

Reliability. First, as for the number of subdimensions, reliability is not materially affected for the specific dimension scores when number of subdimensions increases whereas reliability increases for the general dimension score when increasing the number of subdimensions. This is because increasing the number of subdimensions means more items contributing to the general dimension. Second, reliability increases for the general dimension scores and specific dimension scores when the test length increases. Third, higher correlation between dimensions for the simple structure model results in lower reliability for the specific dimension scores whereas the simulation condition of correlation between dimensions has no effect on the reliability of the general dimension. The proof in Appendix A confirms the same fact that if the simple structure correlations are low, the true specific dimension variation increases, and that leads to higher specific dimension reliability.

Sensitivity. First, lower correlation results in higher sensitivity to strengths and weaknesses. Second, longer length results in higher sensitivity to strengths and weaknesses. Finally, as for number of dimensions, it has little effect on sensitivity as the number of dimensions increases.

Specificity. Lower correlation results in lower specificity. Also, shorter length results in higher specificity. However, increasing the number of dimensions has little effect on specificity.

Limitations

There are several limitations in this thesis. First, when the number of dimensions increases, the amount of computing time increases exponentially. In order to compute the restricted bi-factor model score, the first step is to fit the simple structure model. As we know, the computation becomes much slower when the number of dimensions increases for fitting MIRT.

Second, items are assumed to satisfy the simple structure model in order to fit the restricted bi-factor model. If this assumption cannot be met due to the nature of the assessment, the restricted bi-factor model score cannot be justified.

Third, computing the restricted bi-factor model requires two steps of analysis. That is, the simple structure model needs to be fit first to compute the person scores using MIRT software, such as FLEXMIRT (Cai, 2012). And, the scoring method described in Chapter 3 is used to obtain the general dimension score and the specific dimension scores of the restricted bi-factor model along with the conditional standard errors using a statistical computing program, such as R (R. C., 2012). The software, which computes the restricted bi-factor scores in a one-step process, can be developed. An R shell that incorporates IRT software, such as FLEXMIRT, and the code for the algorithm in Chapter 3 would make it a one-step process. Or, writing an R program that takes the item

parameters and outputs the EAP or MAP scores along with the corresponding general and specific dimension scores would also reduce it to a one-step process.

Fourth, the specificity measure in this study was defined as the proportion who do not have a true specific strength or weakness along the dimension and who are correctly identified as such. Examinees in the "no strength or weakness" category do have some nonzero level of strength or weakness. As test length increases, the analysis becomes more sensitive to those small strengths and weaknesses. As test length increases, sensitivity increases, but users may need to evaluate whether any substantial " strength or weakness is of practical importance, because practically unimportant (but nonzero) levels of strength and weakness become more likely to be identified as "significant" as test length increases.

Finally, in order to set the origin for specific dimensions, it is necessary to decide whether differences in performance across domains are due to differences in person abilities or differences in item difficulties. This decision can have a notable influence on the person parameters along the specific dimensions.

Future Work

A simulation study in which the data violate the simple structure model assumption could be conducted to check the person parameter recovery. Real data violates the simple structure model assumption at some level. When this happens, simulation could establish to what degree violations will have impacts on the person parameter recovery.

Application

There are several advantages to adopting the restricted bi-factor model. First, the restricted bi-factor model ensures that the units of the general dimension and the units of the specific dimensions are the same. For any item *i*, a one unit change in the specific dimension θ^*_{pj} produce the same change in $P(x_{ip}=1|\bar{\theta}_p,\theta^*_{pj})$ as does a one unit change in the general dimension $\bar{\theta}_p$ in Equation 7 in Chapter 3. The benefit to have the same units of the general dimension as the units of the specific dimensions is that specific dimensions expressed in units comparable to those of $\bar{\theta}_p$ facilitate comparisons across the dimensions.

Second, the specific dimensions can be interpreted relative to the general dimension. That is, a positive value suggests item performance above that predicted by the general dimension and a relative strength, whereas negative values suggest performance below that predicted from the general dimension and a relative weakness.

Third, for the specific dimension scores to be useful in practice, the reliabilities must be respectable. For example, reliability of the specific dimension scores need to be higher than 0.7. The simulation study shows the dimension intercorrelations must be lower than 0.3 when test length for each subdimension is more than 15 or the dimension intercorrelations must be lower than 0.6 when test length for each subdimension is more than 30 in order to have reliability for the specific dimension scores higher than 0.7. If the restricted bi-factor model is considered to fit the data, the assessment needs to satisfy those conditions to have a reliable specific dimension score. The number of items

needed to obtain any given specific dimension reliability is a function of both the simple structure dimension correlations.

Fourth, because the restricted bi-factor model provides examinees their strengths and weaknesses from the assessment, scientific analysis, like sensitivity and specificity, should be conducted to evaluate the quality of the assessment. Various decision outcomes require different guidelines the sensitivity. The typical recommendation for sufficient sensitivity is higher than 0.7 or 0.8. The simulation study shows the dimension intercorrelations must be lower than 0.3 when test length is 30 for each subdimension in order to have sensitivity of the specific dimension scores higher than 0.7 when a true weakness is defined as a true specific dimension of 1.00 or greater.

Finally, the standard errors are standard errors on deviations from the overall ability level and hence provide a basis for identifying statistically significant strengths and significant weaknesses.

Final Thoughts

The issue of computing appropriate subscores has been studied for several years. Unless the subscores are appropriately computed, reporting subscores raise questions. This study developed a restricted bi-factor model, which can not only provide appropriate subscores but also examinees' strengths and weaknesses. The restricted bi-factor model provides researchers and practitioners a measurement tool for computing appropriate subscores and providing examinees their strengths and weaknesses thereby improving reporting of test results.

References

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

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

American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*.

Bartholomew, D. J. (1980). Factor analysis for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 293-321.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

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

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

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied psychological measurement, 6*(4), 431-444.

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In M. von Davier & D. Hastedt (Eds.), IERI Monograph Series. Issues and Methodologies in

Large Scale Assessments (Vol. 1, pp. 51-70). Princeton, NJ: IEA-ETS Research Institute.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological methods*, *16*(3), 221.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, *75*(1), 33-57.

Cai, L. (2010). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, *35*(3), 307-335.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581-612.

Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. *Seattle, WA: Vector Psychometric Group, LLC*.

Chib, S., & Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, *49*(4), 327-335.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. Multivariate behavioral research, 41(2), 189.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. *Psychometrika*, *40*(1), 5-32.

Davison, M. L., Chang, Y. F., & Davenport Jr, E. C. (2014). Modeling Configural Patterns in Latent Variable Profiles: Association With an Endogenous Variable. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(1), 81-93.

Davison, M. L., & Davenport Jr, E. C. (2002). Identifying criterion-related patterns of predictor scores using multiple regression. *Psychological methods*, *7*(4), 468.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order IRT model approach. *Applied Psychological Measurement*, *33*(8), 620-639.

DeMars, C. E. (2006). Application of the Bi-Factor Multidimensional Item Response Theory Model to Testlet-Based Tests. *Journal of educational measurement*, *43*(2), 145-168.

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006, April). A comparison of subscale score augmentation methods using empirical data. In *annual meeting of the national council of measurement in education, San Fransisco, CA*.

ELDA Test Administrator Webinar (2014). Baton Rouge, LA: Office of Standards, Assessment and Accountability, Department of Assessment and Accountability, Louisiana Department of Education. http://www.louisianabelieves.com/resources/library/webinars, accessed 3.17.2014.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists*. Psychology Press.

Fukuhara, H., & Kamata, A. (2011). A bifactor multidimensional item response theory model for differential item functioning analysis on testlet-based items. *Applied Psychological Measurement*, *35*(8), 604-622.

Fu, J., & Qu, Y. (2010, April). *A comparison of subscores reporting approaches on simulated data*. Paper presented at the meeting of the National Council on Measurement in Education, Denver, CO.

Gibbons, R D & Hedeker, D. R. (1992). Full-information item bi-factor analysis. Psychometrika, 57(3), 423.

Gibbons, R. D., Bock, R. D., Hedeker, D., Weiss, D. J., Segawa, E., & Bhaumik, D. K. (2007). Full information item bifactor analysis of graded response data. *Applied Psychological Measurement*, 31,4-19.

Gibbons, R. D., Rush, A. J., & Immekus, J. C. (2009). On the psychometric validity of the domains of the PDSQ: An illustration of the bi-factor item response theory model. *Journal of psychiatric research*, *43*(4), 401-410.

Gustafsson, J., & Balke, G. (1993). General and specific abilities as predictors of school achievement. Multivariate Behavioral Research, 28, 407-434.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, *19*, 49–78.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204-229.

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

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. Psychometrika, 2, 41-54.

Immekus, J. C., & Imbrie, P. K. (2008). Dimensionality Assessment Using the Full-Information Item Bifactor Analysis for Graded Response Data An Illustration With the State Metacognitive Inventory. *Educational and Psychological Measurement*, *68*(4), 695-709.

Interpretive guide: *English language development assessment* (2013). Baton Rouge, LA: Office of Standards, Assessment and Accountability, Department of Assessment and Accountability, Louisiana Department of Education. http://www.louisianabelieves.com/docs/assessment/elda-interpretive-guide.pdf?sfvrsn=4, accessed 3.11.2014

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, *38*(1), 32-60.

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183-202.

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores, *Journal of Educational Measurement*, *33*, 129 – 140.

Leading, L. L., & Monaghan, W. (2006). The Facts About Subscores. Princeton, NJ: Educational Testing Services.

Li, Y. & Lissitz, R. W. (2012). Exploring the full-information bifactor model in vertical scaling with construct shift. *Applied Psychological Measurement*, 36(1), 3.

Li, Y., & Rupp, A. A. (2011). Performance of the S− χ2 Statistic for Full-Information Bifactor Models. *Educational and Psychological Measurement*, *71*(6), 986-1005.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). Statistical theories of mental test scores.

Mr. Bayes, & Price, M. (1763). An Essay towards solving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions (1683-1775)*, 370-418.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351-356.

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.

Rasch, G. (1960/1981). Probabilistic Models For Some Intelligence And Attainment Tests Author: George Rasch, Publisher: Univ Of Chicago Pr (Tx).

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation*, *16 Suppl 1*, 19–31.

Reise, S. P., Ventura, J., Keefe, R., Baade, L., Gold, J., Green, M., Kern, R., Mesholam-Gately, R., Nuechterlein, K., Seidman, L. & Bilder, R.,(2011). Bifactor and item response theory analyses of interviewer report scales of cognitive impairment in schizophrenia. *Psychological assessment*, *23*(1), 245.

Reckase, M. D. (1972). Development and application of a multivariate logistic latent trait model. Unpublished doctoral dissertation, Syracuse Univeristy, Syracuse NY.

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

Reckase, M. D., & Hirsch, T. M. (1991). Interpretation of Number-Correct Scores when the True Number of Dimensions Assessed by a Test Is Greater than Two. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

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

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*(1), 51-67.

Rijmen, F. (2010). Formal Relations and an Empirical Comparison among the Bi-Factor, the Testlet, and a Second-Order Multidimensional IRT Model. *Journal of Educational Measurement*, *47*(3), 361–372.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, *22*, 400–407.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores.

Psychometrika monograph, No. 17.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 53-61.

Seo, D. G. (2011). *Application of the Bifactor Model to Computerized Adaptive Testing* (Doctoral dissertation, UNIVERSITY OF MINNESOTA).

Simms, L. J., Grös, D. F., Watson, D., & O'Hara, M. W. (2008). Parsing the general and specific components of depression and anxiety with bifactor modeling. *Depression and anxiety*, *25*(7), E34–46.

Sijtsma, K., & Junker, B. W. (2006). Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika*, *33*(1), 75-102.

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

Spray, J. A., Davey, T. C., Reckase, M. D., Ackerman, T. A., & Carlson, J. E. (1990). *Comparison of two logistic multidimensional item response theory models* (No. ACT-RR-ONR-90-8). AMERICAN COLL TESTING PROGRAM IOWA CITY IA.

Sympson, J. B. (1978). A model for testing with multidimensional items. In *Proceedings of the 1977 computerized adaptive testing conference* (pp. 82-98). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Takane, Y., & Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), 393–408.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscores performance. *Applied measurement in education, 17*(2), 89-112.

Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological methods, 9*(1), 116.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*(4), 479-494.

Yao, L. (2013). Multidimensional Item Response Theory for Score Reporting. *Advances in Modern, International Testing: Transition from Summative to Formative Assessment*.

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

Yung, Y. F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*(2), 113-128.

Zheng, C (2013). *Examination of the Parameter Estimate Bias When Violating the Orthogonality Assumption of the Bifactor Model* (Unpublished doctoral dissertation). University of Kansas, KS.

Appendix A

As covariace between the simple structure model subdimensions increases, fixing other conditions, the true score variances of the specific subdimensions of the restricted bi-factor model decreases (Proof 1) and the true score variance of the general dimension increases (Proof 2). In the proofs, *k* is the constant by which every covariance is increased.

Proof 1

Here, the variance of the specific dimension 1 true scores is designated as σ_1^2. The result is proved for the first specific dimension. In the proof below, $\text{cov}_{jj'}$ is the covariance of simple structure dimensions (j, j') and σ_j^2 is the variance of simple structure dimension *j*. The proof assumes the weights for the first specific dimension in the model in which the general dimension is an equally weighted sum so that, for specific dimension 1,

$w_1 = \frac{J-1}{J}$ and $w_{j\neq 1} = \frac{-1}{J}$.

$$\sigma_1^2 = \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}(\mathrm{cov}_{jj'} + k) \quad \text{(A1)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\,\mathrm{cov}_{jj'} + 2\sum_j\sum_{j'>j} w_j w_{j'}k \quad \text{(A2)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\,\mathrm{cov}_{jj'} + 2\sum_{j'>1} w_1 w_{j'}k + 2\sum_{j>1}\sum_{j'>j} w_j w_{j'}k \quad \text{(A3)}$$

$\because\ w_1 = \dfrac{J-1}{J}$ and $w_j = \dfrac{-1}{J}$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\,\mathrm{cov}_{jj'} + 2\sum_{j'>1}\left(\frac{J-1}{J}\right)\left(\frac{-1}{J}\right)k + 2\sum_{j>1}\sum_{j'>j}\left(\frac{1}{J^2}\right)k \quad \text{(A4)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\,\mathrm{cov}_{jj'} + 2\sum_{j'>1}\left(\frac{-J+1}{J^2}\right)k + 2\sum_{j>1}\sum_{j'>j}\left(\frac{1}{J^2}\right)k \quad \text{(A5)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\,\mathrm{cov}_{jj'} + 2\sum_{j'>1}\left(\frac{-J}{J^2}\right)k + \quad \text{(A6)}$$

$$2\sum_{j'>1}\left(\frac{1}{J^2}\right)k + 2\sum_{j>1}\sum_{j'>j}\left(\frac{1}{J^2}\right)k \quad \text{(A7)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\,\mathrm{cov}_{jj'} + 2\sum_{j'>1}\left(\frac{-1}{J}\right)k + 2\sum_j\sum_{j'>j}\left(\frac{1}{J^2}\right)k \quad \text{(A7)}$$

$\because\ 2\sum_{j'>1}\left(\dfrac{-1}{J}\right)k = -2\dfrac{(J-1)}{J}k$

$\because 2\sum_j\sum_{j'>1}\left(\dfrac{1}{J^2}\right)k = 2k\left(\dfrac{1}{J^2} + \dfrac{1}{J^2} + \dfrac{1}{J^2} + \ldots\right) = 2\left(\dfrac{J(J-1)/2}{J^2}\right)k$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\,\mathrm{cov}_{jj'} - 2\left(\frac{J-1}{J}\right)k + 2\left(\frac{J(J-1)/2}{J^2}\right)k \quad \text{(A8)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\,\mathrm{cov}_{jj'} - 2\left(\frac{J-1}{J}\right)k + \left(\frac{J-1}{J}\right)k \quad \text{(A9)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\,\mathrm{cov}_{jj'} - \left(\frac{J-1}{J}\right)k \quad \text{(A10)}$$

Note. The key term in Equation A10 is $-\left(\dfrac{J-1}{J}\right)k$ because it is the amount by which σ_1^2 is decreased by the increase of each covariance $\mathrm{cov}_{jj'}$ by *k*. Because both *J* and *k* are positive, $-\left(\dfrac{J-1}{J}\right)k$ is negative. Hence, the Proof 1shows that when *k* is added to each covariance, σ_1^2 decreases by the amount $\left(\dfrac{J-1}{J}\right)k$. That is, as the correlation between

subdimensions for the simple structure model increases, fixing other conditions, the true score variances of the specific subdimensions of the restricted bi-factor model decrease. Because the true score variance decreases, and reliability of the specific dimensions is a function of true score variance, the reliability of the specific dimension is expected to decrease.

Proof 2

Here, the variance of the general dimension true scores is designated as σ_g^2. The proof assumes the weights for the general dimension in the model in which the general dimension is an equally weighted sum so that, for the general dimension, $w_{j\neq 1} = \frac{1}{J}$.

$$\sigma_g^2 = \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}(\mathrm{cov}_{jj'} + k) \quad \text{(B1)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\mathrm{cov}_{jj'} + 2\sum_j\sum_{j'>j} w_j w_{j'}k \quad \text{(B2)}$$

$$\because \; w_j = \frac{1}{J}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\mathrm{cov}_{jj'} + 2\sum_j\sum_{j'>j} (\frac{1}{J})(\frac{1}{J})k \quad \text{(B3)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\mathrm{cov}_{jj'} + 2k\sum_j\sum_{j'>j} (\frac{1}{J^2}) \quad \text{(B4)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\mathrm{cov}_{jj'} + 2k(\frac{(J)(J-1)}{2})(\frac{1}{J^2}) \quad \text{(B5)}$$

$$= \sum_j w_j^2\sigma_j^2 + 2\sum_j\sum_{j'>j} w_j w_{j'}\mathrm{cov}_{jj'} + k(\frac{J-1}{J}) \quad \text{(B6)}$$

Note. The key term in Equation B6 is $k(\frac{J-1}{J})$ because it is the amount by which σ_g^2 is increased by the increase of each covariance $\mathrm{cov}_{jj'}$ by *k*. Because both *J* and *k* are positive, $k(\frac{J-1}{J})$ is positive. Hence, the Proof 2shows that when *k* is added to each

covariance, σ_g^2 increases by the amount $\mathbf{k}(\frac{\mathbf{J-1}}{\mathbf{J}})$. That is, as the correlation between subdimensions for the simple structure model increases, fixing other conditions, the true score variance of the general dimension of the restricted bi-factor model increases. Because the true score variance increases, and reliability of the general dimensions is a function of true score variance, the reliability of the general dimension is expected to increase.

Appendix B

Appendix B includes 6 tables for the simulation study. The first five tables contain results of bias values, absolute bias values, RMSE, SE, and average MSE for cell 1 to cell 20 with 14 intervals of the true theta continuum. The last two tables cover reliability values of each dimension score for cell 1 to cell 20 and sensitivity of the strength and weakness and specificity of each dimension score for cell 1 to cell 20.

Table B.1. Bias of Each Dimension Score for Cell 1 to Cell 20

			EAP			MAP				
			Cell 1			Cell 11				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.51	1.13	1.03	1.02	--	0.58	1.15	1.05	1.05	--
2	0.35	0.85	0.82	0.83	--	0.41	0.87	0.85	0.85	--
3	0.26	0.68	0.69	0.66	--	0.3	0.7	0.71	0.67	--
4	0.18	0.51	0.51	0.49	--	0.22	0.52	0.53	0.5	--
5	0.11	0.36	0.35	0.34	--	0.14	0.37	0.36	0.35	--
6	0.06	0.2	0.21	0.2	--	0.08	0.21	0.21	0.2	--
7	0.01	0.07	0.07	0.07	--	0.02	0.07	0.07	0.07	--
8	-0.03	-0.07	-0.07	-0.06	--	-0.03	-0.07	-0.07	-0.07	--
9	-0.06	-0.21	-0.2	-0.2	--	-0.07	-0.21	-0.21	-0.21	--
10	-0.12	-0.35	-0.35	-0.34	--	-0.14	-0.36	-0.36	-0.35	--
11	-0.17	-0.49	-0.49	-0.5	--	-0.19	-0.51	-0.51	-0.52	--
12	-0.22	-0.66	-0.65	-0.65	--	-0.26	-0.68	-0.67	-0.68	--
13	-0.32	-0.8	-0.81	-0.85	--	-0.37	-0.82	-0.83	-0.87	--
14	-0.48	-0.94	-0.99	-0.98	--	-0.54	-0.96	-1.01	-1.01	--
Score			Cell 2			Cell 12				
1	0.32	0.55	0.77	0.87	--	0.38	0.58	0.79	0.9	--
2	0.2	0.59	0.6	0.55	--	0.24	0.61	0.62	0.57	--
3	0.14	0.45	0.46	0.43	--	0.17	0.47	0.48	0.45	--
4	0.09	0.33	0.33	0.31	--	0.12	0.35	0.35	0.32	--
5	0.06	0.22	0.22	0.22	--	0.08	0.24	0.24	0.23	--
6	0.03	0.13	0.13	0.13	--	0.05	0.14	0.14	0.13	--
7	0.01	0.04	0.04	0.04	--	0.02	0.05	0.04	0.04	--
8	-0.01	-0.05	-0.05	-0.04	--	-0.01	-0.05	-0.05	-0.05	--
9	-0.04	-0.13	-0.13	-0.13	--	-0.04	-0.14	-0.14	-0.14	--
10	-0.06	-0.22	-0.22	-0.22	--	-0.07	-0.23	-0.23	-0.23	--
11	-0.09	-0.32	-0.32	-0.32	--	-0.11	-0.34	-0.33	-0.34	--
12	-0.13	-0.44	-0.43	-0.43	--	-0.16	-0.45	-0.45	-0.45	--
13	-0.18	-0.58	-0.56	-0.57	--	-0.22	-0.6	-0.59	-0.6	--
14	-0.27	-0.63	-0.77	-0.7	--	-0.32	-0.66	-0.8	-0.73	--

Table B.1. (Cont.)

	EAP					MAP				
	Cell 3					Cell 13				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.55	0.91	0.86	0.7	--	0.63	0.96	0.91	0.75	--
2	0.38	0.71	0.6	0.57	--	0.45	0.76	0.64	0.6	--
3	0.26	0.55	0.46	0.41	--	0.31	0.59	0.49	0.44	--
4	0.18	0.41	0.33	0.29	--	0.22	0.44	0.36	0.31	--
5	0.11	0.28	0.23	0.21	--	0.14	0.3	0.25	0.22	--
6	0.06	0.16	0.13	0.13	--	0.08	0.18	0.15	0.13	--
7	0.02	0.05	0.04	0.05	--	0.02	0.06	0.05	0.04	--
8	-0.02	-0.06	-0.05	-0.03	--	-0.02	-0.06	-0.04	-0.05	--
9	-0.07	-0.17	-0.13	-0.12	--	-0.09	-0.17	-0.14	-0.15	--
10	-0.12	-0.28	-0.22	-0.23	--	-0.14	-0.28	-0.23	-0.26	--
11	-0.18	-0.39	-0.32	-0.33	--	-0.21	-0.4	-0.34	-0.38	--
12	-0.24	-0.5	-0.42	-0.49	--	-0.28	-0.51	-0.45	-0.54	--
13	-0.33	-0.63	-0.56	-0.64	--	-0.38	-0.64	-0.59	-0.7	--
14	-0.49	-0.7	-0.73	-0.88	--	-0.55	-0.72	-0.78	-0.94	--
Score	Cell 4					Cell 14				
1	0.49	1.12	1.38	0.89	--	0.6	1.15	1.42	0.92	--
2	0.27	0.83	0.81	0.73	--	0.36	0.86	0.85	0.75	--
3	0.16	0.67	0.66	0.57	--	0.24	0.69	0.7	0.58	--
4	0.1	0.49	0.46	0.44	--	0.16	0.51	0.49	0.45	--
5	0.05	0.33	0.3	0.3	--	0.1	0.35	0.33	0.31	--
6	0.03	0.19	0.16	0.18	--	0.05	0.2	0.18	0.18	--
7	0	0.06	0.05	0.06	--	0.02	0.06	0.06	0.06	--
8	-0.02	-0.07	-0.06	-0.05	--	-0.02	-0.07	-0.06	-0.07	--
9	-0.03	-0.19	-0.17	-0.18	--	-0.04	-0.19	-0.17	-0.2	--
10	-0.06	-0.31	-0.27	-0.31	--	-0.08	-0.32	-0.28	-0.34	--
11	-0.09	-0.46	-0.4	-0.46	--	-0.12	-0.47	-0.42	-0.48	--
12	-0.13	-0.6	-0.5	-0.64	--	-0.17	-0.62	-0.52	-0.67	--
13	-0.19	-0.82	-0.64	-0.8	--	-0.24	-0.84	-0.67	-0.83	--
14	-0.34	-0.99	-0.65	-1.1	--	-0.41	-1.02	-0.69	-1.14	--

Table B.1. (Cont.)

	EAP					MAP				
	Cell 5					Cell 15				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.43	0.75	0.8	0.85	0.82	0.51	0.8	0.85	0.9	0.88
2	0.29	0.52	0.58	0.62	0.58	0.36	0.56	0.62	0.67	0.64
3	0.21	0.41	0.46	0.49	0.43	0.26	0.44	0.49	0.52	0.48
4	0.14	0.31	0.33	0.35	0.3	0.18	0.33	0.36	0.38	0.34
5	0.1	0.22	0.22	0.24	0.2	0.12	0.22	0.25	0.26	0.22
6	0.05	0.14	0.13	0.14	0.12	0.06	0.13	0.15	0.15	0.13
7	0.02	0.07	0.04	0.05	0.04	0.02	0.05	0.05	0.06	0.05
8	-0.02	-0.02	-0.05	-0.04	-0.03	-0.03	-0.05	-0.04	-0.04	-0.04
9	-0.05	-0.12	-0.14	-0.15	-0.11	-0.08	-0.15	-0.13	-0.15	-0.13
10	-0.1	-0.23	-0.22	-0.25	-0.2	-0.13	-0.27	-0.23	-0.26	-0.23
11	-0.15	-0.36	-0.31	-0.37	-0.31	-0.2	-0.41	-0.32	-0.39	-0.35
12	-0.22	-0.52	-0.42	-0.49	-0.45	-0.28	-0.57	-0.44	-0.52	-0.49
13	-0.33	-0.71	-0.54	-0.62	-0.61	-0.39	-0.76	-0.57	-0.64	-0.66
14	-0.44	-0.94	-0.75	-0.82	-0.84	-0.52	-0.99	-0.79	-0.86	-0.89
Score	Cell 6					Cell 16				
1	0.29	0.99	1.02	0.8	0.86	0.38	1.03	1.05	0.83	0.89
2	0.15	0.79	0.89	0.68	0.75	0.22	0.82	0.91	0.7	0.78
3	0.11	0.56	0.68	0.54	0.59	0.16	0.59	0.7	0.56	0.61
4	0.08	0.44	0.51	0.4	0.43	0.11	0.46	0.53	0.41	0.45
5	0.05	0.29	0.35	0.28	0.31	0.07	0.31	0.37	0.28	0.33
6	0.03	0.17	0.2	0.17	0.18	0.04	0.18	0.22	0.17	0.19
7	0.02	0.06	0.06	0.07	0.06	0.01	0.06	0.07	0.06	0.07
8	0	-0.05	-0.07	-0.04	-0.06	-0.02	-0.06	-0.06	-0.06	-0.05
9	-0.02	-0.17	-0.21	-0.17	-0.18	-0.05	-0.18	-0.2	-0.2	-0.19
10	-0.05	-0.3	-0.34	-0.32	-0.31	-0.09	-0.31	-0.34	-0.35	-0.32
11	-0.09	-0.45	-0.48	-0.49	-0.46	-0.14	-0.46	-0.49	-0.53	-0.47
12	-0.14	-0.62	-0.65	-0.71	-0.6	-0.2	-0.64	-0.65	-0.76	-0.62
13	-0.22	-0.76	-0.81	-0.9	-0.79	-0.3	-0.79	-0.82	-0.95	-0.81
14	-0.36	-0.98	-0.93	-1.06	-0.96	-0.44	-1.01	-0.95	-1.11	-0.99

Table B.1. (Cont.)

	EAP					MAP				
	Cell 7					Cell 17				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.36	0.7	0.59	0.49	--	0.44	0.75	0.64	0.53	--
2	0.22	0.48	0.38	0.34	--	0.28	0.52	0.42	0.38	--
3	0.15	0.34	0.29	0.25	--	0.19	0.38	0.32	0.28	--
4	0.1	0.25	0.2	0.18	--	0.13	0.28	0.22	0.19	--
5	0.06	0.17	0.13	0.11	--	0.08	0.19	0.15	0.12	--
6	0.04	0.1	0.07	0.07	--	0.05	0.11	0.08	0.07	--
7	0.02	0.03	0.02	0.03	--	0.02	0.03	0.03	0.02	--
8	-0.01	-0.04	-0.03	-0.01	--	-0.02	-0.04	-0.03	-0.03	--
9	-0.03	-0.1	-0.08	-0.06	--	-0.04	-0.1	-0.08	-0.09	--
10	-0.07	-0.17	-0.12	-0.13	--	-0.08	-0.17	-0.13	-0.16	--
11	-0.1	-0.24	-0.19	-0.21	--	-0.12	-0.24	-0.2	-0.24	--
12	-0.13	-0.3	-0.26	-0.29	--	-0.16	-0.3	-0.28	-0.33	--
13	-0.19	-0.36	-0.34	-0.42	--	-0.22	-0.37	-0.37	-0.47	--
14	-0.3	-0.49	-0.48	-0.59	--	-0.35	-0.5	-0.53	-0.65	--
Score	Cell 8					Cell 18				
1	0.31	0.58	0.63	0.89	--	0.41	0.62	0.67	0.91	--
2	0.15	0.61	0.62	0.52	--	0.23	0.64	0.66	0.55	--
3	0.09	0.44	0.44	0.39	--	0.15	0.47	0.47	0.41	--
4	0.05	0.31	0.31	0.28	--	0.1	0.33	0.34	0.29	--
5	0.02	0.21	0.2	0.19	--	0.06	0.22	0.22	0.19	--
6	0.01	0.12	0.1	0.11	--	0.03	0.13	0.12	0.11	--
7	0	0.03	0.03	0.04	--	0.01	0.04	0.04	0.04	--
8	-0.01	-0.04	-0.04	-0.03	--	-0.01	-0.05	-0.04	-0.04	--
9	-0.02	-0.12	-0.1	-0.11	--	-0.03	-0.12	-0.1	-0.13	--
10	-0.03	-0.2	-0.17	-0.2	--	-0.04	-0.21	-0.18	-0.22	--
11	-0.05	-0.29	-0.25	-0.3	--	-0.06	-0.3	-0.26	-0.33	--
12	-0.07	-0.42	-0.33	-0.4	--	-0.1	-0.44	-0.35	-0.43	--
13	-0.1	-0.55	-0.44	-0.57	--	-0.14	-0.58	-0.46	-0.61	--
14	-0.19	-0.29	-0.83	-0.66	--	-0.24	-0.33	-0.86	-0.69	--

Table B.1. (Cont.)

	EAP					MAP				
	Cell 9					Cell 19				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.34	0.52	0.55	0.6	0.58	0.35	0.55	0.58	0.63	0.63
2	0.2	0.34	0.38	0.41	0.38	0.22	0.37	0.41	0.44	0.42
3	0.13	0.23	0.29	0.3	0.27	0.16	0.26	0.31	0.33	0.3
4	0.08	0.18	0.2	0.21	0.18	0.1	0.19	0.22	0.23	0.21
5	0.05	0.12	0.14	0.14	0.11	0.07	0.13	0.15	0.16	0.13
6	0.03	0.08	0.08	0.08	0.07	0.04	0.08	0.09	0.09	0.08
7	0.01	0.04	0.02	0.03	0.02	0.01	0.03	0.03	0.03	0.03
8	0	-0.01	-0.03	-0.02	-0.01	-0.01	-0.02	-0.03	-0.02	-0.02
9	-0.03	-0.07	-0.08	-0.08	-0.06	-0.04	-0.09	-0.08	-0.09	-0.07
10	-0.05	-0.13	-0.13	-0.14	-0.12	-0.08	-0.16	-0.13	-0.15	-0.14
11	-0.09	-0.23	-0.18	-0.22	-0.19	-0.12	-0.26	-0.19	-0.23	-0.22
12	-0.14	-0.33	-0.25	-0.31	-0.29	-0.17	-0.37	-0.26	-0.33	-0.32
13	-0.19	-0.48	-0.34	-0.41	-0.39	-0.22	-0.52	-0.36	-0.42	-0.43
14	-0.35	-0.71	-0.48	-0.56	-0.57	-0.36	-0.74	-0.51	-0.58	-0.61
Score	Cell 10					Cell 20				
1	0.22	0.63	0.74	0.49	0.76	0.24	0.66	0.77	0.52	0.79
2	0.1	0.5	0.59	0.47	0.46	0.14	0.53	0.62	0.49	0.49
3	0.06	0.38	0.45	0.35	0.39	0.09	0.4	0.47	0.37	0.41
4	0.04	0.27	0.33	0.25	0.28	0.06	0.29	0.35	0.26	0.3
5	0.03	0.18	0.23	0.17	0.19	0.04	0.19	0.25	0.18	0.21
6	0.02	0.1	0.13	0.1	0.11	0.02	0.11	0.15	0.1	0.12
7	0.01	0.04	0.04	0.04	0.04	0.01	0.04	0.05	0.04	0.04
8	0	-0.03	-0.04	-0.02	-0.04	-0.01	-0.03	-0.04	-0.04	-0.04
9	-0.01	-0.11	-0.13	-0.11	-0.12	-0.03	-0.11	-0.13	-0.13	-0.12
10	-0.02	-0.19	-0.22	-0.21	-0.2	-0.05	-0.2	-0.22	-0.24	-0.21
11	-0.05	-0.29	-0.31	-0.34	-0.3	-0.08	-0.3	-0.31	-0.38	-0.31
12	-0.09	-0.42	-0.4	-0.49	-0.42	-0.13	-0.43	-0.4	-0.53	-0.43
13	-0.14	-0.5	-0.5	-0.7	-0.53	-0.18	-0.52	-0.5	-0.75	-0.55
14	-0.29	-0.74	-0.56	-0.83	-0.66	-0.29	-0.77	-0.56	-0.88	-0.68

Table B.2. Absolute Bias of Each Dimension Score for Cell 1 to Cell 20

	EAP					MAP				
	Cell 1					Cell 11				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.53	1.13	1.03	1.02	--	0.59	1.15	1.05	1.05	--
2	0.39	0.85	0.82	0.83	--	0.43	0.87	0.85	0.85	--
3	0.32	0.68	0.69	0.66	--	0.35	0.7	0.71	0.67	--
4	0.29	0.51	0.52	0.5	--	0.3	0.53	0.53	0.51	--
5	0.26	0.38	0.38	0.37	--	0.26	0.39	0.38	0.38	--
6	0.24	0.29	0.29	0.28	--	0.24	0.29	0.29	0.28	--
7	0.24	0.24	0.24	0.24	--	0.23	0.24	0.23	0.23	--
8	0.24	0.24	0.24	0.24	--	0.23	0.24	0.24	0.23	--
9	0.24	0.29	0.29	0.28	--	0.24	0.29	0.28	0.29	--
10	0.25	0.38	0.38	0.37	--	0.25	0.39	0.38	0.38	--
11	0.27	0.5	0.5	0.51	--	0.28	0.51	0.51	0.52	--
12	0.31	0.66	0.66	0.66	--	0.32	0.68	0.67	0.68	--
13	0.37	0.8	0.81	0.85	--	0.4	0.82	0.83	0.87	--
14	0.5	0.94	0.99	0.98	--	0.55	0.96	1.01	1.01	--
Score	Cell 2					Cell 12				
1	0.35	0.55	0.78	0.87	--	0.4	0.58	0.8	0.9	--
2	0.27	0.59	0.6	0.55	--	0.29	0.61	0.62	0.57	--
3	0.23	0.46	0.47	0.44	--	0.24	0.47	0.49	0.46	--
4	0.21	0.36	0.36	0.34	--	0.22	0.37	0.37	0.35	--
5	0.2	0.28	0.28	0.28	--	0.2	0.29	0.29	0.28	--
6	0.18	0.23	0.23	0.23	--	0.18	0.23	0.23	0.23	--
7	0.18	0.21	0.21	0.21	--	0.17	0.21	0.21	0.2	--
8	0.18	0.21	0.21	0.21	--	0.17	0.21	0.21	0.21	--
9	0.18	0.23	0.23	0.23	--	0.18	0.23	0.23	0.23	--
10	0.19	0.28	0.28	0.28	--	0.19	0.28	0.28	0.28	--
11	0.2	0.35	0.34	0.35	--	0.21	0.36	0.35	0.36	--
12	0.22	0.44	0.44	0.44	--	0.23	0.46	0.45	0.46	--
13	0.25	0.58	0.57	0.57	--	0.26	0.6	0.59	0.6	--
14	0.32	0.63	0.77	0.7	--	0.35	0.66	0.8	0.73	--

Table B.2. (Cont.)

	EAP					MAP				
	Cell 3					Cell 13				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.55	0.91	0.86	0.71	--	0.63	0.96	0.91	0.75	--
2	0.4	0.72	0.6	0.57	--	0.46	0.76	0.64	0.6	--
3	0.31	0.57	0.47	0.43	--	0.34	0.6	0.5	0.46	--
4	0.26	0.44	0.37	0.34	--	0.28	0.46	0.39	0.35	--
5	0.24	0.35	0.31	0.3	--	0.24	0.36	0.32	0.29	--
6	0.22	0.29	0.27	0.27	--	0.22	0.29	0.27	0.26	--
7	0.21	0.26	0.25	0.25	--	0.2	0.26	0.25	0.24	--
8	0.21	0.27	0.25	0.25	--	0.2	0.26	0.24	0.24	--
9	0.22	0.29	0.27	0.27	--	0.22	0.29	0.26	0.27	--
10	0.24	0.35	0.3	0.31	--	0.24	0.35	0.3	0.32	--
11	0.26	0.42	0.37	0.38	--	0.28	0.43	0.37	0.4	--
12	0.3	0.52	0.44	0.5	--	0.32	0.53	0.46	0.55	--
13	0.37	0.63	0.56	0.65	--	0.4	0.64	0.6	0.7	--
14	0.5	0.71	0.73	0.88	--	0.55	0.72	0.78	0.94	--
Score	Cell 4					Cell 14				
1	0.5	1.12	1.38	0.89	--	0.61	1.15	1.42	0.92	--
2	0.32	0.83	0.81	0.73	--	0.39	0.86	0.85	0.75	--
3	0.27	0.67	0.66	0.57	--	0.3	0.69	0.7	0.58	--
4	0.24	0.5	0.46	0.44	--	0.25	0.51	0.49	0.45	--
5	0.22	0.35	0.33	0.33	--	0.22	0.36	0.34	0.33	--
6	0.2	0.25	0.24	0.25	--	0.2	0.26	0.24	0.25	--
7	0.2	0.21	0.2	0.21	--	0.19	0.21	0.2	0.2	--
8	0.2	0.21	0.21	0.21	--	0.19	0.21	0.2	0.21	--
9	0.2	0.26	0.24	0.25	--	0.19	0.26	0.24	0.26	--
10	0.21	0.34	0.31	0.34	--	0.2	0.35	0.31	0.36	--
11	0.22	0.46	0.41	0.46	--	0.22	0.48	0.42	0.49	--
12	0.24	0.6	0.51	0.65	--	0.24	0.62	0.53	0.68	--
13	0.27	0.82	0.65	0.8	--	0.29	0.84	0.67	0.83	--
14	0.38	0.99	0.65	1.1	--	0.43	1.02	0.69	1.14	--

Table B.2. (Cont.)

	EAP					MAP				
	Cell 5					Cell 15				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.44	0.76	0.8	0.85	0.82	0.52	0.81	0.85	0.9	0.88
2	0.32	0.53	0.59	0.63	0.59	0.38	0.57	0.63	0.67	0.64
3	0.27	0.44	0.47	0.5	0.45	0.29	0.46	0.5	0.53	0.49
4	0.23	0.36	0.38	0.4	0.36	0.24	0.37	0.4	0.42	0.38
5	0.2	0.32	0.32	0.33	0.3	0.2	0.31	0.33	0.34	0.31
6	0.19	0.29	0.28	0.29	0.28	0.18	0.28	0.28	0.29	0.27
7	0.19	0.27	0.27	0.27	0.26	0.18	0.26	0.26	0.27	0.25
8	0.18	0.27	0.27	0.27	0.26	0.18	0.26	0.26	0.27	0.25
9	0.19	0.28	0.28	0.29	0.27	0.19	0.28	0.28	0.29	0.27
10	0.21	0.33	0.32	0.34	0.31	0.22	0.34	0.32	0.34	0.31
11	0.23	0.41	0.37	0.41	0.37	0.25	0.44	0.37	0.42	0.39
12	0.27	0.54	0.45	0.51	0.47	0.31	0.58	0.46	0.53	0.5
13	0.35	0.71	0.55	0.62	0.61	0.4	0.76	0.58	0.65	0.66
14	0.45	0.94	0.75	0.82	0.84	0.52	0.99	0.79	0.86	0.89
Score	Cell 6					Cell 16				
1	0.33	0.99	1.02	0.8	0.86	0.4	1.03	1.05	0.83	0.89
2	0.24	0.79	0.89	0.68	0.75	0.26	0.82	0.91	0.7	0.78
3	0.21	0.57	0.68	0.54	0.59	0.22	0.59	0.7	0.56	0.61
4	0.2	0.45	0.52	0.41	0.44	0.2	0.46	0.54	0.42	0.46
5	0.18	0.32	0.38	0.32	0.34	0.18	0.33	0.39	0.32	0.35
6	0.18	0.26	0.28	0.26	0.26	0.17	0.26	0.28	0.25	0.27
7	0.18	0.22	0.23	0.22	0.22	0.17	0.21	0.23	0.21	0.22
8	0.18	0.22	0.23	0.21	0.22	0.17	0.21	0.22	0.21	0.22
9	0.19	0.26	0.28	0.25	0.26	0.18	0.25	0.28	0.26	0.26
10	0.2	0.33	0.37	0.35	0.34	0.2	0.34	0.37	0.37	0.35
11	0.22	0.45	0.49	0.49	0.47	0.23	0.47	0.49	0.53	0.48
12	0.24	0.62	0.65	0.71	0.61	0.27	0.64	0.65	0.76	0.62
13	0.29	0.76	0.81	0.9	0.79	0.33	0.79	0.82	0.95	0.81
14	0.39	0.98	0.93	1.06	0.96	0.45	1.01	0.95	1.11	0.99

Table B.2. (Cont.)

	EAP					MAP				
	Cell 7					Cell 17				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.37	0.7	0.59	0.49	--	0.44	0.75	0.64	0.53	--
2	0.26	0.49	0.4	0.37	--	0.3	0.53	0.44	0.39	--
3	0.22	0.37	0.33	0.3	--	0.24	0.4	0.35	0.31	--
4	0.19	0.31	0.27	0.25	--	0.2	0.32	0.28	0.26	--
5	0.17	0.27	0.23	0.22	--	0.17	0.27	0.24	0.22	--
6	0.17	0.24	0.22	0.21	--	0.16	0.24	0.21	0.2	--
7	0.16	0.22	0.2	0.2	--	0.16	0.22	0.2	0.2	--
8	0.16	0.23	0.21	0.2	--	0.16	0.22	0.2	0.2	--
9	0.16	0.23	0.21	0.21	--	0.16	0.23	0.21	0.21	--
10	0.18	0.26	0.23	0.23	--	0.18	0.26	0.23	0.24	--
11	0.19	0.3	0.26	0.27	--	0.2	0.3	0.26	0.29	--
12	0.21	0.33	0.3	0.33	--	0.22	0.34	0.31	0.35	--
13	0.24	0.38	0.37	0.44	--	0.26	0.39	0.39	0.48	--
14	0.32	0.49	0.49	0.6	--	0.36	0.51	0.53	0.65	--
Score	Cell 8					Cell 18				
1	0.35	0.58	0.63	0.89	--	0.43	0.62	0.67	0.91	--
2	0.24	0.61	0.63	0.53	--	0.27	0.64	0.66	0.55	--
3	0.21	0.44	0.44	0.4	--	0.22	0.47	0.48	0.41	--
4	0.18	0.33	0.33	0.3	--	0.19	0.34	0.35	0.31	--
5	0.16	0.25	0.24	0.24	--	0.16	0.26	0.25	0.24	--
6	0.15	0.2	0.19	0.2	--	0.15	0.2	0.19	0.19	--
7	0.15	0.18	0.17	0.18	--	0.14	0.18	0.17	0.17	--
8	0.14	0.18	0.17	0.18	--	0.14	0.18	0.17	0.18	--
9	0.14	0.2	0.19	0.2	--	0.14	0.2	0.19	0.2	--
10	0.15	0.24	0.22	0.25	--	0.15	0.25	0.22	0.26	--
11	0.16	0.31	0.28	0.32	--	0.16	0.32	0.29	0.34	--
12	0.17	0.42	0.34	0.41	--	0.18	0.44	0.36	0.44	--
13	0.2	0.55	0.44	0.57	--	0.2	0.58	0.46	0.61	--
14	0.25	0.29	0.83	0.66	--	0.28	0.33	0.86	0.69	--

Table B.2. (Cont.)

	EAP					MAP				
	Cell 9					Cell 19				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.34	0.53	0.56	0.6	0.58	0.35	0.56	0.58	0.63	0.63
2	0.23	0.37	0.41	0.43	0.4	0.25	0.39	0.43	0.46	0.44
3	0.19	0.29	0.33	0.34	0.31	0.2	0.3	0.35	0.36	0.33
4	0.16	0.26	0.28	0.28	0.26	0.17	0.26	0.29	0.29	0.27
5	0.15	0.23	0.25	0.26	0.23	0.15	0.23	0.25	0.26	0.23
6	0.14	0.22	0.23	0.24	0.22	0.14	0.22	0.23	0.23	0.21
7	0.14	0.22	0.22	0.23	0.21	0.13	0.21	0.21	0.22	0.2
8	0.14	0.22	0.22	0.23	0.21	0.13	0.22	0.21	0.22	0.21
9	0.14	0.23	0.22	0.23	0.22	0.14	0.23	0.22	0.23	0.21
10	0.15	0.25	0.24	0.25	0.23	0.16	0.26	0.24	0.25	0.24
11	0.17	0.3	0.26	0.29	0.27	0.18	0.31	0.27	0.3	0.28
12	0.2	0.37	0.3	0.35	0.33	0.22	0.39	0.3	0.36	0.35
13	0.24	0.49	0.36	0.43	0.41	0.26	0.53	0.38	0.44	0.44
14	0.36	0.71	0.49	0.57	0.58	0.37	0.75	0.51	0.58	0.62
Score	Cell 10					Cell 20				
1	0.25	0.63	0.74	0.49	0.76	0.26	0.66	0.77	0.52	0.79
2	0.18	0.51	0.6	0.48	0.46	0.19	0.53	0.62	0.5	0.49
3	0.15	0.39	0.46	0.36	0.4	0.16	0.41	0.48	0.37	0.42
4	0.14	0.3	0.35	0.28	0.31	0.15	0.31	0.36	0.29	0.32
5	0.13	0.24	0.28	0.23	0.25	0.13	0.24	0.29	0.23	0.25
6	0.13	0.2	0.22	0.2	0.21	0.13	0.2	0.23	0.2	0.21
7	0.13	0.19	0.2	0.19	0.19	0.13	0.18	0.2	0.18	0.19
8	0.13	0.18	0.2	0.18	0.19	0.13	0.18	0.2	0.18	0.19
9	0.14	0.2	0.22	0.2	0.21	0.14	0.2	0.22	0.21	0.21
10	0.14	0.24	0.27	0.26	0.25	0.15	0.25	0.27	0.28	0.25
11	0.16	0.31	0.33	0.36	0.32	0.17	0.32	0.33	0.39	0.33
12	0.17	0.42	0.41	0.49	0.42	0.19	0.44	0.41	0.53	0.43
13	0.2	0.5	0.5	0.7	0.53	0.23	0.52	0.5	0.75	0.55
14	0.31	0.74	0.56	0.83	0.66	0.32	0.77	0.56	0.88	0.68

Table B.3. RMSE of Each Dimension Score for Cell 1 to Cell 20

	EAP					MAP				
	Cell 1					Cell 11				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.61	1.15	1.06	1.06	--	0.66	1.17	1.08	1.08	--
2	0.47	0.9	0.86	0.87	--	0.5	0.91	0.88	0.89	--
3	0.4	0.74	0.74	0.71	--	0.42	0.76	0.76	0.73	--
4	0.35	0.58	0.58	0.57	--	0.37	0.59	0.59	0.57	--
5	0.33	0.46	0.45	0.45	--	0.33	0.46	0.45	0.45	--
6	0.3	0.35	0.36	0.35	--	0.3	0.35	0.36	0.35	--
7	0.3	0.3	0.3	0.3	--	0.29	0.29	0.29	0.29	--
8	0.3	0.3	0.3	0.3	--	0.29	0.29	0.3	0.29	--
9	0.3	0.36	0.35	0.35	--	0.29	0.35	0.35	0.35	--
10	0.32	0.46	0.45	0.44	--	0.31	0.46	0.45	0.45	--
11	0.34	0.57	0.57	0.58	--	0.34	0.58	0.58	0.59	--
12	0.38	0.72	0.71	0.71	--	0.39	0.73	0.73	0.73	--
13	0.45	0.83	0.86	0.88	--	0.47	0.85	0.88	0.91	--
14	0.57	0.97	1.02	1.02	--	0.62	0.99	1.05	1.05	--
Score	Cell 2					Cell 12				
1	0.42	0.59	0.84	0.9	--	0.46	0.61	0.87	0.93	--
2	0.33	0.63	0.65	0.6	--	0.35	0.65	0.67	0.62	--
3	0.28	0.52	0.54	0.51	--	0.3	0.53	0.55	0.52	--
4	0.26	0.42	0.42	0.4	--	0.27	0.43	0.43	0.41	--
5	0.24	0.35	0.34	0.34	--	0.25	0.35	0.35	0.35	--
6	0.23	0.29	0.29	0.29	--	0.23	0.29	0.29	0.28	--
7	0.22	0.27	0.26	0.26	--	0.22	0.26	0.26	0.26	--
8	0.22	0.27	0.27	0.26	--	0.22	0.26	0.26	0.26	--
9	0.23	0.29	0.29	0.29	--	0.22	0.29	0.29	0.29	--
10	0.23	0.34	0.34	0.34	--	0.23	0.34	0.34	0.34	--
11	0.26	0.42	0.41	0.41	--	0.26	0.42	0.42	0.42	--
12	0.27	0.51	0.5	0.5	--	0.28	0.52	0.51	0.52	--
13	0.3	0.63	0.62	0.63	--	0.32	0.65	0.64	0.65	--
14	0.38	0.68	0.82	0.75	--	0.42	0.71	0.84	0.78	--

Table B.3. (Cont.)

	EAP					MAP				
	Cell 3					Cell 13				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.61	0.96	0.93	0.76	--	0.69	1.01	0.97	0.8	--
2	0.46	0.78	0.67	0.64	--	0.51	0.82	0.71	0.67	--
3	0.38	0.64	0.54	0.51	--	0.41	0.67	0.57	0.53	--
4	0.32	0.52	0.45	0.42	--	0.34	0.54	0.47	0.43	--
5	0.3	0.43	0.38	0.37	--	0.3	0.44	0.39	0.36	--
6	0.28	0.36	0.34	0.33	--	0.27	0.36	0.34	0.32	--
7	0.27	0.33	0.32	0.32	--	0.26	0.32	0.31	0.3	--
8	0.26	0.33	0.32	0.31	--	0.26	0.33	0.31	0.3	--
9	0.28	0.37	0.34	0.33	--	0.27	0.36	0.33	0.34	--
10	0.29	0.43	0.38	0.38	--	0.3	0.43	0.38	0.4	--
11	0.33	0.5	0.44	0.45	--	0.34	0.51	0.45	0.48	--
12	0.37	0.6	0.52	0.58	--	0.39	0.6	0.53	0.62	--
13	0.44	0.71	0.64	0.72	--	0.47	0.72	0.66	0.76	--
14	0.57	0.77	0.8	0.94	--	0.61	0.78	0.84	0.99	--
Score	Cell 4					Cell 14				
1	0.58	1.14	1.46	0.91	--	0.67	1.17	1.49	0.93	--
2	0.39	0.87	0.85	0.78	--	0.45	0.89	0.88	0.8	--
3	0.33	0.72	0.71	0.63	--	0.36	0.74	0.74	0.64	--
4	0.3	0.55	0.52	0.5	--	0.31	0.57	0.55	0.51	--
5	0.28	0.42	0.39	0.39	--	0.27	0.43	0.41	0.39	--
6	0.26	0.32	0.3	0.31	--	0.25	0.32	0.3	0.31	--
7	0.25	0.27	0.25	0.26	--	0.24	0.26	0.25	0.25	--
8	0.25	0.27	0.26	0.26	--	0.24	0.26	0.25	0.26	--
9	0.25	0.32	0.3	0.31	--	0.24	0.32	0.3	0.32	--
10	0.26	0.41	0.37	0.4	--	0.25	0.41	0.38	0.42	--
11	0.27	0.53	0.47	0.52	--	0.27	0.54	0.48	0.54	--
12	0.29	0.66	0.56	0.69	--	0.3	0.67	0.58	0.72	--
13	0.33	0.86	0.69	0.83	--	0.35	0.88	0.71	0.86	--
14	0.45	1.03	0.65	1.14	--	0.5	1.05	0.69	1.18	--

Table B.3. (Cont.)

	EAP					MAP				
	Cell 5					Cell 15				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.5	0.83	0.86	0.92	0.89	0.57	0.87	0.9	0.96	0.95
2	0.38	0.61	0.66	0.7	0.66	0.43	0.64	0.7	0.73	0.7
3	0.32	0.52	0.55	0.58	0.52	0.35	0.54	0.58	0.61	0.56
4	0.28	0.44	0.46	0.48	0.44	0.29	0.45	0.48	0.5	0.46
5	0.25	0.4	0.4	0.41	0.38	0.25	0.39	0.41	0.42	0.38
6	0.24	0.36	0.35	0.37	0.35	0.23	0.35	0.35	0.36	0.34
7	0.23	0.34	0.33	0.34	0.33	0.22	0.33	0.33	0.33	0.32
8	0.23	0.34	0.34	0.34	0.33	0.22	0.33	0.33	0.33	0.32
9	0.24	0.35	0.36	0.37	0.34	0.24	0.35	0.35	0.36	0.34
10	0.26	0.4	0.4	0.41	0.38	0.27	0.42	0.39	0.42	0.38
11	0.28	0.49	0.45	0.5	0.45	0.3	0.52	0.45	0.5	0.46
12	0.33	0.61	0.53	0.59	0.55	0.36	0.65	0.54	0.6	0.58
13	0.42	0.78	0.63	0.7	0.68	0.46	0.82	0.65	0.72	0.72
14	0.51	0.99	0.82	0.9	0.9	0.58	1.04	0.85	0.92	0.95
Score	Cell 6					Cell 16				
1	0.39	1.03	1.03	0.83	0.89	0.46	1.06	1.06	0.85	0.92
2	0.29	0.83	0.92	0.72	0.79	0.32	0.86	0.95	0.75	0.81
3	0.26	0.62	0.73	0.61	0.65	0.27	0.65	0.75	0.62	0.67
4	0.24	0.51	0.58	0.48	0.51	0.24	0.53	0.59	0.49	0.52
5	0.23	0.39	0.44	0.39	0.41	0.23	0.4	0.46	0.39	0.42
6	0.22	0.32	0.34	0.32	0.33	0.22	0.32	0.35	0.31	0.33
7	0.22	0.27	0.28	0.28	0.28	0.21	0.27	0.28	0.27	0.28
8	0.22	0.27	0.28	0.26	0.28	0.22	0.27	0.28	0.26	0.27
9	0.23	0.32	0.35	0.31	0.32	0.23	0.32	0.34	0.32	0.32
10	0.25	0.4	0.44	0.41	0.41	0.25	0.4	0.43	0.44	0.41
11	0.27	0.52	0.56	0.55	0.53	0.28	0.53	0.56	0.59	0.54
12	0.3	0.67	0.71	0.77	0.66	0.33	0.69	0.71	0.81	0.67
13	0.35	0.81	0.85	0.95	0.83	0.39	0.83	0.86	0.99	0.85
14	0.46	1.04	1.01	1.08	1	0.52	1.07	1.02	1.13	1.02

Table B.3. (Cont.)

	EAP					MAP				
	Cell 7					Cell 17				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.43	0.76	0.66	0.56	--	0.49	0.8	0.7	0.59	--
2	0.32	0.56	0.47	0.43	--	0.35	0.59	0.5	0.46	--
3	0.26	0.44	0.39	0.36	--	0.28	0.47	0.41	0.37	--
4	0.24	0.38	0.33	0.31	--	0.25	0.39	0.34	0.31	--
5	0.22	0.33	0.29	0.28	--	0.22	0.34	0.29	0.27	--
6	0.21	0.3	0.27	0.26	--	0.2	0.29	0.27	0.25	--
7	0.2	0.28	0.26	0.25	--	0.2	0.28	0.25	0.25	--
8	0.2	0.28	0.26	0.26	--	0.2	0.28	0.25	0.25	--
9	0.21	0.3	0.26	0.27	--	0.21	0.29	0.26	0.27	--
10	0.22	0.32	0.28	0.29	--	0.22	0.32	0.28	0.3	--
11	0.24	0.36	0.32	0.34	--	0.24	0.37	0.32	0.35	--
12	0.26	0.4	0.36	0.39	--	0.27	0.41	0.38	0.42	--
13	0.3	0.45	0.43	0.5	--	0.31	0.46	0.46	0.54	--
14	0.38	0.56	0.57	0.67	--	0.42	0.57	0.6	0.71	--
Score	Cell 8					Cell 18				
1	0.42	0.61	0.73	0.89	--	0.49	0.64	0.76	0.92	--
2	0.29	0.68	0.67	0.57	--	0.32	0.7	0.71	0.59	--
3	0.26	0.5	0.51	0.46	--	0.27	0.52	0.54	0.47	--
4	0.23	0.39	0.39	0.36	--	0.23	0.4	0.41	0.37	--
5	0.2	0.31	0.3	0.29	--	0.2	0.32	0.31	0.29	--
6	0.19	0.25	0.24	0.25	--	0.18	0.25	0.24	0.24	--
7	0.18	0.23	0.22	0.22	--	0.18	0.22	0.21	0.22	--
8	0.18	0.23	0.22	0.23	--	0.18	0.22	0.21	0.22	--
9	0.18	0.26	0.24	0.25	--	0.18	0.25	0.23	0.25	--
10	0.19	0.3	0.28	0.3	--	0.19	0.3	0.28	0.31	--
11	0.2	0.36	0.33	0.38	--	0.2	0.38	0.34	0.4	--
12	0.21	0.48	0.39	0.46	--	0.22	0.49	0.41	0.49	--
13	0.24	0.62	0.48	0.62	--	0.25	0.64	0.5	0.65	--
14	0.31	0.29	0.88	0.7	--	0.34	0.33	0.91	0.74	--

Table B.3. (Cont.)

	EAP					MAP				
	Cell 9					Cell 19				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.39	0.61	0.63	0.68	0.67	0.41	0.64	0.64	0.7	0.71
2	0.27	0.44	0.48	0.51	0.47	0.29	0.46	0.49	0.53	0.51
3	0.23	0.36	0.4	0.42	0.38	0.25	0.37	0.42	0.43	0.4
4	0.2	0.32	0.35	0.35	0.32	0.21	0.32	0.36	0.36	0.33
5	0.18	0.29	0.31	0.32	0.29	0.18	0.29	0.31	0.32	0.29
6	0.18	0.28	0.28	0.3	0.27	0.17	0.28	0.28	0.29	0.27
7	0.17	0.28	0.27	0.29	0.26	0.17	0.27	0.27	0.28	0.26
8	0.17	0.28	0.27	0.28	0.26	0.17	0.27	0.27	0.28	0.26
9	0.18	0.28	0.28	0.29	0.27	0.18	0.29	0.28	0.29	0.27
10	0.19	0.31	0.3	0.32	0.29	0.2	0.32	0.3	0.32	0.29
11	0.21	0.37	0.33	0.36	0.33	0.22	0.38	0.33	0.36	0.34
12	0.24	0.44	0.36	0.42	0.4	0.26	0.46	0.37	0.43	0.41
13	0.28	0.57	0.43	0.5	0.48	0.3	0.59	0.44	0.51	0.51
14	0.41	0.79	0.57	0.65	0.65	0.42	0.82	0.58	0.66	0.69
Score	Cell 10					Cell 20				
1	0.31	0.67	0.8	0.55	0.79	0.32	0.71	0.82	0.58	0.82
2	0.22	0.56	0.65	0.55	0.53	0.24	0.58	0.67	0.56	0.55
3	0.19	0.45	0.52	0.42	0.46	0.2	0.47	0.54	0.43	0.48
4	0.18	0.36	0.41	0.34	0.37	0.18	0.37	0.42	0.34	0.38
5	0.17	0.29	0.34	0.29	0.31	0.16	0.3	0.35	0.28	0.31
6	0.16	0.25	0.28	0.25	0.26	0.16	0.25	0.28	0.25	0.26
7	0.16	0.23	0.25	0.24	0.24	0.16	0.23	0.25	0.23	0.24
8	0.16	0.23	0.25	0.23	0.24	0.16	0.23	0.25	0.23	0.23
9	0.17	0.25	0.28	0.25	0.26	0.17	0.25	0.27	0.26	0.26
10	0.18	0.3	0.33	0.32	0.31	0.18	0.3	0.33	0.34	0.31
11	0.2	0.37	0.39	0.42	0.38	0.21	0.38	0.39	0.45	0.38
12	0.22	0.48	0.47	0.56	0.48	0.23	0.5	0.47	0.6	0.49
13	0.25	0.56	0.55	0.78	0.59	0.28	0.57	0.55	0.82	0.6
14	0.38	0.84	0.6	0.93	0.68	0.39	0.86	0.6	0.98	0.7

Table B.4. SE of Each Dimension Score for Cell 1 to Cell 20

	EAP					MAP				
	Cell 1					Cell 11				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.34	0.23	0.25	0.31	--	0.32	0.22	0.25	0.3	--
2	0.32	0.3	0.26	0.29	--	0.3	0.29	0.25	0.28	--
3	0.31	0.3	0.28	0.28	--	0.3	0.29	0.27	0.27	--
4	0.31	0.28	0.28	0.28	--	0.3	0.28	0.27	0.28	--
5	0.32	0.29	0.28	0.29	--	0.31	0.28	0.28	0.28	--
6	0.31	0.29	0.29	0.29	--	0.3	0.29	0.29	0.28	--
7	0.31	0.29	0.29	0.3	--	0.3	0.29	0.29	0.29	--
8	0.31	0.29	0.3	0.29	--	0.3	0.29	0.29	0.29	--
9	0.31	0.29	0.29	0.29	--	0.3	0.29	0.29	0.29	--
10	0.3	0.29	0.29	0.29	--	0.29	0.28	0.28	0.28	--
11	0.3	0.28	0.29	0.29	--	0.29	0.28	0.28	0.28	--
12	0.31	0.27	0.28	0.27	--	0.3	0.27	0.28	0.27	--
13	0.31	0.24	0.27	0.26	--	0.3	0.24	0.27	0.25	--
14	0.32	0.24	0.28	0.27	--	0.31	0.23	0.27	0.27	--
Score	Cell 2					Cell 12				
1	0.31	0.22	0.33	0.3	--	0.3	0.22	0.32	0.3	--
2	0.27	0.25	0.25	0.25	--	0.26	0.25	0.25	0.24	--
3	0.26	0.26	0.27	0.27	--	0.25	0.26	0.27	0.26	--
4	0.26	0.26	0.26	0.26	--	0.25	0.26	0.26	0.26	--
5	0.25	0.27	0.27	0.27	--	0.25	0.26	0.26	0.26	--
6	0.24	0.27	0.26	0.27	--	0.24	0.26	0.26	0.26	--
7	0.24	0.27	0.27	0.27	--	0.24	0.27	0.26	0.26	--
8	0.24	0.27	0.27	0.27	--	0.24	0.27	0.27	0.26	--
9	0.24	0.27	0.27	0.27	--	0.23	0.26	0.26	0.26	--
10	0.24	0.27	0.27	0.26	--	0.24	0.26	0.26	0.26	--
11	0.25	0.27	0.27	0.27	--	0.25	0.26	0.26	0.26	--
12	0.25	0.26	0.26	0.26	--	0.24	0.26	0.25	0.26	--
13	0.25	0.26	0.26	0.26	--	0.24	0.25	0.26	0.25	--
14	0.32	0.24	0.25	0.22	--	0.3	0.24	0.25	0.21	--

Table B.4. (Cont.)

	EAP					MAP				
	Cell 3					Cell 13				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.27	0.33	0.34	0.28	--	0.25	0.31	0.33	0.28	--
2	0.26	0.32	0.31	0.3	--	0.24	0.31	0.3	0.3	--
3	0.28	0.32	0.3	0.31	--	0.26	0.31	0.29	0.3	--
4	0.28	0.33	0.31	0.3	--	0.27	0.32	0.3	0.29	--
5	0.29	0.33	0.31	0.31	--	0.27	0.32	0.3	0.3	--
6	0.28	0.33	0.32	0.32	--	0.27	0.32	0.31	0.3	--
7	0.28	0.33	0.32	0.32	--	0.27	0.32	0.31	0.31	--
8	0.27	0.33	0.32	0.32	--	0.27	0.32	0.31	0.31	--
9	0.28	0.33	0.32	0.32	--	0.27	0.32	0.31	0.31	--
10	0.28	0.33	0.31	0.31	--	0.27	0.33	0.3	0.3	--
11	0.28	0.33	0.31	0.31	--	0.27	0.32	0.3	0.3	--
12	0.28	0.33	0.3	0.3	--	0.28	0.33	0.29	0.3	--
13	0.29	0.33	0.31	0.31	--	0.28	0.33	0.3	0.3	--
14	0.32	0.32	0.3	0.31	--	0.3	0.32	0.29	0.3	--
Score	Cell 4					Cell 14				
1	0.28	0.19	0.4	0.2	--	0.26	0.19	0.4	0.19	--
2	0.29	0.24	0.26	0.27	--	0.27	0.24	0.26	0.26	--
3	0.3	0.26	0.26	0.27	--	0.28	0.25	0.26	0.26	--
4	0.29	0.25	0.25	0.25	--	0.27	0.24	0.25	0.25	--
5	0.28	0.26	0.25	0.25	--	0.27	0.25	0.24	0.25	--
6	0.27	0.26	0.25	0.26	--	0.26	0.25	0.24	0.25	--
7	0.27	0.26	0.25	0.26	--	0.25	0.26	0.25	0.25	--
8	0.26	0.26	0.26	0.26	--	0.26	0.25	0.25	0.26	--
9	0.26	0.26	0.26	0.26	--	0.25	0.25	0.25	0.25	--
10	0.26	0.26	0.26	0.26	--	0.26	0.26	0.25	0.25	--
11	0.27	0.26	0.26	0.25	--	0.26	0.25	0.25	0.25	--
12	0.27	0.25	0.26	0.25	--	0.26	0.24	0.25	0.25	--
13	0.28	0.26	0.25	0.23	--	0.26	0.26	0.25	0.22	--
14	0.32	0.26	0.07	0.2	--	0.29	0.26	0.07	0.2	--

Table B.4. (Cont.)

	EAP					MAP				
	Cell 5					Cell 15				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.25	0.32	0.32	0.33	0.32	0.24	0.31	0.32	0.32	0.31
2	0.26	0.33	0.32	0.31	0.31	0.24	0.31	0.31	0.3	0.3
3	0.25	0.32	0.32	0.32	0.31	0.24	0.3	0.31	0.31	0.29
4	0.25	0.33	0.32	0.33	0.32	0.24	0.31	0.32	0.32	0.31
5	0.24	0.34	0.33	0.34	0.33	0.23	0.32	0.33	0.33	0.31
6	0.25	0.34	0.33	0.34	0.33	0.23	0.33	0.32	0.33	0.32
7	0.25	0.34	0.34	0.35	0.34	0.24	0.33	0.33	0.34	0.32
8	0.24	0.34	0.34	0.35	0.33	0.24	0.33	0.33	0.34	0.32
9	0.25	0.33	0.34	0.34	0.33	0.24	0.32	0.33	0.33	0.32
10	0.26	0.33	0.34	0.34	0.33	0.25	0.32	0.33	0.33	0.31
11	0.25	0.33	0.34	0.34	0.32	0.24	0.32	0.33	0.33	0.31
12	0.26	0.32	0.33	0.32	0.32	0.24	0.31	0.31	0.32	0.31
13	0.27	0.31	0.32	0.33	0.31	0.26	0.3	0.31	0.32	0.3
14	0.24	0.31	0.34	0.35	0.33	0.23	0.3	0.33	0.34	0.32
Score	Cell 6					Cell 16				
1	0.29	0.19	0.19	0.21	0.23	0.27	0.19	0.19	0.21	0.23
2	0.26	0.26	0.26	0.26	0.24	0.24	0.26	0.26	0.25	0.24
3	0.26	0.27	0.28	0.27	0.28	0.24	0.26	0.27	0.27	0.27
4	0.25	0.27	0.27	0.28	0.26	0.23	0.26	0.26	0.27	0.26
5	0.24	0.27	0.27	0.28	0.27	0.23	0.26	0.27	0.27	0.27
6	0.24	0.27	0.28	0.28	0.27	0.23	0.26	0.27	0.27	0.27
7	0.24	0.27	0.28	0.28	0.28	0.23	0.27	0.27	0.26	0.27
8	0.24	0.27	0.28	0.27	0.28	0.23	0.26	0.27	0.25	0.27
9	0.25	0.27	0.28	0.26	0.27	0.24	0.26	0.28	0.25	0.27
10	0.25	0.27	0.28	0.26	0.27	0.24	0.26	0.27	0.25	0.26
11	0.26	0.26	0.27	0.26	0.27	0.25	0.25	0.27	0.25	0.26
12	0.27	0.27	0.28	0.28	0.26	0.26	0.27	0.27	0.28	0.25
13	0.28	0.25	0.26	0.29	0.27	0.27	0.24	0.25	0.29	0.26
14	0.31	0.33	0.37	0.27	0.26	0.29	0.32	0.36	0.26	0.26

Table B.4. (Cont.)

	EAP					MAP				
	Cell 7					Cell 17				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.28	0.29	0.29	0.28	--	0.26	0.28	0.29	0.28	--
2	0.24	0.3	0.28	0.27	--	0.23	0.29	0.27	0.27	--
3	0.23	0.28	0.27	0.26	--	0.22	0.28	0.26	0.26	--
4	0.23	0.29	0.27	0.26	--	0.22	0.29	0.26	0.26	--
5	0.23	0.29	0.27	0.26	--	0.22	0.28	0.26	0.25	--
6	0.22	0.29	0.27	0.26	--	0.22	0.28	0.27	0.26	--
7	0.22	0.29	0.27	0.27	--	0.22	0.28	0.26	0.26	--
8	0.22	0.29	0.27	0.27	--	0.22	0.28	0.26	0.26	--
9	0.22	0.29	0.27	0.27	--	0.22	0.28	0.26	0.26	--
10	0.22	0.29	0.27	0.27	--	0.22	0.28	0.26	0.26	--
11	0.23	0.28	0.27	0.27	--	0.23	0.28	0.26	0.27	--
12	0.24	0.28	0.27	0.27	--	0.23	0.28	0.26	0.27	--
13	0.24	0.29	0.27	0.28	--	0.23	0.29	0.27	0.28	--
14	0.28	0.28	0.28	0.31	--	0.27	0.28	0.27	0.3	--
Score	Cell 8					Cell 18				
1	0.3	0.1	0.31	0.14	--	0.29	0.1	0.31	0.15	--
2	0.26	0.27	0.24	0.23	--	0.24	0.27	0.24	0.23	--
3	0.25	0.24	0.26	0.25	--	0.24	0.23	0.26	0.24	--
4	0.24	0.24	0.24	0.24	--	0.22	0.23	0.23	0.23	--
5	0.22	0.24	0.23	0.23	--	0.21	0.23	0.22	0.22	--
6	0.21	0.23	0.23	0.23	--	0.2	0.23	0.22	0.23	--
7	0.21	0.23	0.23	0.23	--	0.2	0.23	0.22	0.23	--
8	0.2	0.23	0.23	0.23	--	0.2	0.23	0.22	0.23	--
9	0.2	0.24	0.23	0.24	--	0.2	0.23	0.22	0.23	--
10	0.21	0.23	0.23	0.23	--	0.2	0.23	0.22	0.23	--
11	0.21	0.24	0.23	0.23	--	0.21	0.23	0.22	0.23	--
12	0.22	0.23	0.22	0.23	--	0.21	0.23	0.21	0.23	--
13	0.23	0.28	0.19	0.22	--	0.22	0.27	0.19	0.22	--
14	0.31	0	0.23	0.17	--	0.29	0	0.23	0.16	--

Table B.4. (Cont.)

	EAP					MAP				
	Cell 9					Cell 19				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.22	0.3	0.29	0.3	0.32	0.24	0.29	0.28	0.29	0.31
2	0.2	0.29	0.29	0.3	0.29	0.2	0.28	0.28	0.29	0.28
3	0.2	0.28	0.29	0.3	0.28	0.2	0.27	0.28	0.29	0.27
4	0.2	0.28	0.29	0.29	0.27	0.19	0.27	0.29	0.28	0.26
5	0.2	0.28	0.29	0.29	0.28	0.19	0.27	0.28	0.29	0.27
6	0.19	0.28	0.28	0.3	0.28	0.19	0.28	0.28	0.29	0.27
7	0.2	0.29	0.28	0.3	0.28	0.19	0.28	0.28	0.29	0.27
8	0.2	0.29	0.28	0.3	0.28	0.19	0.28	0.28	0.29	0.27
9	0.2	0.29	0.28	0.29	0.27	0.19	0.28	0.28	0.29	0.27
10	0.2	0.29	0.28	0.29	0.28	0.2	0.28	0.28	0.29	0.27
11	0.21	0.29	0.28	0.29	0.27	0.2	0.28	0.27	0.29	0.27
12	0.21	0.29	0.28	0.29	0.28	0.21	0.28	0.27	0.29	0.27
13	0.2	0.3	0.27	0.3	0.29	0.21	0.29	0.26	0.29	0.28
14	0.21	0.34	0.31	0.32	0.32	0.23	0.33	0.3	0.31	0.31
Score	Cell 10					Cell 20				
1	0.23	0.26	0.32	0.24	0.22	0.27	0.25	0.32	0.24	0.22
2	0.2	0.26	0.27	0.28	0.27	0.2	0.25	0.26	0.27	0.26
3	0.2	0.25	0.26	0.24	0.25	0.2	0.25	0.26	0.23	0.24
4	0.2	0.24	0.26	0.24	0.25	0.19	0.24	0.25	0.23	0.25
5	0.19	0.24	0.25	0.24	0.24	0.18	0.23	0.25	0.23	0.24
6	0.19	0.24	0.25	0.24	0.25	0.18	0.24	0.25	0.24	0.24
7	0.19	0.24	0.25	0.24	0.25	0.18	0.24	0.25	0.24	0.24
8	0.19	0.24	0.26	0.24	0.25	0.18	0.24	0.25	0.24	0.24
9	0.19	0.24	0.25	0.24	0.24	0.19	0.24	0.25	0.23	0.24
10	0.2	0.24	0.26	0.25	0.24	0.2	0.23	0.25	0.24	0.24
11	0.21	0.24	0.25	0.25	0.24	0.21	0.24	0.25	0.25	0.24
12	0.22	0.25	0.26	0.28	0.25	0.21	0.24	0.25	0.28	0.24
13	0.22	0.25	0.24	0.34	0.26	0.23	0.24	0.24	0.33	0.25
14	0.25	0.38	0.15	0.39	0.19	0.3	0.37	0.15	0.39	0.19

Table B.5. Average MSE of Each Dimension Score for Cell 1 to Cell 20

	EAP					MAP				
	Cell 1					Cell 11				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.16	0.15	0.16	0.15	--	0.15	0.15	0.15	0.15	--
2	0.14	0.15	0.15	0.15	--	0.13	0.15	0.15	0.14	--
3	0.13	0.15	0.15	0.15	--	0.12	0.15	0.15	0.14	--
4	0.12	0.15	0.15	0.14	--	0.11	0.15	0.15	0.14	--
5	0.11	0.15	0.15	0.14	--	0.11	0.14	0.14	0.14	--
6	0.11	0.15	0.15	0.14	--	0.1	0.14	0.14	0.14	--
7	0.1	0.15	0.14	0.14	--	0.1	0.14	0.14	0.14	--
8	0.1	0.15	0.14	0.14	--	0.1	0.14	0.14	0.14	--
9	0.1	0.15	0.14	0.14	--	0.1	0.14	0.14	0.14	--
10	0.1	0.15	0.14	0.14	--	0.1	0.14	0.14	0.14	--
11	0.11	0.15	0.14	0.14	--	0.1	0.14	0.14	0.14	--
12	0.12	0.15	0.15	0.15	--	0.11	0.14	0.14	0.14	--
13	0.12	0.15	0.15	0.15	--	0.12	0.15	0.14	0.14	--
14	0.14	0.15	0.15	0.15	--	0.13	0.15	0.14	0.15	--
Score	Cell 2					Cell 12				
1	0.1	0.11	0.11	0.1	--	0.09	0.11	0.11	0.1	--
2	0.08	0.1	0.11	0.1	--	0.08	0.1	0.1	0.1	--
3	0.07	0.1	0.1	0.1	--	0.07	0.1	0.1	0.1	--
4	0.07	0.1	0.1	0.09	--	0.07	0.1	0.1	0.09	--
5	0.06	0.1	0.1	0.09	--	0.06	0.09	0.09	0.09	--
6	0.06	0.09	0.09	0.09	--	0.06	0.09	0.09	0.09	--
7	0.05	0.09	0.09	0.09	--	0.05	0.09	0.09	0.09	--
8	0.05	0.09	0.09	0.09	--	0.05	0.09	0.09	0.09	--
9	0.05	0.09	0.09	0.09	--	0.05	0.09	0.09	0.09	--
10	0.06	0.09	0.09	0.09	--	0.05	0.09	0.09	0.09	--
11	0.06	0.1	0.09	0.09	--	0.06	0.09	0.09	0.09	--
12	0.07	0.1	0.1	0.1	--	0.06	0.1	0.09	0.1	--
13	0.07	0.1	0.1	0.1	--	0.07	0.1	0.1	0.1	--
14	0.09	0.1	0.1	0.11	--	0.08	0.1	0.1	0.1	--

Table B.5. (Cont.)

	EAP					MAP				
	Cell 3					Cell 13				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.13	0.18	0.16	0.14	--	0.12	0.18	0.15	0.14	--
2	0.11	0.18	0.15	0.14	--	0.1	0.17	0.15	0.13	--
3	0.1	0.17	0.15	0.13	--	0.1	0.17	0.14	0.13	--
4	0.09	0.17	0.14	0.13	--	0.09	0.16	0.14	0.12	--
5	0.09	0.17	0.14	0.13	--	0.08	0.16	0.13	0.12	--
6	0.08	0.16	0.13	0.13	--	0.08	0.16	0.13	0.12	--
7	0.08	0.16	0.13	0.13	--	0.08	0.16	0.13	0.12	--
8	0.08	0.16	0.13	0.13	--	0.08	0.16	0.13	0.13	--
9	0.08	0.16	0.13	0.13	--	0.08	0.16	0.13	0.13	--
10	0.09	0.16	0.13	0.14	--	0.09	0.16	0.13	0.14	--
11	0.09	0.16	0.13	0.15	--	0.09	0.16	0.13	0.14	--
12	0.1	0.16	0.14	0.15	--	0.09	0.16	0.13	0.15	--
13	0.1	0.16	0.14	0.16	--	0.1	0.16	0.13	0.15	--
14	0.11	0.16	0.14	0.17	--	0.11	0.15	0.14	0.16	--
Score	Cell 4					Cell 14				
1	0.15	0.12	0.13	0.1	--	0.15	0.12	0.12	0.09	--
2	0.13	0.11	0.11	0.1	--	0.12	0.11	0.1	0.1	--
3	0.11	0.11	0.11	0.1	--	0.1	0.11	0.1	0.1	--
4	0.09	0.11	0.1	0.1	--	0.09	0.11	0.1	0.1	--
5	0.08	0.11	0.1	0.1	--	0.08	0.11	0.1	0.1	--
6	0.07	0.11	0.1	0.1	--	0.07	0.11	0.1	0.1	--
7	0.07	0.11	0.1	0.1	--	0.06	0.1	0.09	0.1	--
8	0.06	0.11	0.09	0.1	--	0.06	0.1	0.09	0.1	--
9	0.07	0.11	0.09	0.1	--	0.06	0.1	0.09	0.1	--
10	0.07	0.11	0.09	0.11	--	0.07	0.1	0.09	0.1	--
11	0.07	0.11	0.09	0.11	--	0.07	0.1	0.09	0.11	--
12	0.08	0.11	0.09	0.11	--	0.08	0.1	0.09	0.11	--
13	0.09	0.11	0.1	0.11	--	0.09	0.11	0.09	0.11	--
14	0.11	0.11	0.09	0.12	--	0.1	0.11	0.09	0.11	--

Table B.5. (Cont.)

	EAP					MAP				
	Cell 5					Cell 15				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.09	0.16	0.18	0.18	0.17	0.09	0.15	0.17	0.18	0.16
2	0.08	0.15	0.17	0.18	0.16	0.08	0.15	0.17	0.17	0.15
3	0.07	0.15	0.17	0.17	0.16	0.07	0.14	0.16	0.17	0.15
4	0.07	0.15	0.16	0.17	0.15	0.06	0.14	0.16	0.16	0.14
5	0.06	0.14	0.16	0.17	0.14	0.06	0.14	0.15	0.16	0.14
6	0.06	0.15	0.15	0.16	0.14	0.06	0.14	0.15	0.16	0.13
7	0.06	0.15	0.15	0.16	0.14	0.06	0.14	0.15	0.16	0.13
8	0.06	0.15	0.15	0.16	0.14	0.06	0.15	0.14	0.16	0.13
9	0.06	0.16	0.15	0.16	0.14	0.06	0.16	0.14	0.16	0.14
10	0.07	0.17	0.15	0.17	0.15	0.07	0.17	0.14	0.16	0.14
11	0.07	0.18	0.15	0.17	0.16	0.07	0.17	0.14	0.17	0.15
12	0.08	0.18	0.15	0.18	0.16	0.08	0.18	0.15	0.17	0.16
13	0.09	0.19	0.16	0.18	0.17	0.08	0.19	0.15	0.18	0.16
14	0.1	0.2	0.16	0.18	0.18	0.09	0.2	0.15	0.18	0.17
Score	Cell 6					Cell 16				
1	0.1	0.12	0.14	0.1	0.12	0.09	0.11	0.14	0.1	0.12
2	0.08	0.11	0.14	0.11	0.12	0.07	0.11	0.14	0.1	0.11
3	0.07	0.11	0.14	0.11	0.12	0.06	0.11	0.13	0.1	0.11
4	0.06	0.11	0.14	0.11	0.12	0.06	0.11	0.13	0.1	0.11
5	0.05	0.11	0.13	0.11	0.12	0.05	0.11	0.13	0.1	0.11
6	0.05	0.11	0.13	0.11	0.11	0.05	0.11	0.13	0.1	0.11
7	0.05	0.11	0.13	0.11	0.12	0.05	0.11	0.13	0.11	0.11
8	0.05	0.11	0.13	0.11	0.12	0.05	0.11	0.13	0.11	0.11
9	0.06	0.11	0.13	0.12	0.12	0.06	0.11	0.13	0.11	0.11
10	0.07	0.11	0.13	0.12	0.12	0.06	0.11	0.13	0.12	0.12
11	0.07	0.12	0.13	0.13	0.12	0.07	0.11	0.13	0.12	0.12
12	0.08	0.12	0.13	0.13	0.12	0.08	0.12	0.13	0.13	0.12
13	0.1	0.12	0.13	0.13	0.12	0.09	0.12	0.13	0.12	0.12
14	0.12	0.13	0.13	0.12	0.12	0.11	0.13	0.13	0.12	0.12

Table B.5. (Cont.)

	EAP					MAP				
	Cell 7					Cell 17				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.08	0.13	0.11	0.09	--	0.08	0.12	0.1	0.09	--
2	0.07	0.12	0.1	0.09	--	0.07	0.12	0.1	0.09	--
3	0.06	0.12	0.09	0.08	--	0.06	0.11	0.09	0.08	--
4	0.05	0.11	0.09	0.08	--	0.05	0.11	0.08	0.08	--
5	0.05	0.11	0.08	0.07	--	0.05	0.1	0.08	0.07	--
6	0.05	0.1	0.08	0.07	--	0.04	0.1	0.08	0.07	--
7	0.04	0.1	0.08	0.07	--	0.04	0.1	0.08	0.07	--
8	0.04	0.1	0.08	0.08	--	0.04	0.1	0.07	0.07	--
9	0.05	0.1	0.08	0.08	--	0.05	0.1	0.07	0.08	--
10	0.05	0.1	0.08	0.09	--	0.05	0.1	0.08	0.08	--
11	0.05	0.1	0.08	0.09	--	0.05	0.1	0.08	0.09	--
12	0.06	0.1	0.08	0.1	--	0.05	0.1	0.08	0.1	--
13	0.06	0.1	0.09	0.11	--	0.06	0.09	0.09	0.1	--
14	0.07	0.1	0.1	0.12	--	0.07	0.09	0.09	0.11	--
Score	Cell 8					Cell 18				
1	0.11	0.07	0.08	0.07	--	0.1	0.07	0.08	0.07	--
2	0.08	0.08	0.08	0.07	--	0.08	0.08	0.08	0.07	--
3	0.07	0.08	0.08	0.07	--	0.06	0.08	0.07	0.07	--
4	0.05	0.08	0.07	0.07	--	0.05	0.07	0.07	0.06	--
5	0.04	0.07	0.07	0.06	--	0.04	0.07	0.07	0.06	--
6	0.04	0.07	0.06	0.06	--	0.04	0.07	0.06	0.06	--
7	0.03	0.07	0.06	0.06	--	0.03	0.07	0.06	0.06	--
8	0.03	0.07	0.06	0.06	--	0.03	0.07	0.06	0.06	--
9	0.03	0.07	0.06	0.07	--	0.03	0.06	0.06	0.06	--
10	0.04	0.07	0.06	0.07	--	0.04	0.07	0.06	0.07	--
11	0.04	0.07	0.06	0.07	--	0.04	0.07	0.06	0.07	--
12	0.04	0.07	0.06	0.07	--	0.04	0.07	0.06	0.07	--
13	0.05	0.08	0.06	0.08	--	0.05	0.07	0.06	0.08	--
14	0.07	0.08	0.06	0.08	--	0.06	0.07	0.06	0.07	--

Table B.5. (Cont.)

	EAP					MAP				
	Cell 9					Cell 19				
Score	G	S1	S2	S3	S4	G	S1	S2	S3	S4
1	0.05	0.1	0.11	0.12	0.11	0.06	0.1	0.12	0.12	0.12
2	0.04	0.09	0.11	0.11	0.1	0.05	0.1	0.11	0.11	0.1
3	0.04	0.09	0.1	0.11	0.1	0.04	0.09	0.1	0.11	0.1
4	0.04	0.08	0.1	0.1	0.09	0.04	0.08	0.1	0.1	0.09
5	0.03	0.08	0.09	0.1	0.08	0.03	0.08	0.09	0.1	0.08
6	0.03	0.08	0.09	0.1	0.08	0.03	0.08	0.09	0.09	0.08
7	0.03	0.09	0.09	0.1	0.08	0.03	0.08	0.09	0.09	0.08
8	0.03	0.09	0.09	0.1	0.08	0.03	0.09	0.08	0.09	0.08
9	0.03	0.1	0.08	0.1	0.08	0.03	0.1	0.08	0.1	0.08
10	0.04	0.1	0.09	0.1	0.09	0.04	0.1	0.08	0.1	0.09
11	0.04	0.11	0.09	0.1	0.1	0.04	0.11	0.09	0.1	0.09
12	0.04	0.12	0.09	0.11	0.1	0.05	0.12	0.09	0.11	0.1
13	0.05	0.13	0.09	0.11	0.11	0.05	0.13	0.09	0.11	0.11
14	0.05	0.13	0.1	0.11	0.11	0.06	0.14	0.1	0.12	0.12
Score	Cell 10					Cell 20				
1	0.04	0.08	0.09	0.08	0.08	0.06	0.08	0.09	0.08	0.08
2	0.04	0.07	0.09	0.07	0.08	0.04	0.07	0.09	0.07	0.08
3	0.04	0.07	0.09	0.07	0.08	0.03	0.07	0.09	0.07	0.07
4	0.03	0.07	0.09	0.07	0.07	0.03	0.07	0.09	0.07	0.07
5	0.03	0.07	0.09	0.06	0.07	0.03	0.07	0.09	0.06	0.07
6	0.03	0.07	0.08	0.07	0.07	0.03	0.07	0.08	0.06	0.07
7	0.03	0.07	0.08	0.07	0.07	0.03	0.07	0.08	0.07	0.07
8	0.03	0.07	0.08	0.07	0.07	0.03	0.07	0.08	0.07	0.07
9	0.03	0.07	0.08	0.08	0.07	0.03	0.07	0.08	0.08	0.07
10	0.03	0.07	0.08	0.08	0.08	0.03	0.07	0.08	0.08	0.07
11	0.04	0.08	0.08	0.09	0.08	0.04	0.08	0.08	0.09	0.08
12	0.05	0.08	0.08	0.09	0.08	0.05	0.08	0.08	0.09	0.08
13	0.05	0.08	0.08	0.1	0.08	0.06	0.08	0.08	0.09	0.08
14	0.05	0.08	0.08	0.09	0.09	0.07	0.08	0.09	0.09	0.09

Table B.6. Reliability and MSE of Each Dimension Score for Cell 1 to Cell 20

Dimension	Reliability	MSE	Reliability	MSE
	Cell 1		Cell 11	
G	0.84	0.11	0.84	0.1
S1	0.56	0.15	0.56	0.14
S2	0.57	0.14	0.56	0.14
S3	0.57	0.14	0.57	0.14
	Cell 2		Cell 12	
G	0.91	0.06	0.91	0.06
S1	0.72	0.09	0.71	0.09
S2	0.72	0.09	0.72	0.09
S3	0.72	0.09	0.72	0.09
	Cell 3		Cell 13	
G	0.84	0.09	0.83	0.08
S1	0.65	0.16	0.64	0.16
S2	0.71	0.13	0.71	0.13
S3	0.72	0.13	0.71	0.13
	Cell 4		Cell 14	
G	0.9	0.08	0.89	0.07
S1	0.6	0.11	0.59	0.1
S2	0.64	0.1	0.63	0.09
S3	0.61	0.1	0.61	0.1
	Cell 5		Cell 15	
G	0.87	0.06	0.86	0.06
S1	0.7	0.16	0.69	0.15
S2	0.71	0.15	0.71	0.15
S3	0.68	0.17	0.68	0.16
S4	0.72	0.14	0.72	0.14
	Cell 6		Cell 16	
G	0.91	0.06	0.91	0.06
S1	0.63	0.11	0.62	0.11
S2	0.56	0.13	0.56	0.13
S3	0.62	0.11	0.61	0.11
S4	0.61	0.12	0.61	0.11

Table B.6. (Cont.)

Dimension	Reliability	MSE	Reliability	MSE
	Cell 7		Cell 17	
G	0.91	0.05	0.91	0.05
S1	0.78	0.1	0.78	0.1
S2	0.83	0.08	0.83	0.08
S3	0.83	0.08	0.83	0.08
	Cell 8		Cell 18	
G	0.94	0.04	0.94	0.04
S1	0.74	0.07	0.74	0.07
S2	0.77	0.06	0.77	0.06
S3	0.76	0.06	0.75	0.06
	Cell 9		Cell 19	
G	0.93	0.03	0.93	0.03
S1	0.82	0.09	0.82	0.09
S2	0.83	0.09	0.83	0.09
S3	0.81	0.1	0.81	0.1
S4	0.84	0.08	0.83	0.08
	Cell 10		Cell 20	
G	0.95	0.03	0.95	0.03
S1	0.77	0.07	0.76	0.07
S2	0.72	0.08	0.72	0.08
S3	0.76	0.07	0.75	0.07
S4	0.76	0.07	0.75	0.07

Table B.7. Sensitivity for the Strength and Weakness and Specificity of Each Dimension Score for Cell 1 to Cell 20

	Sensitivity			Sensitivity		
Dimension	Strength	Weakness	Specificity	Strength	Weakness	Specificity
	Cell 1			Cell 11		
S1	0.39	0.42	0.64	0.38	0.41	0.65
S2	0.4	0.42	0.63	0.4	0.41	0.64
S3	0.4	0.42	0.62	0.39	0.42	0.63
	Cell 2			Cell 12		
S1	0.54	0.56	0.49	0.53	0.55	0.5
S2	0.53	0.56	0.49	0.53	0.56	0.49
S3	0.54	0.55	0.48	0.54	0.55	0.49
	Cell 3			Cell 13		
S1	0.58	0.57	0.59	0.58	0.56	0.6
S2	0.6	0.59	0.51	0.59	0.58	0.51
S3	0.61	0.6	0.5	0.59	0.61	0.51
	Cell 4			Cell 14		
S1	0.39	0.43	0.59	0.38	0.42	0.6
S2	0.4	0.45	0.54	0.4	0.44	0.55
S3	0.4	0.44	0.58	0.38	0.44	0.58
	Cell 5			Cell 15		
S1	0.66	0.61	0.52	0.64	0.62	0.53
S2	0.63	0.64	0.53	0.63	0.62	0.53
S3	0.63	0.61	0.55	0.63	0.61	0.56
S4	0.67	0.64	0.49	0.66	0.64	0.5
	Cell 6			Cell 16		
S1	0.47	0.43	0.56	0.46	0.42	0.57
S2	0.41	0.4	0.64	0.41	0.38	0.64
S3	0.48	0.4	0.56	0.47	0.42	0.58
S4	0.45	0.42	0.59	0.44	0.4	0.59
	Cell 7			Cell 17		
S1	0.7	0.69	0.45	0.7	0.68	0.45
S2	0.72	0.71	0.39	0.71	0.71	0.39
S3	0.72	0.71	0.37	0.71	0.72	0.38
	Cell 8			Cell 18		
S1	0.51	0.56	0.45	0.51	0.55	0.45
S2	0.52	0.56	0.41	0.53	0.56	0.41
S3	0.53	0.55	0.43	0.51	0.55	0.44

Table B.7. (Cont.)

	Sensitivity			Sensitivity		
Dimension	Strength	Weakness	Specificity	Strength	Weakness	Specificity
	Cell 9			Cell 19		
S1	0.76	0.73	0.39	0.75	0.73	0.4
S2	0.75	0.76	0.4	0.75	0.75	0.4
S3	0.75	0.73	0.42	0.75	0.73	0.42
S4	0.77	0.75	0.37	0.77	0.75	0.38
	Cell 10			Cell 20		
S1	0.58	0.54	0.42	0.58	0.54	0.43
S2	0.55	0.53	0.49	0.54	0.51	0.49
S3	0.60	0.52	0.42	0.60	0.54	0.43
S4	0.58	0.55	0.45	0.57	0.54	0.45