ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(1); 61-75

Simultaneous Estimation of Overall Score and Subscores Using MIRT, HO-IRT and Bi-factor Model on TIMSS Data

Ayşenur ERDEMİR * Hakan Yavuz ATAR **

Abstract

In educational testing, there is an increasing interest in the simultaneous estimation of the overall scores and subscores. This study aims to compare the reliability and precision of the simultaneous estimation of overall scores and sub-scores using MIRT, HO-IRT and Bi-factor models. TIMSS 2015 mathematics scores have been used as a data set in this study. The TIMSS 2015 mathematics test consists of 35 items, four of which are polytomously scored (0-1-2), and the rest of the items are dichotomously scored (0-1). The four content domains include number (14 items), algebra (9 items), geometry (6 items), and data and change (6 items). Ability parameters were estimated using the BMIRT software. The results showed that the MIRT and HO-IRT methods performed similarly in terms of precision and reliability for subscore estimates. The MIRT maximum information method had the smallest standard error of measurement for the overall score estimates. All three methods performed similarly in terms of the overall score reliability. The findings suggest that among the three methods compared, HO-IRT appears to be a better choice in the simultaneous estimation of the overall score and subscores for the data from TIMSS 2015. Recommendations for the testing practices and future research are provided.

Key Words: TIMSS, subscores, multidimensional item response theory, higher-order item response theory, bi-factor model.

INTRODUCTION

Many tests in educational and psychological testing generally measure more than one ability, which makes them multidimensional inherently (Reckase, 1985; 1997). Tests may be inherently multidimensional due to the intended content or construct structure of the tests (Ackerman, Gierl, & Walker, 2003). Tests consisting of different content domains often measure a primary ability and additional abilities; thus, each item measures the primary ability and one additional secondary ability. Content categories can be considered as the source of secondary abilities. That is, while the primary ability is the estimated overall score, subscores for content categories are considered secondary abilities (DeMars, 2005). Subscores estimated from secondary abilities have been of substantial importance recently (DeMars, 2005; Reckase & Xu, 2015; Sinharay, Haberman, & Wainer, 2011; Wedman & Lyren, 2015). It is because of the potential diagnostic value of the subscores in future remedial work in which students have a chance to know their weaknesses and strengths in different content domains that the test measures (Haberman & Sinharay, 2010). Haberman (2008) and Sinharay (2010) focused on the added value of subscores over the total score by using Classical Test Theory methods. Brennan (2012) suggested the utility index similar to Haberman’s method. Besides, the subscore augmentation method developed by Wainer, Sheehan, and Wang (2000) is used to examine whether getting information from other portions of the test (augmented subscore) estimates the subscore more accurately.

* Res Assist., Gazi University, Gazi Faculty of Education, Ankara-Turkey, erdemiraysenur@gmail.com, ORCID ID: 0000-0001-9656-0878
** Prof. PhD., Gazi University, Gazi Faculty of Education, Ankara-Turkey, hakanatar@gazi.edu.tr, ORCID ID: 0000-0001-5372-1926

To cite this article:
Erdemir, A.. & Atar, H. Y. (2020). Simultaneous estimation of overall score and subscores using MIRT, HO-IRT, and Bi-factor Model on TIMSS data. *Journal of Measurement and Evaluation in Education and Psychology, 11*(1), 61-75. doi: 10.21031/epod.645478

Received: 11.11.2019
Accepted: 25.02.2020

The psychometric quality of subscores is also of importance when they are utilized by policymakers, test takers, and educators for the purpose of diagnosis and admission (Haberman, 2008; Monaghan, 2006). According to the Standard 1.14 of the Standards of Educational and Psychological Testing (2014, p.27), “When a test provides more than one score, the distinctiveness and reliability of the separate scores should be demonstrated.” Over the years, researchers have examined the methods arguing the psychometric quality of subscores (de la Torre & Patz, 2005; DeMars, 2005; Fan, 2016; Haberman, 2008; Haberman & Sinharay, 2010; Longabach, 2015; Md Desa, 2012; Shin, 2007; Sinharay, 2010; Stone, Ye, Zhu & Lane, 2010; Wang, Chen, & Cheng, 2004; Yao, 2014; Yao & Boughton, 2007).

In multidimensional tests, when the overall score is reported, it shows the test-takers' achievement levels concerning the overall construct of the test subject. Subscores, on the other hand, give additional information about the strengths and weaknesses of test-takers in the domain abilities while the overall score presents a general profile of the test-takers. For example, the TOEFL test, which is the English-language test, has four content domains (reading, listening, speaking, and writing). For this test, test-takers receive four subscores related to each skill and a total score as a representative of general English-language ability. Since many tests have a multidimensional structure, the interest in estimating and reporting overall scores and subscores simultaneously has increased (Liu & Liu, 2017). Simultaneous estimation of those scores provides test takers and educators with more detailed information about the primary and secondary ability levels of students (Yao, 2010). More clearly, as opposed to the separate estimation of the primary and secondary abilities, simultaneous estimation means one can have the information on those abilities with one single analysis.

There are studies discussing the methods estimating the overall score and subscores simultaneously (de la Torre & Song, 2009; de la Torre & Song, 2010; Liu, Li, & Liu, 2018; Soysal & Kelecioğlu, 2018; Yao, 2010). In all these studies, it is emphasized that the reliability of scores is very important when the overall scores and subscores need to be reported. Yao (2010) states that the simple averaging method is the most commonly used method to obtain the overall score by averaging the domain scores. She also indicates that simply averaging the domain scores ignores (a) different maximum raw score points of different domains, (b) correlation between the domain abilities, and (c) the possibility of having a different relationship between overall scores and domain scores at different score points. In order to overcome these problems, Yao (2010) proposed using the Multidimensional Item Response Theory (MIRT) maximum information method for the overall score instead of the simple averaging method. The proposed method does not assume any linear relationship between the overall score and domain scores. In the study, subscores were estimated by using MIRT, and the overall scores were estimated by using the MIRT maximum information method. Estimated overall and subscores were compared to those obtained from the Higher-Order Item Response Theory (HO-IRT), Bi-factor, and unidimensional IRT methods. It is found that the MIRT method provides reliable subscores similar to the HO-IRT method and also reliable overall score. The MIRT maximum information method produced overall scores with the smallest standard error of measurement (Yao, 2010).

de la Torre and Song (2009) also proposed using Higher-order Item Response Theory approach for simultaneous estimation of overall and domain abilities. The HO-IRT method assumes a linear relationship between the overall score and the domain score, unlike the MIRT method. In the study, the HO-IRT method was compared with the unidimensional IRT (UIRT) in which the overall ability is estimated using all items ignoring the multidimensional structure of the data, and the domain abilities are estimated using corresponding subsets of items, separately. The findings of the study show that the overall and domain abilities can be estimated more efficiently by using the HO-IRT method. Additionally, in the HO-IRT framework, it is possible to obtain efficient overall and domain ability estimates with small sample sizes and small number of items (de la Torre & Song, 2010).

To estimate the overall score and domain scores based on the bi-factor model, Liu et al. (2018) introduced six methods in the framework of the bi-factor model and compared them with the MIRT method. The weights of the general and domain factors were calculated in different ways in those six bi-factor methods. It is found that the most accurate and reliable overall and domain scores in most conditions were obtained using Bi-factor-M4 and Bi-factor-M6 methods, weights of which were computed using discrimination parameters for a specific domain. In the bi-factor methods, the domain-specific factors are orthogonal to the general factor and each other, unlike the MIRT and HO-IRT methods.

Related research regarding simultaneous estimation of the overall and subscores seems to be few in number (de la Torre & Song, 2010; Liu et al., 2018; Soysal & Kelecioğlu, 2018; Yao, 2010). The present study aims to contribute to the related research. The purpose of the study is to investigate by using which method simultaneous estimation of the overall score and subscores yields more accurate and reliable ability estimates. For this purpose, MIRT, HO-IRT, and bi-factor general model, the most suggested methods in literature, were used in the study. This study also differs from earlier research in that it runs the analysis on mixed-format data, including both dichotomously and polytomously scored items, whereas all other studies used data consisting only dichotomously or polytomously scored items. At this point, using mixed-format data is thought to be important since tests containing a mixture of multiple-choice and constructed-response items are used in many testing situations (Lane, 2005; Yao & Schwarz, 2006).

Ability Estimation with Multiple Dimensions

Multidimensional Item Response Theory

Multidimensional Item Response Theory is a method that provides "a reasonably accurate representation of the relationship between persons' locations in a multidimensional space and the probabilities of their responses to a test item" (Reckase, 2009, p. 53) with a particular mathematical expression. An essential distinction between MIRT models related to the structure of the data is whether the probability of responses to any test item is influenced by one latent dimension or not. If this is the case, the structure of the data is defined as between-item dimensionality (simple-structure). If responses to one item are affected by more than one ability, then, it is denoted as within-item dimensionality (complex structure; Adams, Wilson, & Wang, 1997). In this study, the data were assumed to follow a simple structure because each item was modeled as depending on one specific ability dimension.

Additionally, there are several models within MIRT varying basically in terms of the number of possible score points for the items: MIRT models for dichotomously scored items and MIRT models for polytomously scored items. All of the MIRT models can be considered as generalizations of unidimensional IRT models (Reckase, 1997). However, many tests contain both dichotomously and polytomously scored items on the same test form, which creates a need to use different item response models together (Yao & Schwarz, 2006). TIMSS mathematics achievement test also contains mixed item types. Therefore, in the present study, the TIMSS data were examined using the multidimensional three-parameter logistic (M-3PL) model for dichotomously scored items and the multidimensional two-parameter partial credit model (M-2PPC) applied to polytomously scored items as suggested in the study of Yao & Schwarz (2006). For a dichotomous item j, the probability of a correct response to item j for an examinee with ability $\vec{\boldsymbol{\theta}}_{\mathbf{i}}$ = (θ_{i1}, θ_{i2}, ..., $\theta_{i\mathrm{D}}$) for the M-3PL model (Reckase, 1997) is

$$P_{ij1} = P(x_{ij} = 1 \mid \vec{\boldsymbol{\theta}}_{\mathbf{i}}, \vec{\boldsymbol{\beta}}_{\mathbf{j}}) = \beta_{3j} + \frac{1-\beta_{3j}}{1+e^{(-\vec{\boldsymbol{\beta}}_{\mathbf{2j}} \odot \vec{\boldsymbol{\theta}}_{\mathbf{i}}^{\mathbf{T}} + \beta_{1j})}}, \tag{1}$$

where

x_{ij} = the response of examinee i to item j

$\vec{\boldsymbol{\beta}}_{\mathbf{j}}$ = the parameters for the j^{th} item ($\vec{\beta}_{2j}$, β_{1j}, β_{3j})

$\vec{\boldsymbol{\beta}}_{\mathbf{2j}}$ = a vector of dimension D of item discrimination parameters (β_{2j1}, …, β_{2jD})

β_{1j} = the scale difficulty parameter

β_{3j} = the scale guessing parameter

$\vec{\boldsymbol{\beta}}_{\mathbf{2j}} \odot \vec{\theta}_i^T$ = a dot product of two vectors.

For a polytomous item *j*, the probability of a response *k*−1 to item *j* for an examinee with ability $\vec{\boldsymbol{\theta}}_i$ for the M-2PPC model (Yao & Schwarz, 2006) is

$$P_{ijk} = P(x_{ij} = k - 1 \mid \vec{\boldsymbol{\theta}}_{\mathbf{i}}, \vec{\boldsymbol{\beta}}_{\mathbf{j}}) = \frac{e^{(k-1)\vec{\boldsymbol{\beta}}_{\mathbf{2j}}\odot\vec{\boldsymbol{\theta}}\mathbf{i} - \sum_{t=1}^{k}\beta_{\delta_t j}}}{\sum_{m=1}^{K_j} e^{\left((m-1)\vec{\boldsymbol{\beta}}_{\mathbf{2j}}\odot\vec{\boldsymbol{\theta}}_{\mathbf{i}}^{\mathbf{T}} - \sum_{t=1}^{m}\beta_{\delta_t j}\right)}}, \tag{2}$$

where
x_{ij} = the response of examinee *i* to item *j* (0, …, $K_j - 1$)
$\vec{\boldsymbol{\beta}}_{\mathbf{j}}$ = the parameters for the *j*th item ($\vec{\boldsymbol{\beta}}_{\mathbf{2j}}$, $\beta_{\delta_2 j}$, …, $\beta_{\delta_{K_j} j}$)
$\vec{\boldsymbol{\beta}}_{\mathbf{2j}}$ = a vector of dimension D of item discrimination parameters (β_{2j1}, …, β_{2jD})
$\beta_{\delta_k j}$ = the threshold parameters for *k* = 1, 2, …, K_j; $\beta_{1j} = 0$ and K_j = the number of response categories for the *j*th item.

Higher-Order Item Response Theory

de la Torre and Song (2009) proposed a higher-order multidimensional IRT approach in which overall and domain abilities can be specified simultaneously. In this model, the first order describes domain-specific abilities, while the second-order can be viewed as the overall ability. It is considered that each domain is unidimensional; the second-order ability contains all the domain abilities, so the overall ability is also viewed as unidimensional. de la Torre and Hong (2010) stated that a test is deemed multi-unidimensional in the HO-IRT framework.

The HO-IRT method uses a hierarchical Bayesian framework (de la Torre et al., 2011), and the domain abilities are considered as linear functions of the overall ability, expressed as

$$\theta_i^{(d)} = \lambda^{(d)}\theta_i + \varepsilon_{id}, \tag{3}$$

where
θ_i = the overall ability,
$\theta_i^{(d)}$ = the domain-specific abilities, *d* = 1, 2, …, D,
$\lambda^{(d)}$ = the latent coefficient in regressing the ability *d* on the overall ability,
ε_{id} = the error term following a normal distribution with a mean of zero and variance of $1 - \lambda^{(d)2}$, and $|\lambda^{(d)}| \leq 1$.

The latent regression coefficient, $\lambda^{(d)}$, also means the correlation between the overall and domain abilities. Mathematically, $\lambda^{(d)}$ can have negative values, but it is generally expected to be positive since domain abilities are typically related to the overall ability.

Focusing on estimating abilities of test-takers (Equation 3), the model parameters that need to be estimated are the overall ability, domain abilities, and the latent regression parameters $\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(D)}$. With a hierarchical Bayesian framework, the model formulation is expressed as follows (de la Torre & Song, 2009):

$$\theta_i \sim N(0,1), \tag{4}$$

$$\lambda^{(d)} \sim U(-1.0, 1.0), \tag{5}$$

and

$$\theta_i^{(d)} \mid \theta_i, \lambda^{(d)} \sim N\big(\lambda^{(d)}\theta_i, 1 - \lambda^{(d)2}\big). \tag{6}$$

The model parameters are estimated by using MCMC sampling procedure. First, the overall ability θ_i is sampled from a normal distribution (Equation 4), and the regression coefficient is sampled from a uniform distribution (Equation 5). Then, based on the estimated overall ability and the regression coefficients, the MCMC procedure samples the domain abilities with the sixth equation (de la Torre & Hong, 2010; de la Torre & Song, 2009).

Bi-factor General Model

The bi-factor model (Gibbons & Hedeker, 1992) defines a general factor on which all the items load and domain-specific factors on which the items related to that dimension load. The domain-specific factors are orthogonal to the general factor. The method provides estimates of the overall ability and domain abilities at the same time. It is considered that the domain factors are nuisance traits within the Bi-factor framework, which yields a more meaningful overall ability (DeMars, 2013; Yao, 2010).

Cai, Yang, and Hansen (2011) demonstrated the factor pattern of the standard item bi-factor measurement structure as

$$\begin{pmatrix} a_{10} & a_{11} & 0 \\ a_{20} & a_{21} & 0 \\ a_{30} & a_{31} & 0 \\ a_{40} & 0 & a_{42} \\ a_{50} & 0 & a_{52} \\ a_{60} & 0 & a_{62} \end{pmatrix}.$$

As seen in the pattern, there are six items, one general and two domain-specific factors. The *a*s are the indicators of item discrimination parameters, which are similar to the factor loadings. The first factor is the general factor, and the last two columns refer to the domain factors (Cai et al., 2011).

As defined in Liu et al.'s (2018) study, in the vector of item discrimination parameters, only the one for the general factor (β_{aj}) and one discrimination parameter of *s*th subscale (β_{sj}) have values other than zero. The ability vector of each examinee includes one overall ability for the general factor (θ_{ia}) and domain-specific abilities for S specific factors ($\theta_{i1}, \dots, \theta_{is}, \dots, \theta_{iS}$).

Based on the Bi-factor model, estimation of the overall score and domain scores can be expressed as follows:

$$\theta_{overall} = w_{1a}\theta_{ia} + \sum_{s=1}^{S} w_{1s}\theta_{is} \tag{7}$$

and

$$\theta_{domain\_s} = w_{2a}\theta_{ia} + \ w_{2s}\theta_{is}, \tag{8}$$

where

w_{1a} = weight of the general factor for the overall score

w_{1s} = weight of the domain factors for the overall score

w_{2a} = weight of the general factor for the domain scores

w_{2s} = weight of the domain factors for the domain scores.

Thus, the overall score (Equation 7)) is a weighted composite of the general factor (θ_{ia}) and all domain factors (($\theta_{i1}, \dots, \theta_{is}, \dots, \theta_{iS}$), while the domain score (Equation 8) for the *s*th factor is a weighted composite of the general factor (θ_{ia}) and the relevant domain-specific factor (θ_{is}). In the current study, the Bi-factor general model was employed by using 1 and 0 as the weights, as in the study of Yao (2010): $w_{1a} = 1, w_{1s} = 0$ and $w_{2a} = 0, w_{2s} = 1$. In this method, the general factor represents the overall score, while the domain-specific factors represent subscores.

METHOD

Data Description

Eighth graders' responses to the mathematics test in Trends in International Mathematics and Science Study (TIMSS) 2015 were used in this study. Each country's data from the 1st booklet of mathematics achievement test were merged into a whole data set. The reason behind choosing 1st booklet is that it is the booklet that has the largest number of polytomously-scores items (four items). For handling missing data, the listwise deletion method was utilized because the researchers aimed to analyze the data consisting of the subjects who answered all of the items The final version of the data consists of 5732 students from all the countries who were administered the 1st assessment booklet in TIMSS 2015. Table 1 shows the distribution of scoring types and contents for the chosen test form for the current study.

Table 1. Scoring Types and Content Distribution for The Data

Content domain	Scoring types	Number of items
Number	Dichotomously-scored	11
	Polytomously-scored	3
Algebra	Dichotomously-scored	9
Geometry	Dichotomously-scored	5
	Polytomously-scored	1
Data and Chance	Dichotomously-scored	6

As shown in Table 1, the test has four content domains, which are number (14 items), algebra (9 items), geometry (6 items), and data and change (6 items). The total number of items is 35, four of which are polytomously scored (0-1-2), and the rest of the items are dichotomously scored (0-1).

Data Analysis

Dimensionality analysis

In order to improve interpretations and uses of scores, the dimensional structure of the data is essential to get evidence of validity (Reckase & Xu, 2015). Dimensionality shows the relationship between a test and response patterns, which gives clues about the latent structure measured by the test. Wainer and Thissen (1996) mention the fixed and random forms of dimensionality. While random dimensionality is a concept explaining the possibility of encountering some "unexpected" dimensions, fixed dimensionality is a somewhat "expected" situation. In particular, it is usual to see multidimensionality in scores when the test has multiple content domains. It can be assumed that the data have a multidimensional structure when the test has content domains. Under this circumstance, it is said that it might be more reasonable and effective to use confirmatory dimensionality assessment (Zhang, 2016). Therefore, confirmatory methods were used to assess the dimensionality structure of the data in this study. Confirmatory Factor Analysis (CFA) and content-based confirmatory mode of Poly-DETECT (Zhang & Stout, 1999a, 1999b; Zhang, 2007) were the methods utilized as dimensionality analysis in the current study.

The poly-DETECT analysis was done through the *sirt* package (Robitzsch, 2018). The result of the analysis gives the indices DETECT, ASSI and RATIO. The information about the evaluation of these indices is presented in Table 2 (Jang & Roussos, 2007; Zhang, 2007):

Table 2. Dimensionality Indices of the Poly-DETECT Analysis and Their Evaluation

Index	Critical Values	Explanation
DETECT	DETECT > 1.00	Strong multidimensionality
	.40 < DETECT < 1.00	Moderate multidimensionality
	.20 < DETECT < .40	Weak multidimensionality
	DETECT < .20	Essential unidimensionality
ASSI	ASSI=1	Maximum value under simple structure
	ASSI > .25	Essential deviation from unidimensionality
	ASSI < .25	Essential unidimensionality
RATIO	RATIO=1	Maximum value under simple structure
	RATIO > .36	Essential deviation from unidimensionality
	RATIO < .36	Essential unidimensionality

The DETECT index shows the amount of multidimensionality on a test. The DETECT value of 1 or more indicates strong multidimensionality; values of 0.4 to 1 indicate moderate to large multidimensionality; values below 0.4 indicate moderate to weak multidimensionality, and values below 0.2 indicate unidimensionality. For ASSI and RATIO indices, the critical values are 0.25 and 0.36, respectively. ASSI and RATIO values smaller than those critical values indicate that the data is essentially unidimensional. On the other hand, the data that has the ASSI and RATIO values higher than the critical values are considered to be multidimensional.

MPlus software program was used to conduct the Confirmatory Factor Analysis. Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), and RMSEA (Root Mean Square Error of Approximation) are the fit indices used to test model fit. It is reported that the model fits quite well with the data when CFI and TLI have values more than 0.95, and RMSEA has a value lower than 0.05 (Hu & Bentler, 1999; Tabachnick & Fidell, 2013, p.720-723).

Estimating overall score and subscores

Three estimation methods (MIRT, HO-IRT, and Bi-factor) were used to obtain the overall score (mathematics achievement) and subscores (number, algebra, geometry, and data and chance) for 5732 test takers who were administered the first booklet of TIMSS 2015. Ability parameters for the methods were estimated using the BMIRT software (Yao, 2003; Yao, 2013; Yao, Lewis, & Zhang, 2008). In the present study, the data were analyzed using the M-3PL model for dichotomously-scored items, and the M-2PPC applied to polytomously-scored items for all of the estimation methods. The following are brief explanations of the estimation methods and what they estimate in the context of the current data:

- MIRT: the simple structure MIRT analysis was used to estimate abilities based on four content domains. It gives four thetas (θ), each of which represents single subscore. The overall score was obtained by domain scores using maximum information method as in Yao (2010).
- HO-IRT: It is assumed that there is a linear relationship between the overall score and subscores, so the parameters for the overall ability and domain abilities were estimated simultaneously.
- Bi-factor: The Bi-factor general model estimated five abilities. The first one was the general dimension, and the other four abilities were content-specific dimensions, respectively. In the bi-factor model, content-specific dimensions are orthogonal to each other and the general dimension, and there is no correlation between dimensions.

The default priors of BMIRT software were used for the analyses in this study. The mean and variance of the ability prior distribution were 0.0 and 1.0, respectively. The priors were taken to be lognormal for the discrimination parameters with a mean of 1.5 and variance of 1.5. For the difficulty or threshold parameters, a standard normal distribution with a mean of 0.0 and variance of 1.5 was used. Guessing parameter c had prior beta (α, β) distribution, in which α = 100 and β =400.

Evaluation criteria

The conditional standard error of measurement (cSEM) was used to evaluate the accuracy of overall scores and subscores. The BMIRT program calculated the cSEM values for each student's ability parameters under studied methods estimating the overall and domain scores simultaneously. Then, the analysis of variance (ANOVA) on repeated-measures data for the cSEM was conducted to examine whether there is a significant difference among the mean errors calculated by estimation methods.

The other criterion for the evaluation of methods is reliability. A method proposed by de la Torre & Patz (2005) called Bayesian marginal ability or empirical reliability (Brown & Croudace, 2015) was applied for this study. The reliability of test *d* can be obtained from

$$\rho_d = \frac{var\ (\hat{\theta}_d)}{var\ (\hat{\theta}_d) + \overline{Pvar\ (\hat{\theta}_d)}}. \tag{9}$$

The observed (Equation 10) and marginal posterior (Equation 11) variance of the overall or domain ability estimates are computed from the estimated ability scores $\hat{\theta}$ and their standard errors (SE) in a sample of N test takers:

$$var\ (\hat{\theta}_d) = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{\theta}_i - \bar{\hat{\theta}}\right)^2 \tag{10}$$

$$\overline{Pvar\ (\hat{\theta}_d)} = \frac{1}{N} \sum_{i=1}^{N} SE^2\ (\hat{\theta}_i)\ . \tag{11}$$

For this study, reliability measures for one overall score and four subscores were obtained from the equations above for each studied methods. Higher marginal reliability indicates higher reliability of scores from the methods tested (Md Desa, 2012).

RESULTS

Dimensionality Analysis

Poly-DETECT (confirmatory mode) and Confirmatory Factor Analysis were conducted in order to examine the multidimensionality due to the content domains for mixed-format TIMSS data used in this study. Table 3 shows the results of the content-based Poly-DETECT analysis.

Table 3. The Results of the Poly-DETECT Analysis

Index	Value	Corresponding Classification	
DETECT	0.406	Moderate multidimensionality	.40 < DETECT < 1.00
ASSI	0.459	Essential deviation from unidimensionality	ASSI > .25 RATIO > .36
RATIO	0.522		

As seen in Table 3, the results yielded an essential deviation from unidimensionality in which ASSI = .459 and RATIO = 0.522. DETECT index, which is .406, means moderate multidimensionality. The values of indices obtained from the Poly-DETECT analysis provide evidence of multidimensionality for the current data.

A four-factor model was tested through CFA. The content domains with related items were taken as factors, and the model fit was evaluated. Fit indices for the data and the associated criteria are presented in Table 4.

Table 4. CFA Model Fit Indices and Associated Criteria

Index	Value	Good Fit
TLI	0.974	TLI ≥ 0.95
CFI	0.975	CFI ≥ 0.95
RMSEA	0.037	RMSEA ≤ 0.05

CFI and TLI indicated that the model fits the data well (≥ 0.95). Likewise, the RMSEA value (≤ 0.05) showed a good fit (Table 4). According to the results of CFA, the four-factor model had a good fit with the present data, which supported content-based multidimensionality. After providing evidence of the content-based multidimensionality of the data, the overall and domain abilities were obtained with the aforementioned methods.

Precision of Estimates

The selected three methods (MIRT, HO-IRT, and Bi-factor) for the current study were used through running the BMIRT program to estimate the overall and subscores simultaneously. BMIRT also provided standard errors for the estimated scores. The means for standard errors for the overall and domain ability estimates under each estimation method are summarized in Table 5.

Table 5. The Means and Standard Deviations for the Standard Errors for the Overall and Domain Abilities

	Domain				
Method	Number (14 items)	Algebra (9 items)	Geometry (6 items)	Data and Chance (6items)	Overall
MIRT	0.376 (0.125)	0.511 (0.130)	0.545 (0.142)	0.586 (0.149)	0.295 (0.124)
HO-IRT	0.332 (0.103)	0.410 (0.120)	0.422 (0.133)	0.443 (0.140)	0.474 (0.050)
Bi-factor	0.670 (0.164)	0.820 (0.163)	0.849 (0.168)	0.898 (0.178)	0.322 (0.135)

Table 5 shows the means and standard deviations for the standard errors for each ability. Generally, MIRT and HO-IRT yielded similar results, but the HO-IRT estimation method performed slightly better than MIRT for domain abilities. The Bi-factor model gave the worst standard errors for the domain abilities among all the methods and similar to the MIRT for the overall ability. The repeated-measures ANOVA results whether the difference between standard errors are statistically significant are presented in Table 6.

Table 6. The Repeated-measures ANOVA results for the Standard Errors

Ability	Source	Sum of Squares	*df*	Mean Square	*F*	Partial η^2	Pairwise comparison
Number	Methods	386.536	1.726	223.918	15465.323*	.730	All pairwise
	Error	143.239	9893.087	.014			HOIRT<MIRT<BF
Algebra	Methods	521.582	1.885	276.701	15288.071*	.727	All pairwise
	Error	195.524	10802.949	.018			HOIRT<MIRT<BF
Geometry	Methods	552.440	1.909	289.387	14196.309*	.712	All pairwise
	Error	223.018	10940.494	.020			HOIRT<MIRT<BF
Data and chance	Methods	621.124	1.925	322.731	13418.317*	.701	All pairwise
	Error	265.284	11029.804	.024			HOIRT<MIRT<BF
Overall	Methods	105.937	1.692	62.613	8162.767*	.588	All pairwise
	Error	74.377	9696.490	.008			MIRT<BF<HOIRT

**p* < .001

The repeated-measures ANOVA with a Greenhouse-Geisser correction determined that mean standard errors differed statistically significantly when the estimation method was changed for the domain ability estimates ($F_{(1.726,\ 9893.087)\ \text{number}}$ = 15465.323, *p* < .05, *partial* η^2 = .73; $F_{(1.885,\ 10802.949)\ \text{algebra}}$ = 15288.071, *p* < .05, *partial* η^2 = .727; $F_{(1.909,\ 10940.494)\ \text{geometry}}$ = 14196.309, *p* < .05, *partial* η^2 = .712; $F_{(1.925,\ 11029.804)\ \text{data and chance}}$ = 13418.317, *p* < .05, *partial* η^2 = .701). Post hoc tests using the Bonferroni correction revealed that all pairwise comparisons were statistically significantly different from each other. According to the results in Table 4, the HO-IRT method had the lowest standard errors for all domain abilities, and MIRT had the second-lowest standard errors. Domain abilities from the Bi-factor model were not as accurate as the other two methods.

Therefore, it can be concluded that HO-IRT elicited a statistically significant reduction in standard errors of domain ability estimates. Likewise, the overall ability results showed that the standard errors were significantly affected by the type of estimation method ($F_{(1.692,\ 9696.490)\ \mathrm{overall}}$ = 8162.767, *p* < .05, *partial* η^2 = .588). Post hoc tests using the Bonferroni correction revealed that all pairwise comparisons were significantly different from each other. The HO-IRT had the highest mean for standard errors. The MIRT and Bi-factor model had low and similar standard errors for the overall ability. In general, the three estimation methods were significantly different for all the abilities, including the overall and domain abilities.

Reliability of Scores

The overall and four domain ability estimates from the studied methods were compared in terms of marginal reliability. Estimated reliability coefficients are presented in Table 7.

Table 7. Marginal Reliability Coefficients

	Domain				
Method	Number (14 items)	Algebra (9 items)	Geometry (6 items)	Data and Chance (6items)	Overall
MIRT	0.847	0.722	0.682	0.635	0.816
HO-IRT	0.894	0.838	0.824	0.809	0.815
Bi-factor	0.539	0.301	0.253	0.161	0.876

Table 7 presents the Bayesian marginal reliability of the overall score and subscores based on four content domains. In general, MIRT and HO-IRT had substantially higher reliability across all content domains compared to the reliability of the Bi-factor model. The reliability of the Bi-factor model was extremely low for the domain scores, especially for geometry (i.e., 0.253) and data and chance (i.e. 0.161). In addition, the reliability of domains varied slightly between domains for MIRT and HO-IRT. The reliability coefficient of HO-IRT subscores was for number, 0.894; for algebra, 0.838; for geometry, 0.824, and for data and chance, .809. It can be concluded that HO-IRT was the most reliable method of estimating subscores, followed by MIRT, for all content domains for the data used in the current study. Furthermore, the reliabilities of all methods decreased as the number of items in the domains decreased. The reliability of the overall score was for MIRT, 0.816; for HO-IRT, 0.815, and for Bi-factor, 0.876. Unlike the subscores, the Bi-factor model was the most reliable method for the overall score estimation. The other two methods (MIRT and HO-IRT) also estimated the overall score with high reliability.

DISCUSSION and CONCLUSION

When the overall and domain abilities are reported to the test takers and used by the authorities, it is important to obtain accurate and reliable estimates of the overall score and subscores. The overall scores are useful in reporting the test-takers' general achievement and taking important decisions such as rank-ordering the test takers. On the other hand, the subscores provide test takers, teachers, or policymakers with more diagnostic information such as strengths and weaknesses in each domain. The simultaneous estimation of those scores can be another solution to both of the needs.

This study examined three methods of estimating the overall score and subscores simultaneously in the same model, including MIRT, HO-IRT, and Bi-factor, and compared the reliability and precision of these methods across the overall and domain ability estimates. For this purpose, the real data of mixed item types from TIMSS 2015 were used. The results of Poly-DETECT and CFA provided evidence for the content-based multidimensional structure of the data.

The study showed that the MIRT and HO-IRT methods performed similarly in terms of precision and reliability for subscore estimates. However, HO-IRT had slightly lower standard errors and higher reliability than MIRT. Likewise, de la Torre and Song (2009) stated that domain ability estimates can be more efficient by using the HO-IRT model. In addition, Yao (2010) found that MIRT and HO-IRT were quite similar in terms of estimating subscores. The precise ability estimation and reliable scores by using HO-IRT also supported the use of subscores for reporting for the current data. The Bi-factor general model had the highest standard errors and lowest reliability estimates for the domain scores. Liu et al. (2018) also did not recommend the Bi-factor, the original factor method, for reporting scores. They proposed six other methods of reporting overall and subscores as weighted composite scores of the overall and domain-specific factors in a bi-factor model.

For the overall ability estimation, the MIRT maximum information method and Bi-factor model outperformed the HO-IRT method with regard to standard errors. The MIRT maximum information method had the smallest standard error of measurement for the overall score estimates, as in the study of Yao (2010). While all three methods performed similarly and relatively good in terms of the overall score reliability, the reliability of Bi-factor model was a bit higher than the other two methods.

The analyses of the current study suggested that overall, HO-IRT seems the best solution for the simultaneous estimation of the overall and subscores for the data from TIMSS 2015. Soysal and Kelecioğlu (2018) also recommended the use of HO-IRT in estimation of overall and subscores in their study.

In the present study, only real data were used to examine the relative performance of the three methods, since the true model for the data was not known. Therefore, it is quite possible to get different results for other samples. It is suggested that future research can be done by using other real data. It is also advisable that when the simultaneous estimation of the overall and domain abilities must be done in testing practices, the relative performance of the estimation methods should be checked before reporting the scores to test takers.

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37-51. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Adams, R., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Brennan, R. L. (2012). *Utility indexes for decisions about subscores* (CASMA Research Report No. 33). Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessment. http://www.education.uiowa.edu/docs/default-source/casma---research/33utility-revised.pdf?sfvrsn=2

Brown, A. & Croudace, T. (2015). Scoring and estimating score precision using multidimensional IRT. In Reise, S. P. & Revicki, D. A. (Eds.). *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment (a volume in the Multivariate Applications Series).* New York: Routledge/Taylor & Francis Group.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods*, *16*(3), 221–248. http://doi.org/10.1037/a0023350

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have : A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295–311. https://doi.org/10.3102/10769986030003295

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT model approach. *Applied Psychological Measurement*, *34*, 267-285. https://doi.org/10.1177/0146621608329501

de la Torre, J., & Song, H. (2009). Simultaneous estimation of overall and domain abilities: A higher-order ırt model approach. *Applied Psychological Measurement*, *33*(8), 620–639. http://doi.org/10.1177/0146621608326423 de la Torre, J., Song, H., & Hong, Y. (2011). A comparison of four methods of IRT subscoring. *Applied Psychological Measurement*, *35*(4), 296–316. http://doi.org/10.1177/0146621610378653

DeMars, C.E. (2005, August). *Scoring subscales using multidimensional item response theory models.* Poster presented at the annual meeting of the American Psychology Association. https://eric.ed.gov/?id=ED496242

DeMars, C. E. (2013). A tutorial on interpreting bifactor model scores. *International Journal of Testing, 13*(4), 354-378. https://doi.org/10.1080/15305058.2013.799067

Fan, F. (2016). *Subscore Reliability and Classification Consistency : A Comparison of Five Methods* (Doctoral dissertation, University of Massachusetts Amherst). https://scholarworks.umass.edu/dissertations_2/857/

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57*(3), 423–436. http://doi.org/10.1007/BF02295430

Haberman, S. J. (2008). When can subscale scores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75,* 209– 227. https://doi.org/10.1007/s11336-010-9158-4

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structural analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1- 55. https://doi.org/10.1080/10705519909540118

Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement, 44*(1), 1-21. https://doi.org/10.1111/j.1745-3984.2007.00024.x

Lane, S. (2005, April). *Status and future directions for performance assessments in education.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Liu, Y., Li, Z., & Liu, H. (2018). Reporting Valid and Reliable Overall Scores and Domain Scores Using Bi-Factor Model. *Applied Psychological Measurement, 43*(7), 562–576. https://doi.org/10.1177/0146621618813093

Liu, Y., & Liu, H. (2017). Reporting overall scores and domain scores of bi-factor models. *Acta Psychologica Sinica*, *49*(9), 1234. http://doi.org/10.3724/SP.J.1041.2017.01234

Longabach, T. (2015). *A comparison of subscore reporting methods for a state assessment of English language proficiency* (Doctoral dissertation, University of Kansas). https://kuscholarworks.ku.edu/handle/1808/19517

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

Monaghan, W. (2006). *The facts about subscale scores* (ETS R&D Connections No. 4). Princeton, NJ: Educational Testing Service. https://www.ets.org/Media/Research/pdf/RD_Connections4.pdf

Reckase, M. D. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement, 9*(4), 401-412. https://doi.org/10.1177/014662168500900409

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden &R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271–286). NY: Springer.

Reckase, M. D. (2009). *Multidimensional item response theory (statistics for social and behavioral sciences).* New York: Springer.

Reckase, M. D., & Xu, J.-R. (2015). The Evidence for a Subscore Structure in a Test of English Language Competency for English Language Learners. *Educational and Psychological Measurement*, *75*(5), 805–825. https://doi.org/10.1177/0013164414554416

Robitzsch, A. (2019). Supplementary Item Response Theory Models Version. R-project, Package 'sirt' manual. https://cran.r-project.org/web/packages/sirt/sirt.pdf

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2), 150–174. https://doi.org/10.1111/j.1745-3984.2010.00106.x

Sinharay, S., Haberman, S. J., & Wainer, H. (2011). Do Adjusted Subscores Lack Validity? Don't Blame the Messenger. *Educational and Psychological Measurement*, *71*(5), 789–797. https://doi.org/10.1177/0013164410391782

Soysal, S., & Kelecioğlu, H. (2018). Toplam Test ve Alt Test Puanlarının Kestiriminin Hiyerarşik Madde Tepki Kuramı Modelleri ile Karşılaştırılması. *Journal of Measurement and Evaluation in Education and Psychology , 9*(2), 178-201. https://doi.org/10.21031/epod.404089

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education*, *23*(1), 63–86. https://doi.org/10.1080/08957340903423651

Tabachnick B. G. & Fidel, L. S. (2013). *Using multivariate statistics (4th ed.).* MA: Allyn & Bacon, Inc.

Wainer, H., Sheehan, K. M., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement*, *37*(2), 113–140. https://doi.org/10.1111/j.1745-3984.2000.tb01079.x

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practices*, *15*(1), 22–29. https://doi.org/10.1111/j.1745-3992.1996.tb00803.x

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), 116–136. http://doi.org/10.1037/1082-989X.9.1.116

Wedman, J., & Lyren, P.- E. (2015). Methods for examining the psychometric quality of subscores: A review and application. *Practical Assessment, Research and Evaluation*, *20*(1). https://scholarworks.umass.edu/pare/vol20/iss1/21/

Yao, L. (2003). BMIRT: Bayesian multivariate item response theory [computer software]. Monterey, CA: DefenseManpower Data Center. Downloaded from https://www.bmirt.com/6271.html

Yao, L. (2010). Reporting valid and reliable overall scores and domain scores. *Journal of Educational Measurement*, *47*(3), 339–360. http://doi.org/10.1111/j.1745-3984.2010.00117

Yao, L. (2013). The BMIRT toolkit. Monterey. https://www.bmirt.com/8220.html

Yao, L. (2014). Multidimensional item response theory for score reporting. In Y. Cheng, & H.- H. Chang (Eds.) *Advances in modern international testing: Transition from summative to formative assessment.* Charlotte, NC: Information Age.

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. http://doi.org/10.1177/0146621606291559

Yao, L., Lewis, D., & Zhang, L. (2008, April). *An introduction to the application of BMIRT: Bayesian multivariate item response theory software.* Training session presented at the meeting of the National Council on Measurement in Education, New York, NY.

Yao, L., & Schwarz R. (2006). Amultidimensional partial credit model with associated item and test statistics: An application to mixed format tests. *Applied Psychological Measurement. 30*(6), 469—492. https://doi.org/10.1177/0146621605284537

Zhang, J. (2007). Conditional covariance theory and DETECT for polytomous items. *Psychometrika, 72*(1), 69-91. https://doi.org/10.1007/s11336-004-1257-7

Zhang, J., & Stout, W. (1999a). Conditional covariance structure of generalized compensatory multidimensional items. *Psychometrika, 64*(2), 129-152. https://doi.org/10.1007/BF02294532

Zhang, J., & Stout, W. (1999b). The theoretical DETECT index of dimensionality and its applicationto approximate simple structure. *Psychometrika*, *64*(2), 213-249. https://doi.org/10.1007/BF02294536

Zhang, M. (2016). *Exploring dimensionality of scores for mixed-format tests* (Doctoral Dissertation, University of Iowa). https://ir.uiowa.edu/etd/2171/

Çok Boyutlu MTK, İkinci-düzey MTK ve Bifaktör Modelleri ile TIMSS Verisi için Toplam ve Alt Puanların Birlikte Kestirilmesi

Giriş

Eğitimde ölçme işlemi gerçekleştirilirken bir testin farklı yetenekleri ölçmesi yaygın bir durumdur. Bir testin alt testlerden oluştuğu durumlarda hâlihazırda birçok boyutluluk söz konusudur (Ackerman, Gierl, & Walker, 2003). Bu durumlarda test hem genel yeteneği hem de alt alanlar ile ilgili yetenekleri ölçer. Toplam puana ek olarak alt puanların da raporlanmasına ilişkin artan bir ilgi vardır. Toplam puan genele ilişkin bilgi verirken alt puanlar yanıtlayıcılara güçlü ve zayıf yönlerini detaylı bir şekilde verebilmesi açısından tanılayıcı bir değere sahiptir (Haberman & Sinharay, 2010).

Testlerin çoğunun çok boyutlu bir yapıya sahip olması ve alt alanlardan oluşması, yanıtlayıcılara ve eğitimcilere daha doğru bilgi sağlayan toplam puan ve alt puanların birlikte kestirimine olan ilgiyi arttırmıştır (Liu & Liu, 2017). Az sayıda çalışma toplam puan ve alt puanların birlikte kestirildiği yöntemleri ele almıştır (de la Torre & Song, 2009; Liu, Li, & Liu, 2018; Soysal & Kelecioğlu, 2018; Yao, 2010). De la Torre ve Song (2009) bu puanların birlikte kestiriminin sağlandığı ikinci-düzey madde tepki kuramı (MTK) yöntemini önermişlerdir. Yao (2010) çalışmasında toplam puan ve alt puanların birlikte raporlanabildiği dört yöntemi (tek boyutlu MTK, çok boyutlu MTK, ikinci-düzey MTK ve Bifaktör model) karşılaştırmıştır. Liu ve diğerleri (2018) 6 yeni bifaktör model önermiş ve bunları çok boyutlu MTK yöntemi ile karşılaştırmıştır.

Bu çalışmanın amacı, daha doğru ve güvenilir kestirimler elde etmek amacıyla toplam puan ve alt puanların birlikte kestirildiği yöntemlerin incelenmesidir. Bu kapsamda ele alınan yöntemler, çok boyutlu MTK, ikinci-düzey MTK ve Bifaktör modeldir. Bu çalışmanın az sayıda çalışma bulunan Alana katkı sağlayacağı düşünülmektedir. Ayrıca yapılan çalışmalardan farklı olarak ikili ve çoklu puanlanan maddelerin bir arada kullanıldığı karma-format bir test üzerinden analizlerin gerçekleştirilmiş olması önemli görülmektedir.

Yöntem

Sekizinci sınıflara uygulanan TIMSS 2015 matematik başarı testi birinci kitapçığında yer alan 35 maddeye verilen yanıtlar çalışma verisi olarak kullanılmıştır. Kayıp veri ile baş etme yöntemi olarak liste bazında silme kullanılmış ve kalan 5732 öğrenci verisi analize alınmıştır. TIMSS matematik başarı testi konu temelli dört alt alandan oluşmaktadır: sayılar (14 madde), cebir (9 madde), geometri (6 madde) ve veri ve olasılık (6 madde). Testi oluşturan 35 maddeden dördü çoklu puanlanırken geri kalan 31 madde ikili puanlanmaktadır.

Veri analizi için öncelikle boyutluluk analizi yapılmıştır. Bu amaçla Poly-DETECT ve doğrulayıcı faktör analizleri gerçekleştirilmiştir. İlgili veri için toplam puan ve alt puan kestirimleri ve bunlara ilişkin hatalar, BMIRT programı kullanılarak elde edilmiştir. Yöntemlerin değerlendirilmesi için kriter olarak ele alınan indeksler yetenek kestirimlerine ilişkin standart hatalar ve güvenirlik değerleridir. Standart hata ortalamaları arasındaki fark tekrarlı ölçümler için ANOVA ile değerlendirilirken toplam puan ve alt puanlar için güvenirlik kestirimi marjinal güvenirlik indeksi ile hesaplanmış ve yorumlanmıştır.

Sonuç ve Tartışma

Çalışma verisinin boyut yapısının incelenmesi amacıyla yapılan Poly-DETECT analizi sonuçları tek boyutluluktan sapma olduğunu göstermektedir (DETECT>.40; ASSI>.25; RATIO>.36). Dört alt testin her birinin bir faktör olarak ele alındığı modelin test edildiği doğrulayıcı faktör analizi sonuçları modelin veri ile uyumlu olduğunu göstermektedir (CFI>.95; TLI>.95; RMSEA<.05). Bu bulgular alt alan bazında çok boyutluluğun olduğunu kanıtlamaktadır.

Alt puan bazında yetenek parametrelerine ilişkin hataların ortalamasına bakıldığında çok boyutlu MTK yöntemi ile elde edilen yeteneklerin en düşük hata ile kestirildiği, en yüksek hata ortalamalarının Bifaktör model altında elde edildiği görülmektedir. Toplam puan için ise çok boyutlu MTK ve Bifaktör yöntemlerinin birbirine yakın ve düşük hata ortalamasına sahip olduğu ve ikinci-düzey MTK yönteminin diğer iki kestirim yönteminden az miktarda daha fazla hata ortalaması değerine sahip olduğu sonucuna ulaşılmıştır. Tekrarlı ölçümler için ANOVA sonuçları alt puanlar için elde edilen hata ortalamalarının kestirim yöntemine göre birbirinden anlamlı olarak farklılaştığını göstermektedir estimates ($F_{(1.726,\ 9893.087)\ \text{sayılar}} = 15465.323$, $p < .05$, *kısmi* $\eta 2 = .73$; $F_{(1.885,\ 10802.949)\ \text{cebir}} = 15288.071$, $p < .05$, *kısmi* $\eta 2 = .727$; $F_{(1.909,\ 10940.494)\ \text{geometri}} = 14196.309$, $p < .05$, *kısmi* $\eta 2 = .712$; $F_{(1.925,\ 11029.804)\ \text{very ve olasılık}} = 13418.317$, $p < .05$, *kısmi* $\eta 2 = .701$). Daha sonra yapılan ikili karşılaştırmalar, bütün ikili karşılaştırmalar istatistiksel olarak anlamlı olduğu bulunmuştur. Bu bulgu, alt puanlar için hata ortalamaları dikkate alındığında, ikinci-düzey MTK yönteminin anlamlı olarak diğer yöntemlerden daha az hata ile yetenek kestirimi yaptığını göstermektedir. Çalışma verisi için Bifaktör model ile kestirilen alt puanlar ise diğer iki yöntem kadar doğru değildir. Benzer şekilde, toplam puan bazında ise yetenek parametrelerine ilişkin hataların ortalamaları yöntemlere göre birbirinden anlamlı olarak farklılaşmaktadır ($F_{(1.692,\ 9696.490)\ \text{toplam}} = 8162.767$, $p < .05$, *kısmi* $\eta^2 = .588$). Analiz sonrasında yapılan ikili karşılaştırmalar bütün çiftlerin birbirinden anlamlı olarak farklılaştığını göstermektedir. Çalışma verisi için standart hata ortalaması en yüksek olan yöntem ikinci-düzey MTK'dir. Çok boyutlu MTK ve Bifaktör modele ilişkin standart hata ortalamaları birbirine yakın ve görece düşüktür.

Bir diğer değerlendirme kriteri olan güvenirlik için çalışmada ele alınan bütün yöntemlere göre elde edilen toplam puan ve alt puanlar için marjinal güvenirlik katsayısı hesaplanmıştır. Genel olarak bakıldığında, bütün alt alanlar için çok boyutlu MTK ve ikinci-düzey MTK yöntemleri ile elde edilen puanlara ilişkin güvenirlik değerleri, Bifaktör model ile elde edilen puanlara ilişkin güvenirlik değerlerinden yüksektir. İkinci-düzey MTK ile kestirilen alt puanlara ilişkin güvenirlik kestirimleri diğerlerinden daha yüksek ve hepsi 0,80'den yüksektir. Toplam puanlar için güvenirlik kestirimleri ise çok boyutlu MTK için 0,816, ikinci-düzey MTK için 0.815 ve Bifaktör model için 0.876 olup her üçü için de görece yüksek ve birbirine yakındır. Bifaktör model ile kestirilen güvenirlik ise diğerlerinden biraz daha yüksektir.

Sonuçlar genel olarak ele alındığında, çok boyutlu MTK ve ikinci-düzey MTK yöntemleri, alt puanların kestirim doğruluğu ve güvenirlik açısından benzer özellikler göstermektedir. Fakat ikinci-düzey MTK yöntemi, çok boyutlu MTK yönteminden nispeten daha düşük standart hata ortalamalarına ve daha yüksek güvenirlik kestirimlerine sahiptir. Benzer şekilde, de la Torre ve Song (2009) da çalışmalarında, ikinci-düzey MTK kullanıldığında alt puan kestirimlerinin daha etkili olduğunu belirtmişlerdir. Yao (2010) da çok boyutlu MTK ve ikinci-düzey MTK yöntemlerinin birbirine benzer sonuçlar ürettiğini bulmuştur. Bu çalışma kapsamında Bifaktör genel model, alt puan kestirimleri için en yüksek hataya ve en düşük güvenirliğe sahiptir. Liu ve diğerleri (2018) de elde ettiği sonuçlar ile puanların raporlanmasında orijinal faktör yöntemi olan Bifaktör modelin kullanılmasını tavsiye etmediğini belirtmektedir. Toplam puan kestirimi için ise çalışmada ele alınan üç yöntemin de birbirine yakın değerler vermesine rağmen en düşük hata ile yapılan kestirimin çok boyutlu MTK'ye ait olduğu görülmektedir. Güvenirlik değerleri incelendiğinde ise ilgili üç yöntemin de yüksek güvenirliğe sahip olmakla birlikte en yüksek güvenirlik kestiriminin Bifaktör model ile elde edildiği bulunmuştur.

Özetle, bu çalışma kapsamında gerçekleştirilen analizler, TIMSS 2015 verisi için toplam puan ve alt puanların birlikte kestirildiği yöntemlerden ikinci-düzey MTK yönteminin kullanılmasını önermektedir. Soysal ve Kelecioğlu (2018) da çalışmalarının bulguları doğrultusunda geniş ölçekli testlerde toplam puan ve alt puanların birlikte kestirilmesi için ikinci-düzey MTK'nin kullanılabileceğini önermektedir.

Bu çalışmada, verilere ilişkin gerçek model bilinmediğinden, üç yöntemin göreceli performansını incelemek için yalnızca gerçek veriler kullanılmıştır. Bu nedenle, diğer örneklemler için farklı sonuçlar elde edilmesi olası görünmektedir. Başka gerçek veriler kullanılarak araştırmanın tekrarlanabileceği önerilmektedir. Ayrıca, test uygulamalarında toplam ve alt puanların eşzamanlı olarak kestirilmesi gerektiğinde, puanları yanıtlayıcılara bildirmeden önce ilgili yöntemlerin göreceli performanslarının kontrol edilmesi önerilmektedir.