İSTANBUL UNIVERSITY CERRAHPAŞA

HAYEF: Journal of Education, 20(3); 221-230

DOI: 10.5152/hayef.2023.23065
Received: June 19, 2023
Accepted: July 17, 2023
Publication Date: September 25, 2023

HAYEF: Journal of Education

RESEARCH ARTICLE

Investigating Item Parameter Estimation Accuracy in Multidimensional Polytomous Data Under Various Conditions*

Serap BÜYÜKKIDIK, Hakan Yavuz ATAR

[1]Department of Measurement and Evaluation in Education, Sinop University Faculty of Education, Sinop, Türkiye
[2]Department of Measurement and Evaluation in Education, Gazi University Faculty of Education, Ankara, Türkiye

Abstract

In this study, the root mean square error values of item parameters' estimation in a two-dimensional structure condition were examined under different conditions, considering three and five categories with different algorithms (Expectation–Maximization, Metropolis–Hastings Robbins–Monro, Quasi-Monte Carlo Expectation–Maximization). The simulation conditions included two different sample sizes (1500 and 3000) in a two-dimensional structure, three test lengths (12, 24, and 36), three different interdimensional correlations (0.20, 0.50, and 0.80), and two different category numbers (three and five). Analyses were conducted with three algorithms and the graded response model from the multidimensional item response theory in 36 different conditions with 100 replications. When the errors were examined in terms of the root mean square error, an increase in the number of categories resulted in a partial decrease in most item parameters under the condition of 1500 sample size. For researchers conducting analyses in the polytomous multidimensional item response theory, it is recommended to use as large a sample as possible, at least 24 items, five categories, and the Quasi-Monte Carlo Expectation–Maximization algorithm.

Keywords: Expectation–Maximization, graded response model, Metropolis–Hastings Robbins–Monro, multidimensional item response theory, Quasi-Monte Carlo Expectation–Maximization

Introduction

Important decisions about individual levels, institutional levels, and public policies are made based on the results obtained from measurement tools (Kolen & Brennan, 2014). In the 21st century, multiple-choice tests that score items as either correct or incorrect, that is, 1–0 (binary), are not sufficient for measuring higher-order skills such as problem-solving, critical thinking, and creativity. Therefore, polytomous items provide more information, especially in measuring these skills, compared to dichotomously scored items (Donoghue, 1993; Embretson & Reise, 2000; Lukhele et al., 1994). Moreover, polytomous items are also used for measuring non-cognitive characteristics. Due to the increasing preference for performance-based assessment, which is a constructed response measurement tool, the effectiveness and informativeness of constructed response items in measuring higher-order thinking skills, as well as the widespread use of polytomous items in measuring personality traits and attitudes, the use of item response theory (IRT) models for polytomous scored items has become widespread (Embretson & Reise, 2000; Kim & Cohen, 2002).

Item response theory has various applications in multidimensional structures where the assumption of unidimensionality is violated (Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Hambleton et al., 1991). The advantages of IRT are demonstrated through appropriate model selection (Hambleton & Swaminathan, 1985). Many studies in the literature have shown that the unidimensionality assumption of the obtained data is violated (Lee, 2007). In such cases, unidimensional IRT models may be insufficient, and more complex models are needed like multidimensional IRT (MIRT) (Reckase, 1997).

Measurement tools used in education generally consist of multiple homogeneous subtests, where each item contributes to only one dimension and multiple dimensions are present (Ackerman, 1996). Examples of such structures include international studies such as "Programme for International Student Assessment" (PISA) and "Trends in International Mathematics and Science Study" (TIMSS). For example, in TIMSS, multiple cognitive subtests such as numbers, algebra, data, and probability form a structure, such as mathematics (Bulut, 2013). In cases where each subtest is one-dimensional and these subtests form a multidimensional structure, the literature includes between-item (Hartig & Höhler, 2009; Wang et al., 2004), simple structured (Ackerman et al., 2003; Cole & Paek, 2023; Zhang, 2012), or multi-unidimensional (Kuo, 2015; Kuo & Sheng, 2016; Mun et al., 2019; Sheng, 2005, 2008; Sheng & Wikle, 2007, 2008) are called MIRT models. Since these models are easier to interpret than complex models, most MIRT models are considered within the scope of simple structured IRT models (Ackerman et al., 2003). We used simple structured MIRT models in this research because of their widespread use.

*This research was created from a part of the doctoral thesis titled "Comparison of Parameter Estimation Based on Multidimensional Item Response Theory In Polytomous Items".

Corresponding Author: Serap BÜYÜKKIDIK, **E-mail:** sbuyukkidik@gmail.com

Cite this article as: Büyükkıdık, S., & Atar, H. Y. (2023). Investigating item parameter estimation accuracy in multidimensional polytomous data under various conditions. *HAYEF: Journal of Education*, 20(3), 221-230.

In the literature, there has not been a completely identical study in which graded response model (GRM) and parameter estimations are done with different algorithms in different simulation conditions in simple-structure polytomous data. While MIRT studies are using different algorithms (Cai, 2010; Kuo & Sheng, 2016), it has been observed that algorithms in the R programming language ("Expectation–Maximization (EM), Quasi-Monte Carlo Expectation–Maximization (QMCEM), and Metropolis–Hastings Robbins–Monro (MHRM)") are not used in these studies. Chalmers (2018) stated that the "MHRM" and "QMCEM" algorithms are effective in the case of more than three dimensions, but there is no research yet on which algorithm is more effective in polytomous items in a two-factor structure.

It has been observed that the studies on MIRT in the literature were generally carried out on items scored binary (e.g., Bolt & Lall, 2003; Çakıcı Eser & Gelbal, 2015; Chalmers, 2012; Garnier-Villarreal et al., 2021; Kalkan, 2022; Kose & Demirtasli, 2012; Lee, 2007, 2012; Mun et al., 2019; Özer Özkan, 2014; Sahin et al., 2019; Sünbül, 2011). Only a few of these studies compared the EM, QMCEM, and MHRM algorithms in two categories of data under various simulation conditions (e.g., Kalkan, 2022; Garnier-Villarreal et al., 2021). Kalkan (2022) suggested in his research to use the MHRM algorithm in two-dimensional two-category (binary) data. There are also polytomous MIRT studies (Cai, 2010; Cole & Paek, 2023; Kuo, 2015; Kuo & Sheng, 2016; Gül, 2015; Martelli, 2014; Martelli et al., 2016; Jiang et al., 2016). Among these studies, Kuo and Sheng (2016) used MIRT estimation algorithms (Bock-Aitkin expectation-maximum algorithm (BAEM), adaptive quadrature (AQ), Gibbs sampling, Metropolis–Hastings, Gibbs in Hastings (Hastings-within-Gibbs), blocked Metropolis, and MHRM) in the GRM. They compared using IRTPRO, BMIRT, and MATLAB software. They stated that while the algorithms produce similar results in the simulation condition where the interdimensional correlation is low, the Gibbs algorithm in Hastings produces better results when it is medium and high.

When all these studies in the literature are examined, there is no similar study in which EM, MHRM, and QMCEM algorithms are compared under different conditions by using polytomous data. It is thought that this research will contribute to the literature in terms of providing information on which algorithm will make more accurate estimations in a two-dimensional structure, three- and five-category data. In addition, research on how the differentiation of the number of categories in polytomous items changes under different conditions and algorithms has not been conducted yet.

The accuracy of item parameter estimations was tested under different conditions by using "EM, MHRM, and QMCEM" algorithms included in the "mirt" package in the R programming language. The algorithms used in the research are briefly explained.

Expectation–Maximization

The EM algorithm is an iterative process involving expectation (E) and maximization (M) steps to find the maximum likelihood function (Dempster et al., 1977). Since it is necessary to calculate high-dimensional integrals in probability functions in high-dimensional models, especially in item parameter estimations, in the EM algorithm, the ability to generalize is limited. As the number of dimensions increases, the number of quadrature points increases considerably. Therefore, this algorithm is useless in models with three or four factors (Houts & Cai, 2016). However, there is a need for research on whether the EM algorithm makes good predictions in two-dimensional models.

Metropolis–Hastings Robbins–Monro

Since the MHRM algorithm performs analyses using Robbins–Monro type data augmentation developed by Robbins and Monro and random assignment together, it is recommended by Cai (2010) in high-dimensional models in MIRT analysis. Houts and Cai (2016) stated that analyses were performed in three steps with this algorithm. Martin-Fernandez and Revuelta (2017) stated that MHRM is a useful algorithm in high-dimensional structures that generates maximum likelihood and modal or expected a-posteriori point estimation solutions for marginal likelihood based on MIRT analyses.

Quasi-Monte Carlo Expectation–Maximization

Although Quasi-Monte Carlo and Monte Carlo are handled in similar ways, the QMCEM algorithm is a version of the EM algorithm where the E-step is replaced by the Monte Carlo approach (Jank, 2005). In recent times, QMCEM has found application across diverse domains, including mathematical finance. The existing body of literature highlights the necessity for further exploratory studies employing this algorithm to enhance the research, especially in the measurement and evaluation in education.

Aim of the Current Research

In this study, it has been discussed how the item parameter estimation will be affected if different algorithms are used in the conditions of different sample sizes, correlation, number of categories, and number of items within the scope of polytomous data. For this purpose, answers to the following questions were sought:

What are the root mean square error (RMSE) values of the item parameters estimated in the EM algorithm with the GRM in simple structured two-dimensional data, three- and five-category measurement tools, in the conditions of different sample sizes (1500 and 3000), measurement tool length (12, 24, and 36), and correlation between dimensions (0.2, 0.5, and 0.8)?

What are the RMSE values of the item parameters estimated in the MHRM algorithm with the GRM in simple structured two-dimensional data, three- and five-category measurement tools, in the conditions of different sample sizes (1500 and 3000), measurement tool length (12, 24, and 36), and correlation between dimensions (0.2, 0.5, and 0.8)?

What are the RMSE values of the item parameters estimated in the QMCEM algorithm with the GRM in simple structured two-dimensional data, three- and five-category measurement tools, in the conditions of different sample sizes (1500 and 3000), measurement tool length (12, 24, and 36), and correlation between dimensions (0.2, 0.5, and 0.8)?

How do RMSE values change when different algorithms are used in two-dimensional, three- and five-category data? (See Table 1).

Methods

Type of Research

This research has the characteristics of simulation research in terms of considering item parameter estimations when analyses are performed with a multidimensional GRM by considering different conditions in polytomous items.

Simulation Conditions

Ethic committee approval is not necessary for this research because this is a simulation study. Although the simulation conditions and the

Table 1.
Simulation Conditions Employed in This Research

Sample Size	Test Length	Correlation Between Dimensions	Number of Category
1500	12	0.2	3
3000	24	0.5	5
	36	0.8	

number of replications differ in the studies in the literature, the studies in the literature were used to determine the simulation conditions (see Table 1).

Sample Size

In the domain of simulation studies pertaining to IRT, the scope of sample size consideration typically ranges from 500 to 5000 participants, as evidenced by investigations undertaken by researchers such as Bolt & Lall (2003), Cole & Paek (2023), Çakıcı Eser & Gelbal (2015), de la Torre & Hong (2010), Guo & Choi (2023), Gül (2015), Kuo (2015), Kuo & Sheng (2016), Martelli et al. (2016), Mun et al. (2019), Jiang et al. (2016), Yao & Boughton (2007), Yao (2010), Yavuz & Hambleton (2017), and Zhang (2012). The specific sample size of 1500 was adopted in the work of Bulut (2013), wherein it was deemed sufficient for MIRT applications. In congruence with these researches, the present study pursued simulations employing a sample size ranging from 1500 to 3000 participants.

Fixed Number of Dimensions

While the range of dimensions considered in MIRT simulation studies spans from two to seven dimensions according to existing literature, an observable trend indicates a predominant focus on two and three dimensions (Bulut, 2013; Çakıcı Eser & Gelbal, 2015; de la Torre, 2008, 2009; de la Torre & Hong, 2010; de la Torre & Patz, 2005; de la Torre, Song & Hong, 2011; Guo & Choi, 2023; Gül, 2015 Kalkan, 2022). Within the context of the simulation study, this investigation delved into two-dimensional structures, mirroring the approach adopted in Cai's research (2010).

Measuring Tool Length (Number of Items)

Within the corpus of existing literature, examination of simulation studies reveals a range of item quantities spanning from 10 to 240 (Bolt & Lall, 2003; Bulut, 2013; Cole & Paek, 2023; Çakıcı Eser & Gelbal, 2015; de la Torre, 2008, 2009; de la Torre & Hong, 2010; de la Torre & Patz, 2005; de la Torre et al., 2011; Forero & Maydeu-Olivares, 2009; Garnier-Villarreal et al., 2021; Guo & Choi, 2023; Kalkan, 2022; Kuo, 2015; Kuo & Sheng, 2016; Martelli et al., 2016; Mun et al., 2019; Jiang et al., 2016; Yao, 2010; Yavuz & Hambleton, 2017). This investigation specifically addresses conditions encompassing a minimum of 12 items within a three-dimensional framework, with no less than four items allocated to each dimension. Furthermore, conditions featuring 24 and 36 items were incorporated, thereby accounting for the prevalence of frequently employed items in measurement scales.

Correlation Between Dimensions

In the realm of simulation studies, a spectrum of correlation values ranging from 0 to 0.9 across dimensions has been investigated by various scholars (Bolt & Lall, 2003; Bulut, 2013; de la Torre, 2008, 2009; de la Torre & Hong, 2010; de la Torre & Patz, 2005; de la Torre et al., 2011; Guo & Choi, 2023; Gül, 2015; Jiang et al., 2016; Kalkan, 2022; Koğar, 2014; Kuo, 2015; Kuo & Sheng, 2016; Yao, 2010; Yao & Boughton, 2007; Yavuz & Hambleton, 2017). In this study, simulation was conducted with three different interdimensional correlation values, representing 0.2 for low correlation, 0.5 for medium correlation, and 0.8 for high correlation, respectively. These values were also used in the studies of Kuo (2015) and Kuo and Sheng (2016) with the GRM within the scope of MIRT.

Number of Categories

In IRT simulation studies, mostly binary (1–0) data were used (e.g., Chalmers, 2012; Garnier-Villarreal et al., 2021; Guo & Choi, 2023; Kalkan, 2022; Mun et al., 2019; Sahin et al., 2019). In various investigations, the researchers generated and analyzed datasets encompassing different numbers of categories. For instance, de la Torre (2008) examined datasets involving two, three, and four categories, while Martelli et al. (2016) explored datasets comprising three, four, and five categories. Additionally, Cai (2010) and Cole & Paek (2023) specifically investigated datasets consisting of three categories. In this study, since the **constructed**-response cognitive items in large-scale tests such as PISA and TIMSS are mostly three-category and five-point grading is used more in the scales, the number of categories is considered as three and five, and this simulation study has concentrated on category counts of three and five.

Number of Replications

In the domain of IRT-focused simulation studies, the number of replications varies across different investigations. Specifically, the number of replications adopted across distinct conditions is reported as follows: five replications in the study by Bolt & Lall (2003), 10 replications in studies conducted by Choi (1996), Sheng & Wikle (2007; 2008), Lee (2012), Kuo (2015), Kuo & Sheng (2016), Martelli (2014), and Martelli et al. (2016), 20 replications in research undertaken by Fu et al. (2010), Koğar (2014), Sayın & Gelbal (2016), and Sengul Avsar & Tavsancil (2017), 25 replications in works conducted by Çakıcı Eser & Gelbal (2015), de la Torre & Hong (2010), Gül (2015), and Mun et al. (2019), 30 replications in the study by Jiang et al. (2016), and notably, studies employing 500 replications, as seen in the research by Garnier-Villarreal et al. (2021). Harwell et al. (1996) highlighted that a minimum of 25 replications is advisable in IRT-based Markov Chain Monte Carlo studies. In alignment with this guidance, the present study employed 100 replications, akin to the approach followed in the research conducted by Cole and Paek (2023).

The simulation study is summarized as follows:

Number of conditions: 2 sample sizes (1500, 3000) × 1 dimension (2) × 3 measurement tool lengths (12, 24, 36) × 3 interdimensional correlations (0.2, 0.5, 0.8) × 2 categories (3.5) = 36 conditions
Used models: Simple structured GRM
Algorithms: EM, MHRM, QMCEM
Number of replications: 100

Verification of Generated Data

To verify the data in the study, data were produced for each condition, and fit indices, factor loads, correlations between dimensions, and multivariate normality (relative multivariate kurtosis) were examined. When the fit indices of the models obtained under 36 conditions for the simple structured data were examined, it was confirmed that the structure was provided. All these results confirm that the generated data set has been produced with the desired properties.

Generation and Analysis of Data

Although the a parameter is in the range of $(-\infty,+\infty)$ theoretically, it takes values between −2.80 and +2.80 in applications (Baker, 2001). Cai (2010) carried out a simulation study by producing data by keeping the a parameters in the range of 1.1–2.6 in his simulation study with polytomous data. Jiang, Wang, and Weiss (2016) and Bulut and Sünbül (2017) produced data with a parameter of $a\sim(1.1, 2.8)$ from a uniform distribution in their simulation studies within the scope of MIRT on polytomous data. In this study, data as $a\sim(1.1, 2.8)$ were produced from a uniform distribution of parameter a.

While the parameter b takes values in the range of $(-\infty,+\infty)$ in theory, it usually takes values in the range of $(-3,+3)$ in applications (Baker, 2001). De Ayala (1994) stated that parameter b has values in the range of $[-2,+2]$. In other sources, the b parameter usually takes values between −2.00 and +2.00, and it is stated that items with a b_i value close to −2.00 are very easy and items close to +2.00 are very difficult (Hambleton et al., 1991; Hambleton & Swaminathan, 1985). The initial parameter, denoted as b_1, is designated through a random sampling process from a uniform distribution to ensure the sequential

nature of the b parameter. Specifically, b_1 is drawn from the uniform distribution U(0.67, 2). Subsequently, the remaining b parameters are derived iteratively as follows: $b_2 = b_1 - U(0.67, 1.34)$, $b_3 = b_2 - U(0.67, 1.34)$, $b_4 = b_3 - U(0.67, 1.34)$, and so forth. This approach takes into account the distinctive attributes of GRM within the context of five-category data. Notably, established references from existing literature (Jiang et al., 2016; Bulut & Sünbül, 2017) have been consulted as guidelines for formulating the b parameter in this manner.

In the research, the estimation of the ability parameters of multi-dimensional and categorical simulative data under GRM was made under different conditions. R (studio) programming language was used in the analysis of the data. In the first stage of the research, the distribution characteristics specified through the "mirt" and "MASS" packages were taken into account while generating the data. To perform the analyses faster, parallel calculations (parallel computing) were made using the "doParallel" package and the registerDoParallel( ) command, and all cores in the computer were utilized. In the other step, the RMSE values of the estimated parameter were calculated. The following formulas were used to calculate the RMSE and bias values used in the parameter verification or measurement precision study in the research.

The following formula was used to calculate the RMSE values used in the parameter verification or measurement precision study in the research.

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{K}\left(\hat{\delta}\mathbf{i} - \delta\mathbf{i}\right)^2}{\mathbf{K}}}$$

$\hat{\delta}i$: Estimated parameter value for item i.

δi is the actual parameter value for item i.

K: number of items

Instead of δ value, for example, in two-dimensional and five-category data, calculations are made by replacing the relevant item (a_1, a_2, b_1, b_2, b_3, and b_4). The distance to the 0 point is taken into account in the interpretation of the RMSE, that is, the accuracy of the estimations increases as it approaches 0. In other words, while smaller values closer to zero increase the measurement precision, larger values far from zero decrease the measurement precision. In some investigations, the RMSE value has been observed to be evaluated in relation to a benchmark value of 0.1 (Browne & Cudeck, 1993; DeMars, 2003; Hu & Bentler, 1999).

In the last stage, 100 replications were performed with the (for) command in a loop.

Results

In this section, the results related to the research problems are discussed.

Results Related to the First Sub-Problem

The RMSE values of the item parameters in the EM algorithm for two-dimensional simple structured measurement tools are given in Table 2.

As shown in Table 2, RMSE values varied between 0.04 and 0.14. While the values varied between 0.04 and 0.14 for the a_1 parameter in three-category data, they varied between 0.04 and 0.08 in five-category data. While the values varied between 0.04 and 0.11 for the a_2 parameter in three-category data, they varied between 0.04 and 0.09 in five-category data. While the values varied between 0.04 and 0.07 for parameter b_1 for three-category data, they varied between 0.04 and 0.08 in data with five categories. In three-category data, the b_2 parameter ranged from 0.04 to 0.08; it ranged from 0.04 to 0.07 for five-category items. In items with five categories, the b_3 parameter ranged from 0.04 to 0.08, and the b_4 parameter ranged from 0.05 to 0.10. Increasing the number of items and sample size contributes to measurement precision by causing a decrease in RMSE values. The change in correlation values did not cause a pattern in the change of RMSE values. The highest RMSE value belonged to parameter a_1. In two-dimensional simple structured polytomous data, the b parameter had lower RMSE values than the a parameter.

In Figure 1, the graph formed by considering the RMSE values of the item parameters obtained by the EM algorithm in 1500 and 3000 samples of two-dimensional, three- and five-category measurement tools was given.

Table 2.
RMSE Values of Item Parameters in EM Algorithm in Two-Dimensional Structure

Conditions			**Three-Category Data**				**Five-Category Data**					
Sample Size	**Test Length**	**Correlation Between Dimensions**	a_1	a_2	b_1	b_2	a_1	a_2	b_1	b_2	b_3	b_4
1500	12	0.2	0.14	0.10	0.07	0.08	0.08	0.09	0.07	0.07	0.08	0.09
		0.5	0.10	0.10	0.07	0.08	0.07	0.08	0.07	0.07	0.06	0.10
		0.8	0.10	0.11	0.06	0.08	0.07	0.07	0.06	0.06	0.06	0.10
	24	0.2	0.05	0.10	0.06	0.06	0.07	0.07	0.06	0.05	0.08	0.09
		0.5	0.06	0.11	0.05	0.06	0.07	0.07	0.07	0.05	0.07	0.07
		0.8	0.07	0.11	0.07	0.06	0.07	0.07	0.07	0.06	0.07	0.08
	36	0.2	0.05	0.07	0.06	0.06	0.07	0.06	0.08	0.05	0.07	0.08
		0.5	0.05	0.07	0.05	0.06	0.07	0.06	0.07	0.05	0.08	0.08
		0.8	0.06	0.07	0.06	0.06	0.07	0.05	0.08	0.05	0.07	0.08
3000	12	0.2	0.06	0.07	0.05	0.05	0.07	0.09	0.05	0.06	0.05	0.07
		0.5	0.05	0.07	0.04	0.05	0.06	0.07	0.04	0.06	0.05	0.07
		0.8	0.04	0.08	0.04	0.05	0.05	0.05	0.05	0.04	0.05	0.06
	24	0.2	0.06	0.07	0.05	0.06	0.04	0.08	0.05	0.04	0.06	0.07
		0.5	0.05	0.06	0.05	0.06	0.05	0.07	0.06	0.05	0.05	0.06
		0.8	0.04	0.06	0.05	0.06	0.05	0.06	0.04	0.05	0.06	0.06
	36	0.2	0.05	0.04	0.05	0.04	0.05	0.04	0.05	0.04	0.05	0.07
		0.5	0.06	0.04	0.05	0.04	0.04	0.04	0.05	0.04	0.05	0.07
		0.8	0.06	0.05	0.05	0.05	0.04	0.05	0.05	0.04	0.04	0.05

Note: EM = Expectation–Maximization; RMSE = Root mean square error.

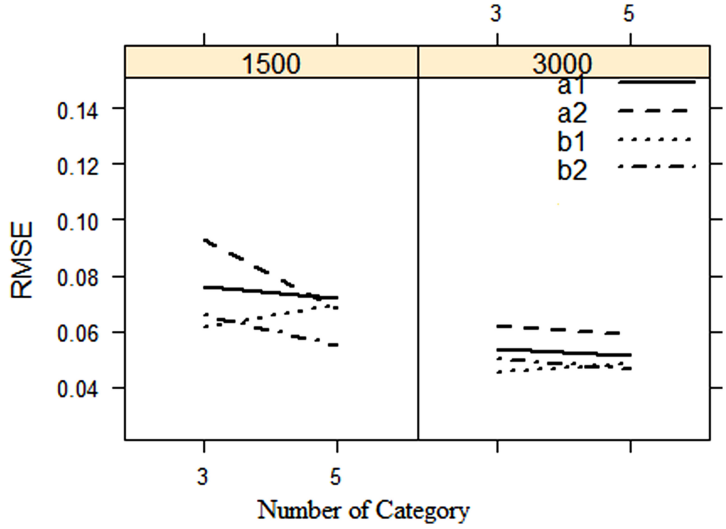

Figure 1.
Analysis of RMSE Values of Item Parameters Estimated by EM Algorithm of Two-Dimensional, Three- (Left) and Five-Category (Right) Measurement Tools by Number of Categories. EM = Expectation–Maximization, RMSE = Root mean square error.

In Figure 1, it was seen that the increase in the number of categories in 1500 samples, except for the a_2 and b_2 parameters, does not cause a decrease in the RMSE values of the item parameters. The change in the number of categories in two-dimensional structures was generally not effective in the change of the RMSE values of the item parameters estimated by the EM algorithm (except for the values of the a_2 and b_2 parameters in 1500 samples).

Results Regarding the Second Sub-Problem

The results for the second research question are given later.

The RMSE values of the item parameters in the MHRM algorithm for two-dimensional simple structured measurement tools are given in Table 3.

As shown in Table 3, RMSE values varied between 0.03 and 0.16. While the values varied between 0.04 and 0.16 for the a_1 parameter in three-category data, they varied between 0.04 and 0.08 in five-category data. While the values varied between 0.05 and 0.11 for the a_2 parameter in three-category data, they varied between 0.03 and 0.09 in five-category data. While the values varied between 0.04 and 0.07 for parameter b_1 for three-category data, they varied between 0.04 and 0.08 in data with five categories. In three-category data, the b_2 parameter ranged from 0.04 to 0.09; it ranged from 0.04 to 0.08 for five-category items. In items with five categories, the b_3 parameter ranged from 0.05 to 0.09, and the b_4 parameter ranged from 0.06 to 0.12. Increasing the number of items and sample size contributed to measurement precision by causing a decrease in RMSE values. The change in correlation values did not cause a pattern in the change of RMSE values. The highest RMSE value belonged to parameter a_1. In two-dimensional simple structured polytomous data, the *b* parameter had lower RMSE values than the *a* parameter.

In Figure 2, the graph formed by considering the RMSE values of the item parameters obtained by the MHRM algorithm in 1500 and

Table 3.
RMSE Values of Item Parameters in MHRM Algorithm in Two-Dimensional Structure

Conditions			**Three-Category Data**				**Five-Category Data**					
Sample Size	**Test Length**	**Correlation Between Dimensions**	a_1	a_2	b_1	b_2	a_1	a_2	b_1	b_2	b_3	b_4
1500	12	0.2	0.16	0.10	0.07	0.09	0.08	0.09	0.07	0.08	0.08	0.10
		0.5	0.11	0.10	0.07	0.08	0.06	0.09	0.07	0.07	0.06	0.11
		0.8	0.10	0.10	0.07	0.08	0.07	0.06	0.06	0.07	0.06	0.12
	24	0.2	0.05	0.10	0.06	0.06	0.07	0.07	0.07	0.06	0.08	0.09
		0.5	0.06	0.11	0.05	0.06	0.06	0.07	0.07	0.05	0.07	0.07
		0.8	0.07	0.11	0.06	0.07	0.07	0.07	0.08	0.07	0.08	0.09
	36	0.2	0.05	0.07	0.06	0.06	0.07	0.06	0.08	0.05	0.09	0.08
		0.5	0.05	0.06	0.05	0.06	0.07	0.06	0.07	0.04	0.08	0.07
		0.8	0.05	0.07	0.07	0.06	0.07	0.05	0.08	0.06	0.09	0.08
3000	12	0.2	0.06	0.08	0.04	0.05	0.06	0.09	0.05	0.07	0.06	0.07
		0.5	0.05	0.08	0.04	0.05	0.06	0.07	0.04	0.06	0.06	0.07
		0.8	0.04	0.07	0.04	0.04	0.05	0.06	0.05	0.04	0.05	0.06
	24	0.2	0.07	0.08	0.05	0.07	0.05	0.07	0.05	0.05	0.06	0.08
		0.5	0.05	0.06	0.04	0.06	0.05	0.07	0.06	0.05	0.06	0.06
		0.8	0.04	0.07	0.05	0.06	0.06	0.05	0.05	0.05	0.06	0.06
	36	0.2	0.05	0.05	0.06	0.05	0.05	0.03	0.05	0.05	0.06	0.07
		0.5	0.06	0.05	0.05	0.04	0.04	0.04	0.06	0.05	0.05	0.07
		0.8	0.06	0.07	0.05	0.05	0.04	0.06	0.06	0.05	0.05	0.07

Note: MHRM = Metropolis–Hastings Robbins–Monro; RMSE = Root mean square error.

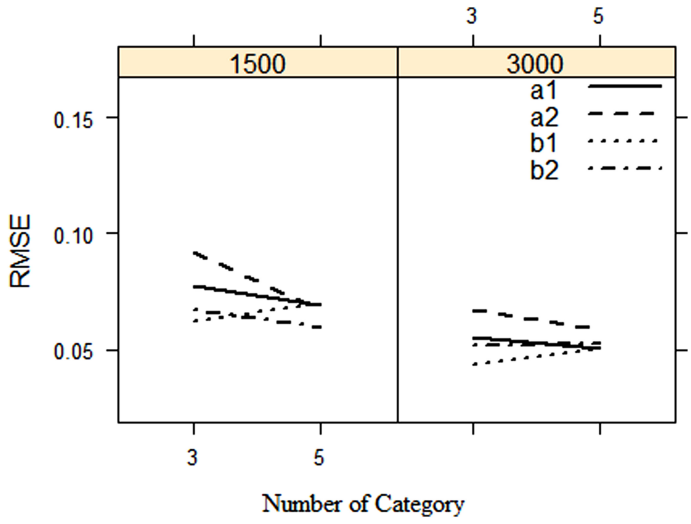

Figure 2.
Analysis of the RMSE Values of the Item Parameters Estimated by the MHRM Algorithm of Two-Dimensional, Three- and Five-Category Measurement Tools by Category Number. MHRM = Metropolis–Hastings Robbins–Monro; RMSE = Root mean square error.

3000 samples of two-dimensional, three- and five-category measurement tools was given.

In Figure 2, it was seen that the increase in the number of categories under conditions other than the a_2 parameter of 1500 samples did not cause a decrease in the RMSE values of the item parameters. In two-dimensional structures, the change in the number of categories was generally not effective in the change of the RMSE values of the item parameters estimated by the MHRM algorithm (except for the values of 1500 sample a_2 parameters).

Results Regarding the Third Sub-Problem

The results for the third research question are given later.Table 4 provides the RMSE values of the item parameters within the framework of the QMCEM algorithm applied to two-dimensional simple structured data.

As shown in Table 4, RMSE values varied between 0.04 and 0.14. While the values varied between 0.04 and 0.14 for the a_1 parameter in three-category data, they varied between 0.04 and 0.08 in five-category data. While the values varied between 0.04 and 0.11 for the a_2 parameter in three-category data, they varied between 0.04 and 0.10 in five-category data. While the values varied between 0.04 and 0.07 for parameter b_1 for three-category data, they varied between 0.04 and 0.08 in data with five categories. In three-category data, the b_2 parameter ranged from 0.04 to 0.08; it ranged from 0.04 to 0.07 for five-category items. In items with five categories, the b_3 parameter ranged from 0.05 to 0.08, and the b_4 parameter ranged from 0.05 to 0.11. Increasing the number of items and sample size contributed to measurement precision by causing a decrease in RMSE values. The change in correlation values did not cause a pattern in the change of RMSE values. The highest RMSE value belonged to parameter a_1. In two-dimensional simple structured polytomous data, the *b* parameter had lower RMSE values than the *a* parameter.

The graph formed by considering the RMSE values of the item parameters obtained by the QMCEM algorithm in 1500 and 3000

Table 4.
RMSE Values of Item Parameters in QMCEM Algorithm in Two-Dimensional Structure

Conditions			Three-Category Data				Five-Category Data					
Sample Size	Test Length	Correlation Between Dimensions	a_1	a_2	b_1	b_2	a_1	a_2	b_1	b_2	b_3	b_4
1500	12	0.2	0.14	0.10	0.07	0.08	0.08	0.10	0.07	0.07	0.08	0.09
		0.5	0.10	0.10	0.07	0.08	0.07	0.09	0.07	0.07	0.06	0.10
		0.8	0.10	0.10	0.06	0.08	0.07	0.07	0.06	0.06	0.06	0.11
	24	0.2	0.05	0.10	0.06	0.06	0.07	0.07	0.06	0.05	0.08	0.09
		0.5	0.06	0.11	0.05	0.06	0.07	0.08	0.07	0.05	0.07	0.07
		0.8	0.07	0.11	0.07	0.06	0.07	0.07	0.07	0.06	0.07	0.08
	36	0.2	0.05	0.07	0.06	0.06	0.07	0.06	0.08	0.05	0.07	0.08
		0.5	0.05	0.07	0.05	0.06	0.07	0.06	0.07	0.04	0.07	0.08
		0.8	0.06	0.07	0.06	0.06	0.07	0.05	0.08	0.05	0.08	0.08
3000	12	0.2	0.06	0.08	0.05	0.05	0.07	0.09	0.05	0.06	0.05	0.07
		0.5	0.05	0.07	0.04	0.05	0.06	0.07	0.04	0.06	0.05	0.07
		0.8	0.04	0.08	0.04	0.05	0.05	0.05	0.05	0.04	0.05	0.06
	24	0.2	0.06	0.07	0.05	0.06	0.04	0.08	0.05	0.04	0.06	0.07
		0.5	0.05	0.06	0.05	0.06	0.05	0.07	0.06	0.05	0.05	0.06
		0.8	0.04	0.07	0.05	0.05	0.05	0.06	0.04	0.05	0.05	0.06
	36	0.2	0.05	0.04	0.05	0.04	0.05	0.04	0.05	0.04	0.05	0.07
		0.5	0.06	0.04	0.05	0.04	0.04	0.04	0.05	0.04	0.05	0.07
		0.8	0.06	0.06	0.05	0.05	0.04	0.05	0.05	0.04	0.05	0.05

Note: RMSE = Root mean square error; QMCEM = Quasi-Monte Carlo Expectation–Maximization.

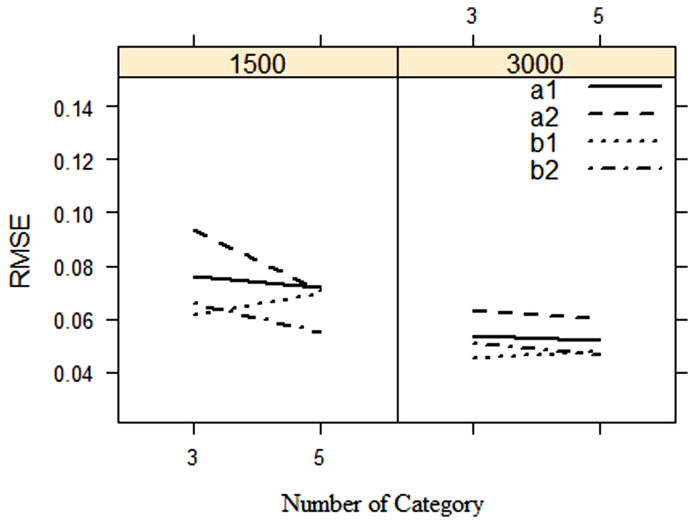

Figure 3.
Analysis of the RMSE Values of the Item Parameters Estimated by the QMCEM Algorithm of Two-Dimensional Measurement Tools According to the Number of Categories. RMSE = Root mean square error; QMCEM = Quasi-Monte Carlo Expectation–Maximization.

samples of two-dimensional, three- and five-category measurement tools is given in Figure 3.

As shown in Figure 3, it was seen that the increase in the number of categories in 1500 samples, except for the a_2 and b_2 parameters, did not cause a decrease in the RMSE values of the item parameters. The change in the number of categories in two-dimensional structures was generally not effective in the change of the RMSE values of the item parameters estimated by the QMCEM algorithm (except for the values of the a_2 and b_2 parameters in 1500 samples).

Results Regarding the Fourth Sub-Problem

Whether the RMSE values of two-dimensional, three- and five-category measurement tools changed when different algorithms were used was discussed in this research problem. In Figure 4, the graph of the RMSE values of the item parameters estimated from the EM, MHRM, QMCEM algorithms, and the two-dimensional, three- and five-category multidimensional GRM data sets is given.

RMSE values in two-dimensional, three- and five-category measurement tools were similar when different algorithms were used.

Discussion and Conclusion and Recommendations

In this study, it was aimed to examine how the RMSE values of item parameter estimations differ under different algorithms under different simulation conditions in simple structured multidimensional data with three and five categories. For this purpose, the change in the number of categories in the two-dimensional structure in EM, MHRM, and QMCEM algorithms with different sample sizes (1500 and 3000), measurement tool length (12, 24, and 36), and correlation between dimensions (0.2, 0.5, and 0.8) was examined to determine what kind of differences it caused in the RMSE values in the item parameter estimations under the conditions.

Generally, augmenting the sample size within the context of a two-dimensional structure led to a convergence of RMSE values toward zero, thereby enhancing the precision of measurement estimations. A comprehensive review of pertinent literature regarding the impact of sample size conditions on parameter estimations (Bolt & Lall, 2003; Cole & Paek, 2023; Çakıcı Eser & Gelbal, 2015; de la Torre & Patz, 2005; de la Torre & Hong, 2010; Gül, 2015; Jiang, Wang & Weiss, 2016; Kuo, 2015, Kuo & Sheng, 2016; Lee, 2012; Martelli et al.,

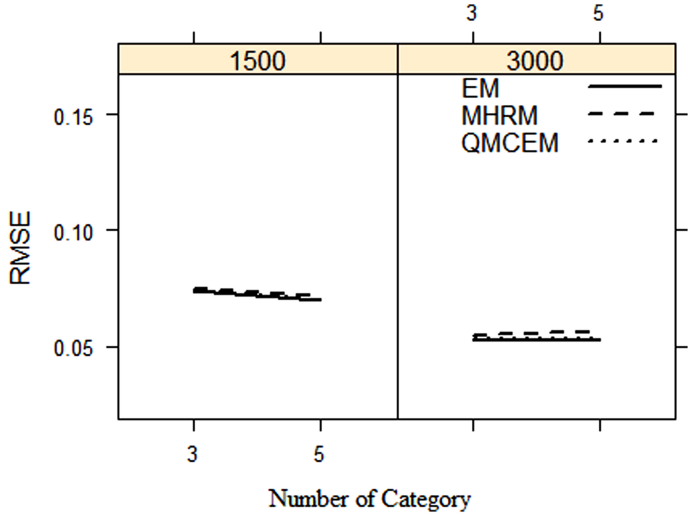

Figure 4.
RMSE Values for Item Parameters Estimated from EM, MHRM, QMCEM Algorithms in Two-Dimensional, Three- and Five-Category MGRM Datasets. EM = Expectation–Maximization; MGRM = Multidimensional grade response model; MHRM = Metropolis–Hastings Robbins–Monro; RMSE = Root mean square error; QMCEM = Quasi-Monte Carlo Expectation–Maximization.

2016; Sheng & Wikle, 2007; Sünbül, 2011; Yao & Boughton, 2007; Yao, 2010; Yavuz & Hambleton, 2017; Zhang, 2012), highlighted a consistent result: as the sample size increases, the RMSE values tend to diminish (Bolt & Lall, 2003; Cole & Paek, 2023; Çakıcı Eser & Gelbal, 2015; DeMars, 2003; de la Torre & Patz, 2005; Gül, 2015; Kuo, 2015; Lee, 2012; Reise & Yu, 1990; Şahin, 2012; Sheng & Wikle, 2007; Sünbül, 2011; Zhang, 2012). The outcomes of this research align with the existing literature and underlined the congruence of findings in this regard.

In addition, Bulut (2013) stated that the sample size of 1500 is sufficient for the parameter validation studies of the MIRT. The RMSE values above 0.10 are considered an indicator of poor fit (Hu & Bentler, 1999). In this study, RMSE values with poor fit were encountered mostly in the conditions of 1500 samples and a low number of items.

Increasing the number of items according to the three algorithms used (EM, MHRM, and QMCEM) led to a decrease in the RMSE values of the item parameter estimations. When similar studies in the literature were examined, it was seen that similar results were reached (Bolt & Lall, 2003; Çakıcı Eser & Gelbal, 2015; Yavuz & Hambleton, 2017). Çakıcı Eser and Gelbal (2015) also suggested using measurement tools with at least 12 items for two-dimensional simple structured data.

One of the simulation conditions in this research was to differentiate the correlation between dimensions. In this study, the differentiation of the correlation between dimensions did not cause a pattern in the RMSE values of the item parameter estimations. When similar studies in the literature were examined, it was seen that the variation of correlation values between dimensions had a different effect from condition to condition (e.g., Ansley & Forsyth, 1985; Bolt & Lall, 2003; Yavuz & Hambleton, 2017; Gül, 2015; Kuo, 2015). These results were consistent with the results of our research.

It was found that EM, MHRM, and QMCEM algorithms performed similarly in testing the accuracy of item parameter estimations in a two-factor structure. Cai (2010) found similar item parameter estimation accuracy when using the EM and MH-RM algorithms in two-dimensional polytomous datasets. Garnier-Villarreal, Merkle, and Magnus (2021) stated in their simulation studies that the EM algorithm works well in a low-factor structure like two, and the MHRM algorithm can be preferred if there are more than four factors. Considering the simulation times in this research, it can be suggested to use the QMCEM algorithm in a two-factor structure for time-saving and good estimations.

Finally, it was found that increasing the number of categories did not cause serious differences, except that several item parameters in different algorithms decreased the RMSE values. However, if researchers who will work with polytomous multidimensional data want to obtain better estimations for each item parameter, it is recommended to work with five-category data. In the literature, there is a need for studies that reveal the effect of the number of categories on the measurement precision under MIRT under different conditions. Future research may focus on this issue.

These studies had several limitations. First, this study focused on simple structured two-dimensional polytomous data. Real and simulation research can be performed on data with a different number of dimensions. Instead of simple structured MIRT models, complex structured MIRT models can be considered. The second limitation concerned simulation conditions. Similar studies can be conducted by considering different sample sizes, correlations between dimensions, and the number of categories. The third limitation was related to the algorithms and programming language used. Research can be carried out using different software and algorithms. The fourth limitation was related to the polytomous MIRT model used. Instead of the GRM, a study can be conducted in which different MIRT models (partial credit model, generalized partial credit model, etc.) are compared. In this study, RMSE values were focused; other studies may include values such as bias.

Peer-review: Externally peer-reviewed.

Author Contributions: Concept – S.B.; Design –S.B.; Supervision – H.Y.A.; Resources – S.B.; Materials – S.B.; Data Generation and/or Processing – S.B.; Analysis and/or Interpretation – S.B.; Literature Search – S.B.; Writing Manuscript –S.B.; Critical Review – H.Y.A.

Declaration of Interests: The authors have no conflict of interest to declare.

Funding: The authors declared that this study has received no financial support.

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*(4), 311–329. [CrossRef]

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement*, *22*(3), 37–51. [CrossRef]

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*(1), 37–48. [CrossRef]

Baker, F. B. (2001). *The basics of item response theory* (2nd ed). ERIC Claringhouse of assessment and evaluation.

Bolt, D. M., & Lall, V. F. (2003). Estimation of compensatory and noncompensatory multidimensional item response models using Markov chain Monte Carlo. *Applied Psychological Measurement*, *27*(6), 395–414. [CrossRef]

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.). *Testing structural equation models*. Sage.

Bulut, O. (2013). *Between-person and within-person subscore reliability: Comparison of unidimensional and multidimensional IRT models* (Unpublished Doctoral Dissertation), University of Minnesota Faculty of The Graduate School.

Bulut, O., & Sünbül, S. (2017). Monte Carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(3), 266–287. [CrossRef]

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

Çakıcı Eser, D., & Gelbal, S. (2015). Farklı boyutluluk özelliklerindeki basit ve karmaşık yapılı testlerin çok boyutlu madde tepki kuramına dayalı parametre kestirimlerinin incelenmesi. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *6*(2), 331–350. [CrossRef]

Chalmers, P. (2018). *Package mirt*. https://cran.r-project.org/web/packages/mirt/mirt.pdf.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. [CrossRef]

Choi, S.-W. (1996). *A response dichotomization technique for item parameter estimation of the multidimensional graded response model* (Unpublished Doctoral Dissertation). Faculty of the Graduate School of The University of Texas at Austin.

Cole, K., & Paek, I. (2023). SAS PROC IRT and the R mirt Package: A comparison of model parameter estimation for multidimensional IRT models. *Proceedings of the IRT and the R Mirt package: A Comparison of model parameter estimation for multidimensional IRT models*. *Psych*, *5*(2), 416–426. [CrossRef]

De Ayala, R. J. (1994). The influence of multidimensionality on the graded response model. *Applied Psychological Measurement*, *18*(2), 155–170. [CrossRef]

de la Torre, J. (2008). Multidimensional scoring of abilities: The ordered polytomous response case. *Applied Psychological Measurement*, *32*(5), 355–370. [CrossRef]

de la Torre, J. (2009). Improving the quality of ability estimates through multidimensional scoring and incorporation of ancillary variables. *Applied Psychological Measurement*, *33*(6), 465–485. [CrossRef]

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order IRT approach. *Applied Psychological Measurement*, *34*(4), 267–285. [CrossRef]

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. [CrossRef]

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

DeMars, C. E. (2003). Sample size and the recovery of nominal response model item parameters. *Applied Psychological Measurement*, *27*(4), 275–288. [CrossRef]

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, *39*(1), 1–22. [CrossRef]

Donoghue, J. R. (1993). An empirical examination of the IRT information in polytomously scored reading items under the generalized partial credit model. *Journal of Educational Measurement*, *31*(4), 295–311. [CrossRef]

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

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, *14*(3), 275–299. [CrossRef]

Fu, Z. H., Tao, J., & Shi, N. Z. (2010). Bayesian estimation of the multidimensional graded response model with nonignorable missing data. *Journal of Statistical Computation and Simulation*, *80*(11), 1237–1252. [CrossRef]

Garnier-Villarreal, M., Merkle, E. C., & Magnus, B. E. (2021). Between-item multidimensional IRT: How far can the estimation methods go? *Psych*, *3*(3), 404–421. [CrossRef]

Gül, E. (2015). *Tek boyutlu ve çok boyutlu madde tepki kuramına göre çok boyutlu yapıların incelenmesi* (Unpublished Doctoral Dissertation). Ankara University.

Guo, W., & Choi, Y. J. (2023). Assessing dimensionality of IRT models using traditional and revised parallel analyses. *Educational and Psychological Measurement*, *83*(3), 609–629. [CrossRef]

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2–3), 57–63. [CrossRef]

Harwell, M., Stone, C. A., Hsu, T.-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125. [CrossRef]

Houts, C. R., & Cai, L. (2016). *flexMIRT user's manual version 3.5: Flexible multilevel multidimensional item analysis and test scoring*. Vector Psychometric Group.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. [CrossRef]

Jank, W. (2005). Quasi-Monte Carlo sampling to improve the efficiency of Monte Carlo EM. *Computational Statistics and Data Analysis*, *48*(4), 685–701. [CrossRef]

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample size requirements for estimation of item parameters in the multidimensional graded response model. *Frontiers in Psychology*, *7*(109), 109. [CrossRef]

Kalkan, Ö. K. (2022). The comparison of estimation methods for the four-parameter logistic item response theory model. *Measurement: Interdisciplinary Research and Perspectives*, *20*(2), 73–90. [CrossRef]

Kim, S., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*(1), 25–41. [CrossRef]

Koğar, H. (2014). *Madde tepki kuramının farklı uygulamalarından elde edilen parametrelerin ve model uyumlarının örneklem büyüklüğü ve test uzunluğu açısından karşılaştırılması* (Unpublished Doctoral Dissertation). Hacettepe University

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scalling, and linking* (3rd ed). Springer.

Kose, I. A., & Demirtasli, N. C. (2012). Comparison of unidimensional and multidimensional models based on item response theory in terms of both variables of test length and sample size. *Procedia – Social and Behavioral Sciences*, *46*, 135–140. [CrossRef]

Kuo, T. C., & Sheng, Y. (2016). A comparison of estimation methods for a multi-unidimensional graded response IRT model. *Frontiers in Psychology*, *7*(880), 880. [CrossRef]

Kuo, T. C. (2015). *Bayesıan estımatıon of multı-unıdımensıonal graded response IRT models* (Unpublished Doctoral Dissertation). Southern Illinois University Carbondale.

Lee, J. (2012). *Multidimensional item response theory: An investigation of interaction effects between factors on item parameter recovery using Markov chain Monte Carlo* (Unpublished Doctoral Dissertation). Michigan State University.

Lee, S. H. (2007). *Multidimensional item response theory: A SAS MDIRT MACRO and emprical study of PIAT MATH Test* (Unpublished Doctoral Dissertation). The University of Oklahoma.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple choice, contructed response, and examinee selected items on two achievement tests. *Journal of Educational Measurement*, *31*(3), 234–250. [CrossRef]

Martelli, I. (2014). *Multidimensional item response theory models with general and specific latent traits for ordinal data* (Unpublished Doctoral Dissertation). Alma Mater Studiorum Università di Bologna.

Martelli, I., Matteucci, M., & Mignani, S. (2016). Bayesian estimation of a multidimensional additive graded response model for correlated traits. *Communications in Statistics – Simulation and Computation*, *45*(5), 1636–1654. [CrossRef]

Martin-Fernandez, M., & Revuelta, J. (2017). Bayesian estimation of multidimensional item response models. A comparison of analytic and simulation algorithms. *Psicologica: International Journal of Methodology and Experimental Psychology*, *38*(1), 25–55.

Mun, E. Y., Huo, Y., White, H. R., Suzuki, S., & De la Torre, J. (2019). Multivariate higher-order IRT model and MCMC algorithm for linking individual participant data from multiple studies. *Frontiers in Psychology*, *10*, 1328. [CrossRef]

Özer Özkan, Y. (2014). Öğrenci başarılarının belirlenmesi sınavından klasik test kuramı, tek ve çok boyutlu madde tepki kuramı modelleri ile kestirilen başarı puanlarının karşılaştırılması. *International Journal of Human Sciences / Uluslararası İnsan Bilimleri Dergisi*, *11*(1), 20–44. [CrossRef]

Reckase, M. D. (1997). The past and future of multidimensional item response theory. *Applied Psychological Measurement*, *21*(1), 25–36. [CrossRef]

Reise, S. P., & Yu, J. (1990). Parameter recovery in graded response model using MULTILOG. *Journal of Educational Measurement*, *27*(2), 133–144. [CrossRef]

Sahin, S. G., Gelbal, S., & Walker, C. M. (2019). Examining parameter estimation when treating semi-mixed multidimensional constructs as unidimensional. *Journal of Applied Measurement*, *20*(3), 310–325.

Sayın, A., & Gelbal, S. (2016). Yapısal eşitlik modellemesinde parametrelerin klasik test kuramı ve madde tepki kuramına göre sınırlandırılmasının uyum indekslerine etkisi. *Uluslararası Eğitim Bilim ve Teknoloji Dergisi*, *2*(2), 57–71.

Sengul Avsar, A., & Tavsancil, E. (2017). Examination of polytomous items' psychometric properties according to nonparametric item response theory models in different test conditions. *Educational Sciences: Theory and Practice*, *17*(2), 493–514. [CrossRef]

Sheng, Y. (2005). *Bayesian analysis of hierarchical IRT models: Comparing and combining the unidimensional & multi-unidimensional IRT models* (Unpublished Doctoral Dissertation). Faculty of the Graduate School University of Missouri-Columbia.

Sheng, Y. (2008). A MATLAB package for Markov Chain Monto Carlo with a multi-unidimensional IRT model. *Journal of Statistical Software*, *28*(10), 1–20. [CrossRef]

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

Sheng, Y., & Wikle, C. K. (2008). Bayesian multidimensional IRT models with a hierarchical structure. *Educational and Psychological Measurement*, *68*(3), 413–430. [CrossRef]

Sünbül, Ö. (2011). *Çeşitli boyutluluk özelliklerine sahip yapılarda madde parametrelerinin değişmezliğinin Klasik test teorisi, tek boyutlu madde tepki*

kuramı, çok boyutlu madde tepki kuramı çerçevesinde incelenmesi (Unpublished Doctoral Dissertation). Mersin University.

Şahin, A. (2012). Madde Tepki Kuramı'nda test uzunluğu ve örneklem büyüklüğünün model veri uyumu, madde parametreleri ve standart hata değerlerine etkisinin incelenmesi (Unpublished Doctoral Dissertation). Hacettepe University.

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. [CrossRef]

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

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. [CrossRef]

Yavuz, G., & Hambleton, R. K. (2017). Comparative analyses of MIRT models and software (BMIRT and flexMIRT). *Educational and Psychological Measurement*, *77*(2), 263–274. [CrossRef]

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