Original Article

Received/Accepted Dates

17.01.2025/08.02.2025

DOI 10.52096/usbd.9.41.10

International Journal of Social Sciences

Uluslararası Sosyal Bilimler Dergisi

www.sobider.net e-ISSN: 2548-0685

IJSS

# Examining The Effect of Sample Size and Test Length on Parameter Estimates in The Multidimensional Item Response Theory Model Using Different Algorithms

**Dr. Ömer DOĞAN**

Ministry of National Education, TÜRKİYE, 64omerdogan64@gmail.com ORCID: 0000-000-5169-520X

**Abstract**

Multidimensional Item Response Theory (MIRT) offers a robust framework for modeling complex latent traits, addressing the limitations of unidimensional models in educational and psychological assessments. This study aims to examine the estimation performance of three widely used algorithms—Expectation-Maximization (EM), Stochastic EM (SEM), and Metropolis-Hastings Robbins-Monro (MH-RM)—under varying test lengths (10, 20, and 40 items) and sample sizes (N = 1000 and 5000). Employing a Monte Carlo simulation design, datasets were generated in R to reflect multidimensional structures. Item and ability parameters were estimated, and estimation accuracy was evaluated using Root Mean Square Error (RMSE) and bias statistics. Findings indicate that under a sample size of 1000, MH-RM performs most effectively for discrimination parameters (a) in longer tests, while SEM is preferable for shorter tests. EM yielded the least biased estimates for difficulty (d) and guessing (c) parameters. At N = 5000, both EM and MH-RM demonstrated lower RMSE values, while SEM showed higher error rates. Overall, increasing sample size and test length led to improved estimation accuracy across all methods. Results corroborate previous findings and highlight the superior performance of MH-RM in high-dimensional estimation, particularly in large-scale testing scenarios. These outcomes provide practical guidance for researchers in selecting suitable estimation techniques under different test conditions.

**Keywords:** Multidimensional Item Response Theory, Expectation Maximization, Metropolis-Hastings Robbins-Monro, Stochastic Expectation Maximization.

## Introduction

Classical Test Theory (CTT) has been used by researchers for many years. However, due to its various limitations and difficult-to-meet assumptions, along with developments in and accessibility of computer technologies, Item Response Theory (IRT) has come to the forefront. IRT provides a methodological basis for model-based measurement. Item responses are modeled as a function of individual ability levels and item characteristics. Item difficulties and individual ability levels can be defined on a common scale (Embretson & Reise, 2000). Unidimensional IRT has also been used for a long time in the fields of education and psychology; however, it is very difficult to fully meet the unidimensionality assumption in achievement and ability tests (Hambleton et al., 1991; Reckase, 1997). This is because many educational and psychological tests are inherently multidimensional, meaning that they measure two or more dimensions or constructs.

Multidimensional Item Response Theory (MIRT) is used to model the relationship between two or more unobserved variables—conceptually defined as constructs or dimensions—and the probability that a test taker will answer a specific test item correctly. In contrast, unidimensional item response theory models the relationship between an unobserved construct or dimension and the probability that a test taker will answer a specific test item correctly (Ackerman et al., 2003). MIRT can be considered either as a special case of multivariate statistical analysis—particularly factor analysis or structural equation modeling—or as an extension of unidimensional IRT. These two conceptualizations of MIRT shape how the methodology is applied and how the results of MIRT analyses are interpreted (Reckase, 2009).

MIRT is a highly interesting methodology for assessing competencies in educational contexts. If competencies are considered as tendencies to perform successfully within a specific real-life context, then modeling these tendencies as multidimensional constructs may be necessary and desirable (Hartig & Höhler, 2009). MIRT is the result of the convergence of ideas from various fields such as psychology, education, test development, psychometrics, and statistics. Two general themes lie beneath the influence of these ideas on the development of MIRT. The first theme is that as our understanding of these fields increases, it becomes apparent that things are more complex than initially thought. The second theme is that complexity can be represented by models or theories; however, these models and theories are idealizations of reality. Since they are idealizations, they may likely be proven wrong when tested using a large number of observations. Nevertheless, models can provide useful approaches for many practical applications (Reckase, 2009).

In MIRT studies, normal ogive models are often used to describe the relationship between the probability of answering an item correctly and the position in multidimensional space. Today, statistical estimation programs also perform estimations based on ogive models. The three-parameter form of the normal ogive model extended for multidimensionality is given in Equation 1:

Equation 1. $$P\left(U_{ij}=1\middle|\theta_j,\alpha_{i,}\gamma_{i,}d_i\right)=\gamma_i+(1-\gamma_i)\frac{1}{\sqrt{2\pi}}\int_{-z_i(\theta_j)}^{\infty}e^{\frac{-t^2}{2}}dt$$

$Uij$= The response of individual $j$ to item $i$ (1 or 0)

$\theta j$= The ability level of individual $j$

$\alpha i$= The discrimination parameter of item $i$

$\gamma i$= The lower asymptote (guessing parameter) of item $i$

$di$= The intercept of item $i$

The expression $-zi(\theta j)$ in the equation corresponds to $\alpha i\theta j' + di$. Other terms are as previously defined. If the guessing parameter $\gamma$ in the equation is equal to zero, the model reduces to a two-parameter form. In addition, the $d$ parameter in the equation represents the intercept, which is obtained through the interaction of the $\alpha$ and β parameters (Reckase, 2009).

IRT models are capable of estimating at least two types of parameters: individual parameters and item parameters. For complex models such as the two-parameter, three-parameter, and multidimensional IRT models, consistent estimations depend largely on the characteristics of the items (Embretson & Reise, 2000). Accordingly, various algorithms are used for parameter estimation in IRT models. Among these, three algorithms used in the scope of this study are the most well-known: the Expectation-Maximization (EM) algorithm, the Stochastic EM (SEM) algorithm, and the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm.

The EM algorithm is widely used for parameter estimation in data-driven process modeling. EM is an algorithm used for maximum likelihood estimation of parameters and ensures the convergence of the likelihood function (Sammaknejad et al., 2019). The EM algorithm operates iteratively through two steps: the E (expectation) step and the M (maximization) step. In the E step, the expected value of the likelihood function arising from the distribution of observed and missing data is computed under the condition of the current estimates of the parameters. In the M step, a new set of parameters is estimated by maximizing the expected value of the likelihood function (Koch, 2013).

The EM algorithm is a fundamental tool for maximum likelihood estimation of parameters in latent variable models. However, with the extraordinary growth in large datasets in recent years, the EM algorithm has rapidly become impractical, as it requires processing the entire dataset at each E-step. For this reason, Cappé and Moulines (2009) proposed the Stochastic EM (SEM) algorithm for exponential family models. This algorithm reduces the time complexity of the E-step by approximating the full-batch expectation using an exponential moving average over data minibatches. SEM has been used in various applications, including natural language processing, topic modeling, and hidden Markov models (Chen et al., 2018). By integrating a stochastic sampling procedure into the estimation process, SEM eliminates the need to compute the pseudo-likelihood function and avoids the associated analytical maximization problems. Therefore, the sequence of parameter estimates generated by SEM constitutes a discrete-time stochastic process that forms an ergodic and homogeneous Markov chain, which does not converge pointwise or almost surely, but instead converges to a single stationary distribution (Moser et al., 2005).

The MH-RM algorithm represents a synthesis of the Markov Chain Monte Carlo method, which is widely used in Bayesian statistics, and the Robbins-Monro stochastic approximation algorithm, which is well known in the optimization literature (Cai, 2008). The MH-RM algorithm is an extension of the Stochastic Approximation EM algorithm, combining the Metropolis–Hastings algorithm with the Robbins–Monro stochastic approximation algorithm (Yang & Cai, 2014). In a sense, the MH-RM algorithm was designed to address the "curse of dimensionality," which limits the applicability of the numerically intensive Bock–Aitkin EM in multidimensional IRT. This represents one of the first applications of the MH-RM algorithm to non-normal latent variable models and significantly extends the applicability of MH-RM as a general estimation approach for maximum marginal likelihood estimation in IRT modeling (Monroe & Cai, 2014).

One of the main problems regarding parameter estimation techniques in MIRT models is the continued uncertainty surrounding the relationships between the accuracy of item and individual parameter estimates and test conditions such as the number of items and sample size. In addition to previous studies indicating that larger sample sizes improve parameter estimates, it is also important to determine how this improvement is influenced by changes in the number of items in a test. Examining the estimation algorithms in the context of varying test lengths and sample sizes is expected to contribute to the literature and serve as a guide for researchers working in this field.

Accordingly, this study investigates how item and individual parameter estimations conducted under Multidimensional Item Response Theory (using EM, MH-RM, and SEM algorithms) are affected by sample size (1000–5000) and test length (10–20–40). Within this scope, when the sample size is 1000 and test lengths are 10, 20, and 40, the Root Mean Square Error (RMSE) and bias values related to item and individual parameter estimates are examined. Similarly, when the sample size is 5000 and test lengths are 10, 20, and 40, changes in the RMSE and bias values of item and individual parameter estimates are analyzed separately for each estimation algorithm.

## METHODS

### Research Type

This study is a Monte Carlo simulation in which datasets for comparing the selected estimation algorithms are generated using the R programming language (R Core Team, 2025). The variables that may influence algorithm performance are controlled through simulated data, allowing for a systematic evaluation of the strengths and limitations of each algorithm. In this regard, the study qualifies as basic research, as it aims to contribute to the development and refinement of methodological approaches.

### Data

Within the scope of the study, datasets were generated in a simulated, multidimensional format in a computer environment, with test lengths set at 10, 20, and 40 items, and sample sizes set at 1000 and 5000.

### Data Analysis

The datasets generated through simulation were analyzed to estimate item and individual parameters. These datasets were created and processed using R programming language scripts that reflected the structure of the test. After the estimation procedures were completed, goodness-of-recovery (GOR) statistics were calculated by comparing the estimated and true values of item and individual parameters. GOR statistics are computed to examine the effects of different conditions and to test the stability of the estimation. In this study, two GOR statistics were calculated for each parameter: Root Mean Square Error (RMSE) and bias. RMSE is defined as the square root of the average of the squared differences between the true and estimated parameter values.

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(\bar{\tau}_{nj} - \tau_{nj})^2}{N}}$$

In the formula, the term $N$ represents the number of parameters included in the test. When RMSE is calculated for item parameters, $N$ refers to the number of items in the test; when calculated for individual parameters, $N$ refers to the number of individuals. Here, $\mathrm{T}_{nj}$ denotes the true value of the $j$th parameter for the $n$th item, while , $\bar{\tau}_{nj}$ denotes the estimated value of the $j$th parameter for the $n$th item.

Bias is defined as the difference between the average estimated value of a parameter and its true value. In this study, bias values were calculated for each item and each individual parameter.

$$Bias \quad (\tau_j) = \left(\frac{\sum_{r=1}^{R}\bar{\tau}_{jr}}{R}\right) - \tau_j$$

In the formula, $R$ represents the number of replications. Here, $\bar{\tau}_{jr}$ denotes the estimated value of parameter $j$ in the $r$th replication, while $\tau_j$ is the true value of parameter $j$. According to the formula, the difference between the average of the estimated values for each parameter across all replications and the true parameter value was computed. In this way, the number of bias values obtained corresponds to the number of items for item parameters and the number of individuals for individual parameters (Maris, 1999).

## FINDINGS

As part of the study, bias and RMSE values for individual and item parameters estimated using the EM, SEM, and MH-RM algorithms were first reported for tests consisting of 10, 20, and 40 items under a sample size of 1000. Table 1 presents the values obtained for the 10-item test.

**Table 1.** *RMSE and bias values for algorithms when the sample size is 1000 and the test length is 10.*

| Sample Size | Number of Items | Algorithm | Bias a1 | Bias a2 | Bias d | Bias c | RMSE a1 | RMSE a2 | RMSE d | RMSE c |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000 | 10 | EM | -2.79 | -1.38 | -0.04 | -0.05 | 2.88 | 1.54 | 0.46 | 0.13 |
| | | MH-RM | -3.14 | -1.62 | 0.18 | -0.03 | 3.16 | 1.69 | 0.45 | 0.13 |
| | | SEM | -2.64 | -1.31 | -0.13 | -0.05 | 2.70 | 1.39 | 1.94 | 0.11 |

When examining the RMSE and bias values related to item and individual parameter estimations for the dataset with a sample size of 1000 and test length of 10 items (as presented in Table 1), it is observed that the lowest bias for the discrimination parameters (*a*) was obtained using the SEM algorithm, the lowest bias for the *d* parameter was obtained using the EM algorithm, and the lowest bias for the *c* parameter was achieved with the MH-RM algorithm. In terms of RMSE values, the smallest errors for the *a* and *c* parameters were produced by the SEM algorithm, while the lowest RMSE for the *d* parameter was obtained using the MH-RM algorithm. Table 2 presents the values obtained for the 20-item test under a sample size of 1000.

**Table 2.** *RMSE and bias values for algorithms when the sample size is 1000 and the test length is 20.*

| Sample Size | Number of Items | Algorithm | Bias a1 | Bias a2 | Bias d | Bias c | RMSE a1 | RMSE a2 | RMSE d | RMSE c |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000 | 20 | EM | -2.72 | -1.28 | 0.19 | -0.01 | 2.78 | 1.44 | 0.46 | 0.12 |
| | | MH-RM | -2.61 | -1.26 | 0.16 | -0.04 | 2.69 | 1.34 | 0.56 | 0.11 |
| | | SEM | -2.62 | -1.31 | -0.13 | -0.05 | 2.70 | 1.40 | 1.94 | 0.11 |

When examining the RMSE and bias values related to item and individual parameter estimations for the dataset with a sample size of 1000 and a test length of 20 items (as presented in Table 2), it is observed that the lowest bias for the discrimination parameters (*a*) was obtained using the MH-RM algorithm, the lowest bias for the *d* parameter was achieved with the SEM algorithm, and the lowest bias for the *c* parameter was produced by the EM algorithm. Regarding RMSE values, the MH-RM algorithm yielded the smallest error for the *a* parameters, the EM algorithm produced the lowest RMSE for the *d* parameter, and the SEM algorithm yielded the lowest RMSE for the *c* parameter. Table 3 presents the values obtained for the 40-item test under a sample size of 1000.

**Table 3.** *RMSE and bias values for algorithms when the sample size is 1000 and the test length is 40.*

| Sample Size | Number of Items | Algorithm | Bias a1 | Bias a2 | Bias d | Bias c | RMSE a1 | RMSE a2 | RMSE d | RMSE c |
|---|---|---|---|---|---|---|---|---|---|---|
| 1000 | 40 | EM | -2.71 | -1.27 | 0.04 | -0.03 | 2.87 | 1.75 | 0.47 | 0.11 |
| | | MH-RM | -2.60 | -1.25 | 0.14 | -0.04 | 2.69 | 1.36 | 1.60 | 0.12 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | SEM | -2.63 | -1.26 | -0.22 | -0.03 | 2.69 | 1.40 | 1.82 | 0.10 |

When examining the RMSE and bias values related to item and individual parameter estimations for the dataset with a sample size of 1000 and a test length of 40 items (as presented in Table 3), it is observed that the lowest bias for the discrimination parameters (*a*) was obtained using the MH-RM algorithm, while the lowest bias for the *d* and *c* parameters was achieved with the EM algorithm. Regarding RMSE values, the smallest errors for the *a* parameters were produced by the MH-RM algorithm, the lowest RMSE for the *d* parameter was obtained using the EM algorithm, and the SEM algorithm yielded the lowest RMSE for the *c* parameter. Table 4 presents the values obtained for the 10-item test under a sample size of 5000.

**Table 4.** *RMSE and bias values for algorithms when the sample size is 5000 and the test length is 10.*

| Sample Size | Number of Items | Algorithm | Bias a1 | Bias a2 | Bias d | Bias c | RMSE a1 | RMSE a2 | RMSE d | RMSE c |
|---|---|---|---|---|---|---|---|---|---|---|
| | | EM | -2.68 | -1.49 | -0.04 | 0.01 | 2.69 | 1.54 | 0.34 | 0.13 |
| 5000 | 10 | | | | | | | | | |
| | | MH-RM | -2,86 | -1.63 | -0.09 | 0.01 | 2.73 | 1.87 | 0.18 | 0.09 |
| | | SEM | -2.79 | -1.60 | 0.09 | -0.02 | 2.80 | 1.68 | 1.59 | 0.12 |

When examining the RMSE and bias values related to item and individual parameter estimations for the dataset with a sample size of 5000 and a test length of 10 items (as presented in Table 4), it is observed that the lowest bias for the discrimination parameters (*a*), as well as for the *d* and *c* parameters, was obtained using the EM algorithm. Regarding RMSE values, the smallest error for the *a* parameters was produced by the EM algorithm, while the lowest RMSE for the *d* and *c* parameters was obtained using the MH-RM algorithm. Table 5 presents the values obtained for the 20-item test under a sample size of 5000.

**Table 5.** *RMSE and bias values for algorithms when the sample size is 5000 and the test length is 20.*

| Sample Size | Number of Items | Algorithm | Bias a1 | Bias a2 | Bias d | Bias c | RMSE a1 | RMSE a2 | RMSE d | RMSE c |
|---|---|---|---|---|---|---|---|---|---|---|
| 5000 | 20 | EM | -2.79 | -1.57 | -0.05 | -0.01 | 2.88 | 1.65 | 0.21 | 0.06 |
| | | MH-RM | -2.74 | -1.52 | 0.01 | -0.01 | 2.80 | 1.60 | 0.20 | 0.08 |
| | | SEM | 2.76 | -1.46 | -0.09 | 0.08 | 0.45 | 0.90 | 1.53 | 0.08 |

When examining the RMSE and bias values related to item and individual parameter estimations for the dataset with a sample size of 5000 and a test length of 20 items (as presented in Table 5), it is observed that the lowest bias for the discrimination parameters (*a*) was obtained using the SEM algorithm, while the lowest bias for the *d* and *c* parameters was again achieved with the EM algorithm. In terms of RMSE values, the SEM algorithm yielded the smallest error for the *a* parameters, the MH-RM algorithm produced the lowest RMSE for the *d* parameter, and the EM algorithm provided the lowest RMSE for the *c* parameter. Table 6 presents the values obtained for the 40-item test under a sample size of 5000.

**Table 6.** *RMSE and bias values for algorithms when the sample size is 5000 and the test length is 40.*

| Sample Size | Number of Items | Algorithm | Bias a1 | Bias a2 | Bias d | Bias c | RMSE a1 | RMSE a2 | RMSE d | RMSE c |
|---|---|---|---|---|---|---|---|---|---|---|

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 5000 | 40 | EM | -2.65 | -1.52 | -0.06 | -0.01 | 2.70 | 1.58 | 0.19 | 0.08 |
| | | MH-RM | -2.76 | -1.54 | 0.05 | 0.001 | 2.82 | 1.58 | 0.37 | 0.09 |
| | | SEM | -2.71 | -1.59 | 0.06 | -0.01 | 2.84 | 1.64 | 1.96 | 0.10 |

When examining the RMSE and bias values related to item and individual parameter estimations for the dataset with a sample size of 5000 and a test length of 40 items (as presented in Table 6), it is observed that the lowest bias for the discrimination parameters (*a*) was obtained using the EM algorithm, while the lowest bias for the *d* and *c* parameters was achieved using the MH-RM algorithm. Regarding RMSE values, the EM algorithm yielded the smallest errors for both the *a* parameters and the *d* and *c* parameters.

## CONCLUSION AND DISCUSSION

This study aimed to examine how item and individual parameter estimation algorithms used within the framework of Multidimensional Item Response Theory (EM, MH-RM, SEM) are affected by sample size (1000–5000) and test length (10–20–40).

When the sample size was 1000, it was observed that increasing the number of items in the EM algorithm reduced the bias in the discrimination parameters (*a*) but increased the bias in the *d* parameters. The RMSE values for the *d* parameter did not show any notable change. No consistent pattern was found for the other parameters. In the MH-RM algorithm, increasing the number of items led to a decrease in bias for both the *a* and *d* parameters. While RMSE values for the *a* parameter decreased, RMSE values for the *d* parameter increased. In the SEM algorithm, no significant change was observed in either bias or RMSE for the *a* and *d* parameters. For the *c* parameters, no substantial variation related to item count was observed across any of the algorithms.

When the sample size was 5000, increasing the number of items did not lead to a significant change in the bias of the discrimination parameters (*a*) for the EM and MH-RM algorithms, whereas the SEM algorithm produced lower discrimination parameter estimates as item count increased. For the *d* and *c* parameters, no notable change in bias was observed across any of the algorithms as a function of item count. However, RMSE values obtained using the SEM algorithm were found to be considerably higher compared to those of the other algorithms.

In light of the findings, it can be stated that under a sample size of 1000, the MH-RM estimation method is the most effective for estimating the *a* parameter in tests with 20 and 40 items, whereas the SEM method is more effective for tests with 10 items. Moreover, the lowest bias for the *d* and *c* parameters was observed with the EM algorithm. While all three estimation methods yielded effective results for RMSE values depending on the parameter, the MH-RM algorithm generally produced estimates with lower error compared to the other two.

Cai (2010) conducted a series of simulation studies for multidimensional item response models and demonstrated that the parameter recovery performance of the MH-RM method was nearly equivalent to that of the numerical quadrature-based EM algorithm. Liu (2022), in a study comparing efficient MH-RM (eMH-RM), EM, and Metropolis-Hastings (MH) algorithms in terms of parameter recovery and computational time, found that eMH-RM was considerably more efficient computationally and tended to produce better estimates than the EM algorithm.

Under the condition of a sample size of 5000, the EM estimation method was found to produce more accurate estimates in terms of bias, while both the EM and MH-RM methods yielded lower errors in terms of RMSE. Büyükkıdık (2018) conducted a study comparing the bias and RMSE values of item and ability parameters estimated using the multidimensional graded response model under various conditions, including sample sizes (1500 and 3000), number of items (12, 24, 36), inter-dimensional correlations (0.20, 0.50, and 0.80), number of response categories (three and five), and number of dimensions (two and three). In that study, the performance of different algorithms (EM, MHRM, QMCEM) was examined. The findings indicated that for all three algorithms, RMSE values for item and ability parameters generally decreased as sample size and number of items increased. However, for three-dimensional data, the highest RMSE values for item parameters were generally observed with the EM algorithm.

In a study by Yavuz (2014), the accuracy of item parameter recovery was investigated using three estimation techniques (BA-EM, MH-RM, and MCMC) under varying test conditions: number of dimensions (3 and 5), test lengths (15, 30, and 60), sample sizes (1000, 2000, and 5000), and inter-dimensional correlation values (0.0, 0.3, 0.6, and 0.9). In larger samples and longer tests, RMSE and bias values obtained from all techniques were found to be quite similar.

Arslan (2019) aimed to examine the dimensional structure of actual test data based on parameter estimation techniques and model fit indices. The study focused on mixed-format tests across unidimensional IRT models, two- and three-dimensional MIRT models, and bifactor models, taking into account item types, content domains, and cognitive processes. In this study, the BA-EM and MH-RM algorithms were used for item parameter estimation, while the Empirical Cross Products and Supplemented EM techniques were used for standard error estimation. The test data analyzed with the Generalized Partial Credit Model (GPCM), MH-RM, and Supplemented EM techniques yielded more accurate results. However, the data did not fit well with the BA-EM algorithm and the Graded Response Model (GRM). The study concluded that among all models, MH-RM was more appropriate for item parameter estimation, Supplemented EM for standard error estimation, and GPCM for modeling polytomous item structures.

In studies conducted using item parameter estimation techniques within the MIRT framework, it has been found that increasing the sample size leads to more accurate and less biased estimates of item parameters. This conclusion is also consistent with the general findings of similar studies (e.g., DeMars, 2002; Lee, 2007; Wang & Chen, 2005). A similar result was also reported by Kose and Demirtaslı (2012). While in unidimensional IRT models, increasing sample size and test length does not significantly affect item parameter estimates, analyses of model fit indices in MIRT models indicate that both sample size and test length have a clear impact on model fit. Consistent with the present study, their findings also suggest that working with longer tests and larger sample sizes under MIRT improves estimation accuracy and yields more precise measurement outcomes.

In light of these results, it can be concluded that researchers may benefit from adjusting their choice of estimation algorithm based on the number of items and the sample size. However, the effect of item count changes was not found to be particularly strong in this study, which may be due to the relatively close range of item numbers used across the different test conditions.

**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), 37–53. https://doi.org/10.1111/j.1745-3992.2003.tb00136.x

Arslan, A.K. (2019). *Karma yapıdaki geniş ölçekli naep fen bilimleri testinin boyutluluk yapısının çok boyutlu madde tepki kuramına göre incelenmesi.* [Unpublished Doctoral Dissertation, Ankara University].

Büyükkıdık, S. (2018). *Çok Kategorili Maddelerde Çok Boyutlu Madde Tepki Kuramına Dayalı Parametre Kestirimlerinin Karşılaştırılması.* [Unpublished Doctoral Dissertation, Gazi University].

Cai, L. (2008). *A metropolis-hastings robbins-monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model*. [Unpublished Doctoral Dissertation]. University of North Carolina.

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

Chen, J., Zhu, J., Teh, Y. W., & Zhang, T. (2018). Stochastic expectation maximization with variance reduction. *Advances in Neural Information Processing Systems*, *31*.

DeMars, C. (2002). Incomplete data and item parameter estimates under JMLE and MML. *Applied Measurement in Education, 15*, 1531.

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

Hambleton, R. K., Swaminathan, H. & Rogers, H. (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*, 57-63. https://doi.org/10.1016/J.STUEDUC.2009.10.002.

Koch, K.R. (2013). Robust estimation by expectation maximization algorithm. *J Geod 87*, 107–116 https://doi.org/10.1007/s00190-012-0582-3

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, https://doi.org/10.1016/j.sbspro.2012.05.082.

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].

Liu, C.-W. (2022). Efficient metropolis-hastings robbins-monro algorithm for high-dimensional diagnostic classification models. *Applied Psychological Measurement*, *46*(8), 662-674. https://doi.org/10.1177/01466216221123981

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

Monroe, S., & Cai, L. (2014). Estimation of a ramsay-curve item response theory model by the metropolis–hastings robbins–monro algorithm. *Educational and Psychological Measurement*, *74*(2), 343-369. https://doi.org/10.1177/0013164413499344

Moser, G., Zerubia, J., & Serpico, S. B. (2005). Dictionary-based stochastic expectation-maximization for SAR amplitude probability density function estimation. *IEEE Transactions on Geoscience and Remote Sensing*, *44*(1), 188-200.

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 2024.04.1) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Reckase, M.D. (1997). *Models for multidimensional tests and hierarchically structured training materials*. Technical Report. The American College Testing Program. Iowa City, Iowa

Reckase, M.D. (2009) *Multidimensional Item Response Theory*. Springer, New York. https://doi.org/10.1007/978-0-387-89976-3

Sammaknejad, N., Zhao, Y. & Huang, B. (2019). A review of the expectation maximization algorithm in data-driven process identification, *Journal of Process Control, 73*, 123-136, https://doi.org/10.1016/j.jprocont.2018.12.010.

Wang, W. C. & Chen, C. T. (2005). Item parameter recovery, standard error estimates, and fit statistics of the Winsteps program for the family of Rasch models. *Educational and Psychological Measurement*, *65,* 376-404.

Yavuz, G. (2014). *Çok Boyutlu Madde Tepki Kuramı Modelleri ve Paket Programları için Karşılaştırmalı Analizler.* [Unpublished Doctoral Dissertation, Hacettepe University].