

## An Investigation of Item Response Theory Parameter Estimations and Reliability in Multiple Groups

### Çoklu Gruplarda Madde Tepki Kuramı Parametre Kestirimi ve Güvenirliğinin İncelenmesi\*

Serap BÜYÜKKIDIK<sup>1</sup>, Hatice İNAL<sup>2</sup>

<sup>1</sup> Sinop University, Faculty of Education, Measurement and Evaluation in Education. e-mail: sbuyukkidik@gmail.com

<sup>2</sup> Mehmet Akif Ersoy University, Faculty of Education, Measurement and Evaluation in Education. e-mail: e-posta:hinal@mehmetakif.edu.tr

*Makale Türü/Article Types: Araştırma Makalesi/ Research Article*

*Makalenin Geliş Tarihi: 11.11.2022*

*Yayına Kabul Tarihi: 29.05.2023*

#### ABSTRACT

This study aimed to investigate the parameters estimation of item response theory (IRT) and their reliability in the context of binary data across multiple groups derived from the same population. Within the scope of the research, 2017 (April) mathematics subtest of the Transition from Primary to Secondary Education exam (TPSEE) was used. The dataset encompassed 7500 students as a single-sample subgroup and 3750 students distributed across two subgroups. In the research, IRT assumptions were examined first. After meeting the assumptions, item and ability estimations were performed with 1PLM, 2PLM, 3PLM, and 4PLM for dichotomous data. When the model data fits were examined, it was found that the best fit was obtained with 3PLM in all conditions. It was observed that the item parameters did not differ significantly as the sample changed. The *a* and *b* parameters differ according to the different IRT models. While there is a partial difference between the ability parameters as the samples change, there are also differences as the models change. Minor differences have been observed among the ability parameters obtained through ability estimation methods (Expected A Posteriori (EAP) and Maximum A Posteriori (MAP)). The marginal reliability coefficients were similar in all conditions. It is recommended that researchers perform parameter estimation with which have the best model data fit out of 3PLM or 4PLM to provide more information while performing analysis in IRT.

**Keywords:** IRT, Transition from Primary to Secondary Education, Multi-group, parameter estimation

---

\***Reference:** Büyükkıdık, S., & İnal, H. (2023). An investigation of item response theory parameter estimations and reliability in multiple groups. *Gazi University Journal of Gazi Education Faculty*, 43(2), 825-855.

**ÖZ**

*Bu çalışmada, aynı evrendeki çoklu gruplardan elde edilen ikili verilerde madde tepki kuramı (MTK) parametre kestirimi ve güvenilirliğinin incelenmesi amaçlanmıştır. Araştırma kapsamında TEOG 2017 (Nisan) matematik alt testi kullanılmıştır. Araştırma 7500 kişilik bir alt grupta ve 3750 kişilik iki alt grupta yer alan öğrencilerin verileri ile gerçekleştirilmiştir. Araştırmada öncelikle MTK varsayımları incelenmiştir. Varsayımlar sağlandıktan sonra, ikili puanlanan veriler için 1PLM, 2PLM, 3PLM ve 4PLM ile madde ve yetenek kestirimleri gerçekleştirilmiştir. Model veri uyumları incelendiğinde her koşulda en iyi uyumun 3PLM ile elde edildiği görülmüştür. Örneklem değişikçe madde parametrelerinin önemli ölçüde farklılaşmadığı gözlemlenmiştir. a ve b parametrelerinin farklı MTK modellerine göre farklılık gösterdiği bulgusuna ulaşılmıştır. Yetenek parametreleri arasında örneklemler değişikçe kısmi farklılık bulunurken, kullanılan modeller değişikçe de farklılık olduğu bulunmuştur. Yetenek kestirim yöntemlerine (Beklenen A Posteriori (EAP) ve Maksimum A Posteriori (MAP)) göre elde edilen yetenek parametreleri arasında bazı küçük farklılıkların olduğu görülmüştür. Marjinal güvenilirlik katsayıları tüm koşullarda benzerlik göstermiştir. Bu çalışmadan yola çıkarak, MTK'de analiz yaparken daha fazla bilgi sağlamak için araştırmacıların 3PLM veya 4PLM'den en iyi model veri uyumuna sahip olan modelle parametre kestirimi yapmaları önerilir.*

**Anahtar Sözcükler:** MTK, Temel eğitimden ortaöğretime geçiş, çoklu-grup, parametre kestirimi

**INTRODUCTION**

Measuring success in mathematics education can handle with various measurement tools. The measurement tools can be used in different ways depending on their purpose. Whichever measurement tool is used, three features should not be overlooked in measurement. The three qualities a measurement were validity, reliability, and usability, as will often be seen in the literature. There are various theories and models based on these theories in determining the psychometric properties of measurement.

With the No Child Left Behind (NCLB) Law adopted in the USA in 2001, schools account for all students to produce positive outcomes (U.S. Department of Education, 2001). To improve student performance and meet growing expectations, schools' situations are determined by standardized, high-stakes, and often multiple-choice assessments (Lembke & Stecker, 2007). Multiple-choice tests, which are objective and economical in terms of scoring, are frequently encountered in national and international practices. Even in assessments such as Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS), multiple-

choice tests are used. These large scale international assessments are implemented through international organizations and Turkey is one of country which participating these assessments. International data obtained from these assessments provide results for accountability. In Turkey, accountability is also a subject in the light of the national data. The transition from Primary to Secondary School Education exam (TPSEE) data is one of these national data. TPSEE was one of the exams that determines Turkish students' success nationally. TPSEE starting from the 2013-2014 academic year held by Ministry of Education in Turkey, is one of the periodic exams held in the 8th grade for six basic courses. Success of these six basic courses was measured by this exam. Collecting validity and reliability evidence and performing data analysis by using different models based on different theories is an important issue.

Various theories are used to reveal the psychometric properties of measurement. Two of these are the Classical Test Theory (CTT) dating back to the beginning of the 20th century and the Item Response Theory (IRT), which claims that the CTT has removed the limitations of the items depending on the groups, and which continues to develop since the mid-20th century. Classical Test Theory was used in contrast to its modern concept in IRT or modern test theory (Wu et al., 2016). IRT, commonly known as latent trait theory, strong true score theory, and modern mental test theory, can be identified by various terminologies (Kaplan & Saccuzzo, 1997). There are different models in IRT. So results getting from IRT can be differentiated from one model to another. This situation should be tested using different samples and models.

The number of response categories for items holds significant importance in the determination of parametric unidimensional IRT models (Edelen, & Reeve, 2007). Multiple choice tests are binary scored and there are several models developed for these tests scored 1-0, considering the number of parameters in the item response theory. Logistics models (PLM) with 1, 2, and 3 parameters are the most frequently used, and it is also possible to make estimates based on 4PLM, which produces the upper asymptote parameter (Edelen, & Reeve, 2007). 4PLM was created by Barton and Lord (1981) with the addition of the di parameter to 3PLM. With 4PLM, high-ability respondents take into

account the possibility of making a mistake in answering an easy item. With the addition of the upper asymptote with a value less than 1.00, it ensures that a respondent with a high ability level does not change significantly in the ability scale if it responds incorrectly to an easy item.

The parametric models discussed within the research scope of IRT were explained below.

### **One Parameter Logistics Model (1PLM)**

Danish mathematician George Rasch introduced a different approach in IRT in the 1950s. The logistic function obtained from the item characteristic curve used the normal ogive function (Han & Hambleton, 2014, p. 12). One parameter logistics model is one of the most widely used models in IRT. For one parameter logistic model, the item characteristic curve is given as in the equation (Hambleton et al., 1991, p. 12).

$$P_j(\theta_i) = \frac{e^{(\theta_i - b_j)}}{1 + e^{(\theta_i - b_j)}} \quad i=1,2,\dots,n \quad (1)$$

In the first equation:

The probability that a respondent who is selected randomly at the level of  $P_j(\theta)$ ,  $\theta$  will respond correctly to item  $i$ ,

$b_j$  = the difficulty parameter of item  $i$ ,

$n$  = number of items in the test

$e$  = is a constant number with a value of 2,718.

In this model, the discrimination parameter ( $a$ ) of all items is the same and the pseudo-guessing parameter ( $c$ ) is considered zero. However, the difficulty parameters of the items in test ( $b$ ) vary according to the item (Hambleton et al., 1991, p. 13). In the Rasch model,  $a$  parameter value is taken as 1.00, and in one parameter logistic model, an estimated value of  $a$ , i.e. an average value is used (Baker, 2001, p. 25; Embretson & Reise, 2000, p. 69; Kolen & Brennan, 2014, p. 175).

In the case of one parameter logistic model, when the probability of an item being answered correctly is 0.5, the value corresponding to the  $\theta$  ability level is the item difficulty index:  $b$  parameter. As the  $b$  parameter value of the item increases, the level of ability that individuals must have in order to respond correctly to that item increases. When the value of the parameter  $b$  is taken so that the group's ability average is zero and the standard deviation is one, parameter  $b_i$  usually gets values between -2.00 and +2.00; Items with a value close to -2.00 are very easy, items close to + 2.00 are very difficult (DeMars, 2010, p. 21; Hambleton et al., 1991, p. 13; Hambleton and Swaminathan, 1985, p. 36). The values that the difficulty parameter can be in the range of  $(-\infty, +\infty)$ , while in practice it is generally in the range of -3 to +3 (Baker, 2001, p. 22).

### Two Parameters Logistics Model (2PLM)

Lord (1952) developed the two-parameter item response model based on the cumulative normal distribution (normal ogive) for the first time. Birnbaum (1968) has replaced the two-parameter normal ogive function as a form of item characteristic function (Hambleton et al., 1991, p. 13).

The two-parameter logistics model is the generalized version of 1PLM. Instead of a fixed discrimination parameter in all items in the 1PLM model, each item has its discrimination parameter in 2PLM. Therefore, the model is explained mathematically as follows (Han & Hambleton, 2014, p. 12).

$$P_j(\theta_i) = \frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}} \quad i=1,2,\dots,n \quad (2)$$

$P_j(\theta) = \theta$  individual at the skill level the possibility of answering the item correctly ith item,

$b_j$  = difficulty parameter of item  $i$ ,

$a_j$  = discrimination parameter of item  $i$ ,

$n$  the number of items in the test,

$D = 1.7$  is the scaling factor.

The slope or discrimination parameter ( $a$ ) is theoretically in the range  $(-\infty, +\infty)$ , but in applications takes values in the range of 2.80 to +2.80 (Baker, 2001, p. 22). According to Hambleton et al. (1991, p. 15)  $a$  (discrimination) parameter usually gets a value between 0 and 2.00, and When a parameter value gets close to 2.00, the discrimination increases. So, higher values of  $a$  parameter indicate higher discrimination in IRT like CTT (DeMars, 2010, p. 5).

### Three Parameters Logistics Model (3PLM)

By adding the pseudo-guessing parameter to 2PLM by Birnbaum (1968), a third parameter was added to the model, and 3PLM was created (Baker, 2001, p. 28). The three-parameter logistic model allows the lower asymptote of the item characteristic curve to be different from zero. This model is suitable, even if the tested ones are at a fairly low proficiency level, for example, when they answer a multiple choice item with chance (Han & Hambleton, 2014, p. 13). In this model

$$P_j(\theta_i) = c_j + (-c_j) \frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}} \quad i=1,2,\dots,n \quad (3)$$

$P_j(\theta)$ ,  $b_j$ ,  $a_j$ ,  $n$  and  $D$  are explained in two-parameter model. The added parameter of the  $c_i$  (pseudo-chance-level) in the model represents the probability of responders with a low ability level to correctly answer the item and provides item characteristic curves with a low asymptote different from zero (Hambleton et al., 1991, p. 17; Hambleton & Swaminathan, 1985, pp. 37-38). 2PLM is the special version of 3PLM when  $c = 0$ , and the Rasch model is the special version of 2PLM when  $a = 1$  (Baker, 2010, p. 25; Han & Hambleton, 2014, p. 13).

The pseudo guessing parameter ( $c_j$ ) is theoretically gets values in the range  $[0, 1.0]$ , but in practice, it is stated that " $c$ " values higher than 0.35, where this range is out of  $[0, 0.35]$ ,

are not accepted (Baker, 2001, pp. 28-29). While the lower asymptote or  $c$ -parameter takes values between 0 and 1 in theory, it usually takes values between 0 and 0.3 in real data (DeMars, 2010, p. 21).

#### Four Parameters Logistics Model (4PLM)

Barton and Lord (1981) developed 4PLM by adding the probability of high-level respondents making mistakes when answering the easy item, namely the  $d_i$  parameter corresponding to the upper asymptote to 3PLM. The model is explained mathematically with the following equation:

$$P_j(\theta_i) = c_j + (d_j - c_j) \cdot \frac{e^{Da_j(\theta_i - b_j)}}{1 + e^{Da_j(\theta_i - b_j)}} \quad i=1,2,\dots,n \quad (4)$$

In this equation, the upper asymptote, represented by  $d_j$ , is slipping parameter. The value of this parameter is in the range [0, 1.0] in theory. The fact that the  $d$  parameter is considerably lower than 1.00 indicates that respondents with high ability levels are more likely to answer this item incorrectly due to carelessness and similar reasons.

#### Current Investigations Related to IRT Parameter Estimation

There are a lot of studies using IRT in different data sets (e.g., Erdemir & Önen, 2019; Doğruöz & Akın Arıkan, 2020; Kalkan, 2022; Yalçın, 2018). Some of them are simulation studies based on different conditions (e.g., Kalkan, 2022), and some of them include model comparison based on real data with only one sample (e.g., Erdemir & Önen, 2019; Doğruöz & Akın Arıkan, 2020; Yalçın, 2018).

Yalçın (2018) aimed to compare model fit for Rasch, 2PLM, 3PLM, 4PLM and Mixed IRT. It was found that the MixIRT model with two parameters and three latent classes has best model data fit values. Erdemir and Önen (2019) conducted a study to compare item and ability parameter estimation for 1PLM, 2PLM, 3PLM, and 4PLM. It was found that 4PLM was the better fitting model than 1PLM, 2PLM and 3PLM as a result of this study. Doğruöz and Akın Arıkan (2020) compare ability estimation for 3PLM,

and 4PLM. The result of this study indicated that WLE estimation method model was found best a algorithm for the 4PLM IRT ability parameters.

Kalkan (2022) aimed to examine the performance of expectation-maximization (EM), Quasi-Monte Carlo EM (QMCEM), and Metropolis-Hastings Robbins-Monro (MH-RM) estimation methods for the item parameters in the 4PLM IRT model under the manipulated conditions, including test length, the number of factors and the correlation between factors. The result of this study indicated that none of the methods were found best algorithm among the estimation methods for the estimation of 4PL item parameters based on all conditions.

Considering all the studies which includes 4PLM, no study was found that examined the differentiation of parameters and reliability of all models into multiple groups, models, and methods by using 4PLM. In this respect, it can be said that this research will contribute to the literature.

Thus, the aim of this research can be explained as follows:

In this study, for the math subtest of the national transition examination (TPSEE) which is conducted for transition from primary to secondary school education model data fit of 1PLM, 2PLM, 3PLM, and 4PLM models were compared, and item and ability parameters related to the best fit model were estimated. In addition, the marginal reliability coefficient was calculated within all four models in multiple groups.

In line with the purpose of the research, the research problems are as follows:

1. Considering the data (TPSEE 2017 April) set as completely and randomly assigned two groups (in multiple groups), which one of the 1PLM, 2PLM, 3PLM, and 4PLM provides the best model-data fit?
2. What are the item and ability parameters in multiple groups according to 1PLM, 2PLM, 3PLM, and 4PLM?
3. Do the predicted item parameters differ in multiple groups according to 1PLM, 2PLM, 3PLM, and 4PLM?



4. Do the ability parameters estimated in subgroups according to 1PLM, 2PLM, 3PLM, and 4PLM differ according to different parameter estimation methods?
5. How are the marginal reliability coefficients obtained according to 1PLM, 2PLM, 3PLM, and 4PLM in multiple groups?

## **METHOD**

### **Research Method**

As previously stated, the primary objective of this investigation is to compare the estimations derived from 1PLM, 2PLM, 3PLM, and 4PLM models across various clusters of a national transition examination. With this aim in mind, the research adopted a descriptive research design to elucidate the prevailing circumstances.

### **Sample**

For accurate parameter estimates, it was recommended different sample sizes based on IRT model (Kean & Reilly, 2014). In this study, the sample size was large enough considering the adequacy of the sample size for the convergence of parameters to a solution. The sample of the research consists of 7500 randomly selected 8th-grade students who took the TPSEE 2017 April and took booklet A. To be able to analyze between subgroups, the full sample was randomly divided into two sub-group as 3750 students.

### **Data Collection Tool**

For the data collection tool, mathematical subtest of TPSEE 2017 April was used in the research. TPSEE 2017 April is national exam for 8th grade students from Turkey. TPSEE includes different subtests. The mathematics subtest consists of 20 questions.

### **Data Analysis**

Before analyzing data, data were randomly divided into two groups with the "picked" command in the R (R Core Team, 2021) software. In the analysis of the data, first of all,

the IRT assumptions were tested. Parallel analysis was performed for unidimensionality. The scree plots obtained as a result of the parallel analysis of the data collection tool were given in Figure 1.

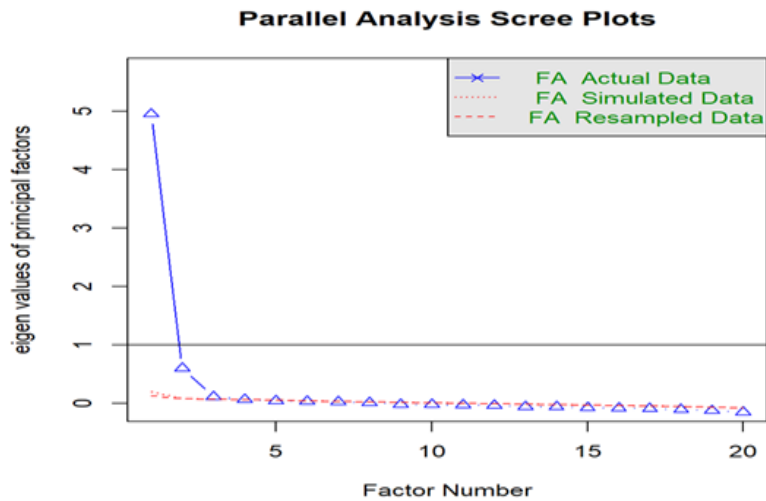


Figure 1. Parallel Analysis Scree Plots

When Figure 1 examined, it is seen that math subtest meet the unidimensionality assumption.

For local independence assumption, Yen's  $Q_3$  test was conducted. For Yen's  $Q_3$  test, residual correlations were found below the critical value of 0.20 ( $Q_{3\min}=-0.113$ ,  $Q_{3\max}=0.089$ ). This demonstrates that local independence assumption was met. Then, data analysis was started. Analysis of the data based on IRT was carried out with the R (R Core Team, 2021) software in the Supplementary Item Response Theory Models (sirt) (Robitzsch, 2021) package program. First of all, model data fit was tested for each sample. In the second stage, item and ability parameters were calculated for each sample. Ability parameters were handled using EAP and MAP estimation methods. In the third stage, whether the item parameters differ from sample to sample and model to model was examined with the Multi way ANOVA. From Multi way ANOVA results, effect sizes were interpreted based on Cohen (1988). Eta squared was interpreted as "negligible",

“small” “medium” and “large” respectively if Eta squared was “<0,01”; “0,01-0.06”; “0.06-0.14” and “>0.14”. In the fourth stage, it was tested whether the ability estimations differed significantly in multiple groups. In the last stage, the marginal reliability coefficient obtained for the measurements in each group was reported. Finally, variance analysis of IRT parameters obtained was performed according to the factors discussed in the study.

## RESULTS

In this section, the findings related to each sub-problem were given in order.

### Model Data Fit Findings

In Table 1, model fit indices obtained by analyzing the data in three groups with 1PLM, 2PLM, 3PLM, and 4PLM were given.

**Table 1.** Model-Data Fit Comparison For All Conditions

IRT Model	Sample	np	Deviance	Fit index			
				<i>AIC</i>	<i>BIC</i>	<i>CAIC</i>	<i>AICC</i>
1PLM	7500	21	163578.59	163620.59	163765.97	163786.97	163620.71
	3750-X	21	81961.80	82003.80	82134.62	82155.62	82004.05
	3750-Y	21	81600.21	81642.21	81773.03	81794.03	81642.46
2PLM	7500	40	161883.54	161963.54	162240.45	162280.45	161963.98
	3750-X	40	81179.46	81259.46	81508.64	81548.64	81260.34
	3750-Y	40	80661.98	80741.98	80991.16	81031.16	80742.86
3PLM	7500	60	158691.44	158811.44	159226.80	159286.80	158812.43
	3750-X	60	79629.95	79749.95	80123.72	80183.72	79751.93
	3750-Y	60	79004.96	79124.96	79498.73	79558.73	79126.94
4PLM	7500	80	158738.89	158898.89	159452.70	159532.70	158900.64
	3750-X	80	79658.39	79818.39	80316.75	80396.75	79821.92
	3750-Y	80	79027.44	79187.44	79685.80	79765.80	79190.97

When the deviance, AIC, BIC, CAIC, AICc indices in Table 1 were examined, it was seen that the best fit was in 3PLM. This is also true for the 7500, 3750-X, and 3750-Y samples. Considering the multi-groups from the same population, it was found that the fit indices obtained from the y sample of 3750 students showed a better fit than the indexes obtained from the X sample of 3750 students.

**Findings for Item and Ability Parameter Estimations**

Within the framework of the second sub-problem, the item parameters obtained for the sample of 7500 students were given in Table 2.

**Table 2.** The Item Parameters Obtained for the Sample of 7500 Students

Items	1PLM		2PLM		3PLM			4PLM			
	b	a	b	a	b	c	a	b	c	d	
item 1	-1.56	1.61	-1.03	1.95	-0.53	0.29	1.88	-0.61	0.25	1	
item 2	-1.52	2.58	-0.85	4.37	-0.37	0.31	3.77	-0.47	0.25	1	
item 3	0.65	1.49	0.43	4.24	0.75	0.16	6.64	0.66	0.17	0.95	
item 4	-0.94	1.77	-0.61	2.77	-0.02	0.29	2.56	-0.11	0.25	1	
item 5	-0.02	1.36	-0.02	4.76	0.63	0.29	4.05	0.56	0.25	1	
item 6	-1.03	1.9	-0.65	2.58	-0.19	0.25	2.6	-0.2	0.25	1	
item 7	0.24	0.85	0.28	2.03	0.95	0.27	2.06	0.8	0.25	0.95	
item 8	-3.1	1.95	-1.79	1.87	-1.94	0	2.11	-1.87	0	1	
item 9	-0.65	2.07	-0.42	4.66	0.16	0.29	4.29	0.09	0.25	1	
item 10	0.25	2.17	0.09	4.75	0.45	0.15	4.76	0.42	0.15	1	
item 11	0.35	1.38	0.25	3.72	0.7	0.21	5.43	0.61	0.22	0.95	
item 12	0.14	1.49	0.08	4.12	0.6	0.23	4.33	0.58	0.23	1	
item 13	0.45	1.23	0.35	3.01	0.8	0.21	3	0.78	0.21	1	
item 14	0.1	0.94	0.13	2.92	0.88	0.31	2.18	0.75	0.25	1	
item 15	-0.64	1.39	-0.46	2.37	0.22	0.3	2.12	0.11	0.25	1	
item 16	-0.49	1.92	-0.34	4.22	0.24	0.27	3.94	0.19	0.25	1	

item 17	-1.5	1.61	-0.99	1.54	-0.99	0	1.55	-0.99	0	1
item 18	-1.45	2.09	-0.87	2.49	-0.52	0.22	2.51	-0.53	0.22	1
item 19	0.06	1.32	0.05	2.7	0.57	0.23	2.7	0.55	0.23	1
item 20	0.51	0.88	0.53	1.77	1.01	0.21	1.94	0.91	0.21	0.95
Mean	-0.51	1.61	-0.29	3.14	0.17	0.22	3.22	0.11	0.21	0.99

When Table 2 was examined, it was seen that the highest  $b$  parameter was 0.65 and the lowest  $b$  parameter was -3.10 in 1PLM for a sample of 7500 students. For 2PLM, the  $b$  parameter had the highest value of 0.53 and the lowest value of -1.79. In 3PLM, the  $b$  parameter took values between -1.94 and 1.01. In 4PLM, the values of parameter  $b$  range from -1.87 to 0.91. When  $a$  parameter was examined in 2PLM, 3PLM, and 4PLM, the highest values were generally obtained in 4PLM. In 2PLM, the values of parameter  $a$  ranged from 0.85 to 2.58. In 3PLM, the value range of parameter  $a$  was [1.54, 4.76]. In 4PLM, the values of parameter  $a$  were between 1.55 and 6.64. When  $c$  parameters were examined in 3PLM and 4PLM, values were generally close to each other. While  $c$  parameter values were between 0.00 and 0.31 in 3PLM, it was between 0.00 and 0.25 in 4PLM. The  $d$  parameter estimated in 4PLM took values between 0.95 and 1.00. The item with the highest probability of incorrect answers due to carelessness has the lowest  $d$  parameter. While the lowest  $d$  parameter was in the 3rd, 7th, 11th, and 20th items, the  $d$  parameter of 16 items was estimated as 1.00.

The item parameters obtained for the X sample of 3750 students were given in Table 3.

**Table 3.** The Item Parameters Obtained for the X Sample of 3750 Students

Items	1PLM	2PLM		3PLM			4PLM			
	b	a	b	a	b	c	a	b	c	d
item 1	-1.54	1.63	-1.01	2.21	-0.38	0.35	1.93	-0.59	0.25	1
item 2	-1.5	2.43	-0.86	4.3	-0.34	0.33	3.59	-0.47	0.25	1
item 3	0.65	1.49	0.44	4.02	0.75	0.16	6.63	0.64	0.17	0.93
item 4	-0.89	1.8	-0.58	2.93	0.01	0.3	2.66	-0.09	0.25	1

item 5	-0.05	1.26	-0.03	4.43	0.66	0.3	3.49	0.55	0.25	1
item 6	-0.99	1.75	-0.64	2.41	-0.14	0.26	2.4	-0.17	0.25	1
item 7	0.21	0.85	0.25	2.01	0.94	0.27	1.88	0.85	0.25	0.99
item 8	-3.1	1.9	-1.82	1.8	-1.98	0	1.94	-1.94	0	1
item 9	-0.6	2.03	-0.4	4.26	0.16	0.28	3.97	0.1	0.25	1
item 10	0.27	2.2	0.1	4.85	0.45	0.15	4.84	0.42	0.15	1
item 11	0.4	1.41	0.28	3.91	0.71	0.2	6.63	0.61	0.21	0.94
item 12	0.16	1.43	0.1	3.76	0.62	0.23	4.14	0.58	0.24	0.99
item 13	0.49	1.27	0.37	2.91	0.78	0.19	2.96	0.75	0.19	0.99
item 14	0.08	0.93	0.11	2.88	0.88	0.31	2.12	0.74	0.25	1
item 15	-0.62	1.42	-0.44	2.36	0.2	0.29	2.19	0.11	0.25	1
item 16	-0.51	1.87	-0.35	3.82	0.22	0.27	3.65	0.16	0.25	1
item 17	-1.5	1.64	-0.98	1.56	-0.98	0	1.56	-0.99	0	1
item 18	-1.44	1.97	-0.88	2.34	-0.54	0.21	2.36	-0.55	0.21	1
item 19	0.1	1.33	0.07	2.62	0.57	0.22	2.64	0.55	0.22	1
item 20	0.53	0.92	0.53	1.7	0.97	0.19	2.28	0.71	0.21	0.86
Mean	-0.49	1.58	-0.29	3.05	0.18	0.23	3.19	0.1	0.2	0.99

When Table 3 was examined, it was seen that the highest  $b$  parameter was 0.65 and the lowest  $b$  parameter was -3.10 in 1PLM for the X sample of 3750 students. For 2PLM, the  $b$  parameter had the highest value was 0.53 and the lowest value was -1.82. In 3PLM, the  $b$  parameter took values between -1.98 and 0.97. In 4PLM, the values of parameter  $b$  range from -1.94 to 0.85. When  $a$  parameter was examined in 2PLM, 3PLM, and 4PLM, the highest values were generally obtained in 4 PLM. In 2 PLM, the values of parameter  $a$  range from 0.85 to 2.43. In 3 PLM, the value range of parameter " $a$ " was [1.56, 4.85]. In 4PLM, the values of parameter  $a$  were between 1.56 and 6.63. When  $c$  parameters were examined in 3PLM and 4PLM, values were generally close to each other. While  $c$  parameter values were between 0.00 and 0.35 in 3 PLM, it was between 0.00 and 0.25 in 4PLM. The  $d$  parameter estimated in 4PLM, on the other hand, took values between 0.86 and 1.00. While the lowest  $d$  parameter was in the 20th item, the  $d$  parameter of 14 items

was estimated as 1.00. The item with the highest probability of answering incorrectly due to carelessness was item 20.

The item parameters obtained for the Y sample of 3750 students were given in Table 4.

**Table 4.** The Item Parameters Obtained for the Y Sample of 3750 Students

Items	1PLM		2PLM		3PLM			4PLM			
	b	a	b	a	b	c	a	b	c	d	
item 1	-1.57	1.59	-1.04	1.65	-0.81	0.14	1.68	-0.78	0.16	1	
item 2	-1.53	2.75	-0.84	4.45	-0.39	0.3	3.98	-0.46	0.25	1	
item 3	0.65	1.49	0.43	4.45	0.76	0.17	6.63	0.67	0.17	0.95	
item 4	-0.98	1.74	-0.64	2.62	-0.05	0.29	2.47	-0.13	0.25	1	
item 5	0.02	1.47	0	5.15	0.61	0.27	4.73	0.57	0.25	1	
item 6	-1.08	2.08	-0.67	2.79	-0.22	0.24	2.79	-0.23	0.24	1	
item 7	0.27	0.86	0.3	2.04	0.96	0.26	2.17	0.8	0.25	0.94	
item 8	-3.1	2	-1.77	1.95	-1.89	0	2.27	-1.81	0	0.99	
item 9	-0.69	2.1	-0.45	5.1	0.16	0.3	4.61	0.09	0.25	1	
item 10	0.22	2.14	0.07	4.69	0.45	0.15	4.74	0.43	0.15	1	
item 11	0.3	1.36	0.21	3.58	0.69	0.22	4.88	0.61	0.22	0.95	
item 12	0.13	1.55	0.07	4.53	0.59	0.23	4.6	0.58	0.23	1	
item 13	0.41	1.18	0.33	3.13	0.83	0.22	3.12	0.81	0.22	1	
item 14	0.12	0.95	0.14	2.94	0.88	0.3	2.24	0.77	0.25	1	
item 15	-0.67	1.35	-0.48	2.39	0.25	0.31	2.06	0.11	0.25	1	
item 16	-0.48	1.97	-0.33	4.71	0.27	0.28	4.31	0.21	0.25	1	
item 17	-1.51	1.58	-1	1.53	-0.99	0	1.53	-0.99	0	1	
item 18	-1.47	2.21	-0.86	2.67	-0.5	0.23	2.69	-0.5	0.23	1	
item 19	0.03	1.31	0.02	2.79	0.58	0.24	2.78	0.56	0.23	1	
item 20	0.5	0.84	0.54	1.86	1.06	0.23	2.56	0.8	0.24	0.86	
Mean	-0.52	1.63	-0.3	3.25	0.16	0.22	3.34	0.11	0.21	0.98	

When Table 4 was examined, it was seen that the highest  $b$  parameter was 0.65 and the lowest  $b$  parameter was -3.10 in 1PLM for the Y sample of 3750 students. For 2PLM, the  $b$  parameter had the highest value was 0.54 and the lowest value was -1.77. In 3PLM, the  $b$  parameter took values between -1.89 and 1.06. In 4PLM, the values of parameter  $b$  range from -1.81 to 0.81. When  $a$  parameter was examined in 2PLM, 3PLM, and 4PLM, the highest values were generally obtained in 4PLM. The values of parameter  $a$  in 2PLM were between 0.86 and 2.75. In 3PLM, the value range of parameter  $a$  was [1.53,5.15]. In 4PLM, the values of parameter  $a$  were between 1.53 and 6.63. When  $c$  parameters were examined in 3PLM and 4PLM, values were generally close to each other. While  $c$  parameter values were between 0.00 and 0.31 in 3PLM, it was between 0.00 and 0.25 in 4PLM. The  $d$  parameter estimated in 4PLM, on the other hand, took values between 0.86 and 1.00. While the lowest  $d$  parameter was in the 20th item, the  $d$  parameter of 15 items was estimated as 1.00. The item with the highest probability of answering incorrectly due to carelessness was item 20.

In Table 5, the average, minimum, and maximum cut-off values of the ability parameter were given.

**Table 5.** Descriptive Statistics for Ability Parameters

IRT Model	Sample	Ability Estimation Method					
		EAP			MAP		
		Mean	Min	Max	Mean	Min	Max
1PLM	7500	0	-3.47	2.7	-0.03	-3.6	2.4
	3750-X	0	-3.45	2.7	-0.04	-3.6	2.4
	3750-Y	0	-2.97	2.7	-0.01	-3	2.4
2PLM	7500	0	-2.32	1.87	-0.03	-2.4	1.8
	3750-X	0	-2.33	1.88	-0.02	-2.4	1.8
	3750-Y	0	-1.96	1.86	-0.03	-1.8	1.8



	7500	0,01	-1.9	1.79	0.06	-1.8	1.8
3PLM	3750-X	0,01	-1.89	1.8	0.06	-1.8	1
	3750-Y	0,02	-1.92	1.79	0.07	-1.8	1.8
	7500	0	-1.94	1.73	0.04	-1.8	1.8
4PLM	3750-X	0	-1.94	1.72	0.04	-1.8	1.8
	3750-Y	0,01	-1.95	1.71	0.06	-1.8	1.8

In Table 5, it has been seen that the average values are close to zero and very close to each other if the ability estimations are made with the EAP and MAP methods. The minimum value of the ability parameter estimated by the EAP estimation was found to be -3.47 for 7500 samples. The highest ability estimate estimated by the EAP estimation was obtained in 1PLM for all samples. When the estimations obtained by the MAP method were examined, the highest ability parameter was estimated at 1PLM for all samples, while the lowest ability parameter was -3.60 in 1PLM for 7500 and 3750 X samples.

#### Findings Regarding the Differentiation of Item Parameters

Multi-way analysis of variance was applied to determine the differences between  $a$  parameters according to different IRT models in multiple groups. The obtained results were given in Table 6.

**Table 6.** Investigation of the Differentiation of Parameter  $a$  in Multiple Groups and Models

Source	F	$p$	Partial Eta Squared	Observed Power
Sample	.234	.791	.003	.086
Model	44.870	.000	.344	1.000
Sample*Model	.028	.998	.001	.055

When Table 6 was examined, a significant difference was found between  $a$  parameters obtained from 2PLM and 3PLM and 4PLM ( $F(2, 179)=44.870$ ;  $p<0.05$ ). The  $a$  parameter

was underestimated in 2PLM. The parameter  $a$  estimated in multiple groups did not differ significantly from sample to sample ( $F(2, 179)=0.234$ ;  $p>0.05$ ).

When the effect sizes in Table 6 were examined, we can say that the effect of sample and sample\*model on  $a$  parameter was non-significant and the effect of model is large based on Cohen's (1988) criteria for effect size.

Multi-way analysis of variance was applied to determine the differences between  $b$  parameters according to different IRT models in multiple groups. The obtained results were given in Table 7.

**Table 7.** Investigation of the Differentiation of Parameter  $b$  in Multiple Groups and Models

Source	F	$p$	Partial Eta Squared	Observed Power
Sample	.005	.995	.000	.051
Model	10.512	.000	.122	.999
Sample*Model	.002	1.000	.000	.050

When Table 7 was examined, a significant difference was found between the  $b$  parameters obtained from 1PLM and 3PLM and 4PLM, 2PLM and 3PLM and 4PLM ( $F(3, 239)=10.512$ ;  $p<0.05$ ). The  $b$  parameter was underestimated in 1PLM and 2PLM. The  $b$  parameter estimated in multiple groups did not differ significantly from sample to sample ( $F(2, 239)=0.005$ ;  $p>0.05$ ).

When the effect sizes in Table 7 were examined, the effect of sample and sample\*model on the  $b$  parameter was found non-significant and the effect of model is medium based on Cohen's (1988) criteria for effect size.

Multi-way analysis of variance was applied to determine the differences between  $c$  parameters according to different IRT models in multiple groups. The obtained results were given in Table 8.

**Table 8.** Investigation of the Differentiation of Parameter  $c$  in Multiple Groups and Models

Source	F	$p$	Partial Eta Squared	Observed Power
Sample	.027	.974	.000	.054
Model	1.295	.258	.011	.204
Sample*Model	.013	.987	.000	.052

When Table 8 was examined, no significant difference was found between the  $c$  parameters obtained from different IRT models and multiple groups ( $F(2, 119)=0.027$ ;  $p>0.05$ ). The estimations of  $c$  parameter according to 3PLM and 4PLM did not show a significant difference ( $F(1, 119)=1.295$ ;  $p>0.05$ ). In addition, the  $c$  parameter estimations did not show a significant difference in the samples of 7500 and 3750 students ( $p>0.05$ ).

When the effect sizes in Table 8 were examined, according to the classification developed by Cohen (1988) for the effect size, all effect of variance sources on the  $c$  parameter was found non-significant.

#### **Findings Related to the Differentiation of Ability Parameters in Subgroups**

Multi-way analysis of variance was applied to determine the differences between ability parameters according to different ability estimation methods and IRT models on a sample of 7500 students. The obtained results were given in Table 9.

**Table 9.** Investigation of the Differentiation of Ability Parameter Estimations for Sample of 7500 Students According to the IRT Models and Ability Estimation Methods

Source	F	<i>p</i>	Partial Eta Squared	Observed Power
Model	9.418	.000	.000	.997
Ability Estimation Method	1.298	.255	.000	.207
Model*Ability Estimation Method	7.198	.000	.000	.983

When Table 9 was examined, a significant difference was found between the ability parameters obtained from 1PLM and 3PLM and 4PLM, 2PLM and 3PLM and 4PLM ( $F(3, 59999)=9.418$ ;  $p<0.05$ ). The ability parameter was underestimated at 1PLM and 2PLM compared to 3PLM and 4PLM. The ability parameter estimated according to the EAP and MAP method did not show a significant difference ( $F(1, 59999)=1.298$ ;  $p>0.05$ ).

When the effect sizes in Table 9 were examined, all effect of variance sources on the ability parameters was found non-significant based on the classification developed by Cohen (1988) for the effect size.

Multi-way analysis of variance was applied to determine the differences between ability parameters according to different ability estimation methods, IRT models, and samples for samples of 3750 students. The obtained results were given in Table 10.

**Table 10.** Investigation of the Differentiation of Ability Parameter Estimations for Samples of 3750 Students According to the Samples, Models, and Ability Estimation Methods

Source	F	<i>p</i>	Partial Eta Squared	Observed Power
Model	100.782	.000	.005	1000
Ability Estimation Method	5.303	.021	.000	.634
Sample	51.107	.000	.001	1000
Model* Ability Estimation Method	11.228	.000	.001	.999
Model*Sample	62.164	.000	.003	1000
Ability Estimation Method* Sample	.449	.503	.000	.103
Model* Ability Estimation Method*Sample	1.218	.301	.000	.330

When Table 10 was examined, the ability parameter estimation between models other than 1PLM and 2PLM in two samples of 3750 students differs significantly from each other ( $F(3, 59999)=100.782$ ;  $p<0.05$ ). At the same time, the ability parameters obtained from different samples and different ability parameter estimation methods also show significant differences ( $p<0.05$ ).

When the effect sizes in Table 10 were examined, all effect of variance sources on the ability parameters was found non-significant.

#### **Findings for Marginal Reliability Coefficient**

The marginal reliability coefficients for the EAP estimation were given in Table 11.

**Table 11.** Marginal Reliability Coefficients for EAP Estimation

IRT Model	Sample	Marginal Reliability Coefficient
	7500	0.855
1PLM	3750-X	0.855
	3750-Y	0.856
	7500	0.865
2PLM	3750-X	0.864
	3750-Y	0.866
	7500	0.851
3PLM	3750-X	0.846
	3750-Y	0.857
	7500	0.852
4PLM	3750-X	0.848
	3750-Y	0.857

When Table 11 was examined, the highest reliability coefficient was found to be 0.866 for the 3750 Y sample. The lowest reliability coefficient (0.846) was estimated in 3PLM for the 3750 X sample. Considering the marginal reliability coefficients, it was seen that values were very close to each other under all conditions.

## DISCUSSION AND CONCLUSION

In this study, 1PLM, 2PLM, 3PLM, and 4PLM comparisons were done based on data of the national transition examination. For this, first of all, the validity and reliability analysis of the measurements obtained from randomly selected 7500 students who responded to the mathematics subtest in the TPSEE 2017 April dataset and when the dataset was randomly divided into two were performed. Then, one-dimensionality and local independence assumptions, which are the basic assumptions of parametric one-dimensional IRT, were examined. After meeting the assumptions, model fit indices in 1PLM, 2PLM, 3PLM, and 4PLM were examined and item and ability parameters based

on EAP and MAP were calculated for each model. In addition, reliability analysis was performed and reported in IRT for each model.

As a result of the findings, it was seen that the best model-data fit in the data set used was in 3PLM. When the studies carried out on different data in the literature were examined, it was seen that the results may differ. Publications about 4PLM, whose use has increased since the 2000s, have been published on different data sets in the literature (Barton & Lord, 1981; Erdemir & Önen, 2019; Feuerstahler & Waller, 2014; Liao et al,2012; Loken & Rulison, 2010; Magis, 2013; Reise & Waller, 2003; Rulison & Loken, 2009; Rupp, 2003, Waller & Reise, 2010, Yalçın, 2018; Yen et al., 2012). When all these studies were examined, it has been revealed in many studies that 4PLM has better model data fit than other dichotomous models (Barton & Lord, 1981; Erdemir & Önen, 2019; Loken & Rulison, 2010; Rulison & Loken, 2009; Rupp, 2003; Waller & Reise, 2010). In computerized adaptive test (CAT) studies, it was observed that the ability estimation was obtained with 4PLM with a lower standard error (Magis, 2013; Yen et al., 2012).

In Erdemir and Önen (2019)'s study, when the overall model data fit was handled with 5 different methods (-2LL, AIC, BIC,  $\chi^2$ , and SMRSR), four of the methods indicated that 4PLM showed a relatively better fit. Loken and Rulison (2010) compared the freely estimated parameter  $d$  with the parametric IRT model in their study with real and simulated data and found that 4PLM exhibited a better fit. However, Barton and Lord (1981), who pioneered studies on 4PLM, found that the 3PLM model had a better model fit index than the 4PLM. However, in these studies conducted in the early 1980s, the  $d$  parameters were not freely estimated. Yalçın (2018), who also uses the MixIRT model, reached a parallel conclusion with Barton and Lord (1981). While interpreting the results obtained from the studies, the characteristics of the data set should not be ignored. In the case of considering the  $d$  parameter as a constant value in the 4PLM model, the results obtained differ from the study of Barton and Lord (1981) in the literature.

In this study, while item parameters did not differ from sample to sample, ability parameters is seen as differing partially in samples of 3750 students. But when the effect size of sample on ability parameter examined according to the classification developed

by Cohen (1988) for the effect size, it was found non-significant difference. Therefore, this result is similar from the study of Fan (1998), who found that parameter invariance was provided in both theories (IRT and CTT). There are many studies in the literature on the invariance of parameters obtained using IRT (e.g. Fan, 1998; Fan & Ping, 1999; Kelkar, Wightman, & Luecht, 2000; Doğan & Tezbaşaran, 2003; Acar & Kelecioğlu, 2008; Custer et al., 2008; Adedoyin, Nenty & Chilisa, 2008; Immekus & Maller, 2009; Adedoyin, 2010; Galdin & Laurencelle, 2010; Sünbül & Erkuş, 2013; Doğan & Kılıç, 2017). When these studies were examined, it is seen that the assumption of parameter invariance in IRT was largely met (e.g. Fan, 1998; Fan & Ping, 1999; Kelkar, Wightman, & Luecht, 2000; Acar & Kelecioğlu, 2008; Custer et al., 2008; Adedoyin, Nenty & Chilisa, 2008; Adedoyin, 2010; Sünbül & Erkuş, 2013). In some studies, it was found that item parameter invariance was not fully achieved (Doğan & Tezbaşaran, 2003; Immekus & Maller, 2009; Galdin & Laurencelle, 2010; Doğan & Kılıç, 2017). In a study, it was found that ability parameter invariance was provided to a greater extent than item parameter invariance (Doğan & Kılıç, 2017). This study is not a parameter invariance study, but an investigation of IRT parameter estimations and reliability in multiple groups can show us how parameters and reliability coefficients can be differed by different groups from the same universe.

Another result of the study was that the item parameters ( $a$  and  $b$ ) partially differed according to the model used. The  $c$  parameter, on the other hand, did not differ according to the sampling (3750-X, 3750-Y, and 7500) and the model used (3PLM and 4PLM). Ability parameter estimations did not differ according to the method (EAP-MAP) used in 7500 samples. In 3750 samples, it seem like there were differences when the model, method, sample together, and their interaction were considered. But if the effect sizes on ability parameter in 3750 samples was examined, according to the classification developed by Cohen (1988) for the effect size, it was found non-significant.

One of the contributions of current research is to handle parameter estimations based on different models and estimation methods in sub-groups or multiple groups. This research did not address parameter invariance in subgroups with methods such as the IRT



Likelihood Ratio Test. Within the scope of the research, it was only examined whether the parameters differed in different conditions. In future research, parameter invariance can be tested in subgroups using methods such as the IRT Likelihood Ratio Test. A limitation of this research is the analysis of dichotomous data obtained from the mathematics subtest in the TPSEE 2017 exam. Similar studies can be performed on different datasets (polytomous, dichotomous or mixed) and different sample size. In addition, different models can be tested under different simulation conditions. Similar studies can be conducted to compare nonparametric IRT models and bayesian models. IRT comparisons can be done on multidimensional data sets.

## REFERENCES

- Acar, T., & Kelecioğlu, H. (2008). Genelleştirilmiş aşamalı doğrusal model ile rasch modelinin parametre değişmezliğinin karşılaştırılması. *1st National Congress of Measurement and Evaluation in Education and Psychology*, 14-16 Mayıs, Ankara, 181-193.
- Adedoyin, O. O. (2010). Investigating the invariance of person parameter estimates based on classical test and item response theories. *International Journal of Education Science*, 2(2), 107-113. <https://doi.org/10.1080/09751122.2010.11889987>
- Adedoyin, O. O., Nenty, H. J., & Chilasa, B. (2008). Investigating the invariance of item difficulty parameter estimates based on CTT and IRT. *Educational Research and Review*, 3(2), 83-93. <https://doi.org/10.5897/ERR.9000209>
- Baker, F. B. (2001). *The basics of item response theory*. United States of America: ERIC Clearinghouse on Assessment and Evaluation.
- Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item response model. *Research Bulletin*, 81-20.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores* (pp. 392-479). Reading, MA: Addison-Wesley.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd edition. New York: Academic Press.
- Custer, M., Sharairi, S., Yamazaki, K., Signatur, D., Swift, D., & Frey, S. (2008). A paradox between IRT invariance and model-data fit when utilizing the one-parameter and three-parameter models. *Annual Meeting of the American Educational Research Association*, 24-28 March, New York, 70-71.
- DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.
- Doğan, N., & Kılıç, A. F. (2017). Madde tepki kuramı yetenek ve madde parametreleri kestirimlerinin değişmezliğinin incelenmesi. ss 297-314. Demirel, Ö., Dinçer, S., ed. *Küreselleşen Dünyada Eğitim*, Pegem Akademi, Ankara.
- Doğan, N., & Tezbaşaran, A. A. (2003). Klasik test kuramı ve örtük özellikler kuramının örneklem bağlamında karşılaştırılması. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi*, 25(25), 58-67.
- Doğruöz, E., & Arıkan, Ç. A. (2020). Comparison of different ability estimation methods based on 3 and 4PL item response theory. *PAU Journal of Education* 50, 50-69. <https://doi.org/10.9779/pauefd.585774>
- Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(1), 5-18. <https://doi.org/10.1007/s11136-007-9198-0>
- Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. New Jersey: Lawrence Erlbaum.

- Erdemir, A., & Önen, E. (2019). Bir, iki, üç ve dört parametrelili lojistik madde tepki kuramı modellerinin karşılaştırılması [Comparison of 1PL, 2PL, 3PL and 4PL item response theory models]. *e-Turkish Studies*, 14(1), 307-332. <https://doi.org/10.7827/TurkishStudies.14745>
- Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357–381. <https://doi.org/10.1177/0013164498058003001>
- Fan, X., & Ping, Y. (1999). Assessing the effect of model-data misfit on the invariance. *Journal of Mathematical and Statistical Psychology*, 42, 139-167.
- Feuerstahler, L. M., & Waller, N. G. (2014). Estimation of the 4-parameter model with marginal maximum likelihood. *Multivariate behavioral research*, 49(3), 285-285. <https://doi.org/10.1080/00273171.2014.912889>
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer Nijhoff.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. London: Sage.
- Han, K. T., & Hambleton, R. K. (2014). *User's manual for WINGEN 3: windows software that generates IRT model parameters and item responses* (Center for Educational Assessment Report No. 642). Amherst, MA: University of Massachusetts.
- 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. <https://doi.org/10.1080/15366367.2021.1897398>
- Kaplan, R. M. & Saccuzo, D. P. (1997). *Psychological testing: principles, applications and issues*. Pacific Grove: Brooks Cole Pub. Company.
- Kean, J., & Reilly, J. (2014). Item response theory. *Handbook for clinical research: Design, statistics and implementation*, 195-198.
- Kelkar, V., Wightman, L.F., & Luecht, R.M. (2000). Evaluation of the IRT parameter Invariance property for the MCAT. *Annual Meeting of the National Council on Measurement in Education*, 25-27 April, New Orleans.
- Kolen, M. J. & Brennan, R. L. (2014). *Test equating, scalling, and linking*. (third edition). USA: Springer.
- Lembke, E. & Stecker, P. (2007). *Curriculum-based measurement in mathematics: an evidence-based formative assessment procedure*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.
- Liao, W., Ho, R., & Yen, Y. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality*, 40(10), 1679–1694. <https://doi.org/10.2224/sbp.2012.40.10.1679>
- Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *The British Journal of Mathematical and Statistical Psychology*, 63(3), 509–25. <https://doi.org/10.1348/000711009X474502>

- Lord, F. M. (1952). *A theory of test scores* (Psychometric Monograph No. 7). Iowa City, IA: Psychometric Society, 35.
- Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement, 37*(4), 304-315. <https://doi.org/10.1177/0146621613475471>
- R Core Team. (2021). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.
- Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*(2), 164-184. <https://doi.org/10.1037/1082-989X.8.2.164>
- Robitzsch, A. (2021). sirt: Supplementary item response theory models. R package version 3.11-21, <https://cran.r-project.org/web/packages/sirt/sirt.pdf>
- Rulison, K. L., & Loken, E. (2009). I've fallen and i can't get up: can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*(2), 83-101. <https://doi.org/10.1177/0146621608324023>
- Rupp, A. A. (2003). Item response modeling with BILOG-MG and MULTILOG for windows. *International Journal of Testing, 3*(4), 365-384. [https://doi.org/10.1207/S15327574IJT0304\\_5](https://doi.org/10.1207/S15327574IJT0304_5)
- Sünbül, Ö., & Erkuş, A. (2013). Madde parametrelerinin değişmezliğinin çeşitli boyutluluk özelliği gösteren yapılar da madde tepki kuramına göre incelenmesi. *Mersin Üniversitesi Eğitim Fakültesi Dergisi, 9*(2), 378- 398.
- U. S. Department of Education (2001). The elementary and secondary education act (The No Child Left Behind Act of 2001). Retrieved September 3, 2019, from <http://www.ed.gov/legislation/ESEA02>
- Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with nonstandard item response theory models: fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. S. E. Embretson (Ed.), In *Measuring psychological constructs: Advances in modelbased approaches* (147-173). Washington, DC, US: American Psychological Association. <http://dx.doi.org/10.1037/12074-007>
- Wu, M., Tam, H. P., & Jen, T. H. (2016). Classical test theory. In *Educational measurement for applied researchers* (pp. 73-90). Springer, Singapore.
- Yalçın, S. (2018). Data fit comparison of mixture item response theory models and traditional models. *International Journal of Assessment Tools in Education, 5*(2), 301-313. <https://doi.org/10.21449/ijate.402806>
- Yen, Y., Ho, R., Liao, W., & Chen, L. (2012). Reducing the impact of inappropriate items on reviewable computerized adaptive testing. *Educational Technology & Society, 15*, 231-243.

## GENİŞ ÖZET

*Şans başarısını göz ardı etmesi, test merkezli olması, tüm beceri aralığı için tek bir hata tahmini yapması, güvenilirliği tahmin etmek için paralel testlere ihtiyaç duyması, madde istatistiklerinin gruba bağlı olması ve yeteneği tahmin etmenin teste bağlı olması Klasik Test Teorisi'nin dezavantaj ve sınırlamalardan bazılarıdır (Embretson ve Reise, 2000; Hambleton vd., 1991). Madde Tepki Kuramı'nın (MTK) avantajları ise, bireysel yetenek parametresini tahmin etmesi ve parametreleri tahmin ederken gruptan ve koşullardan bağımsız özelliğine sahip olmasıdır (DeMars, 2010).*

*Parametrik tek boyutlu MTK modellerini belirlerken dikkat edilmesi gereken noktalardan biri de madde cevap kategori sayısıdır (Edelen ve Reeve, 2007). Çoktan seçmeli testler ikili puanlanır ve bu testler için MTK'deki parametre sayısı dikkate alınarak 1-0 puanlanan çeşitli modeller vardır. 1, 2 ve 3 parametrelili lojistik modeller (PLM) en sık kullanılanlardır ve üst asimptot parametresini üreten 4PLM'ye dayalı tahminler yapmak da mümkündür (Edelen ve Reeve, 2007). 4PLM, Barton ve Lord (1981) tarafından d; parametresinin 3PLM'ye eklenmesiyle oluşturulmuştur. 4PLM ile yüksek yetenekli katılımcıların, kolay bir maddeyi yanıtlatırken hata yapma olasılığını hesaba katmaktadır. 1.00'dan küçük bir değere sahip üst asimptot eklenmesiyle, yetenek düzeyi yüksek olan bir katılımcının kolay bir maddeye yanlış yanıt vermesi durumunda yetenek ölçeğinde önemli ölçüde değişmemesini sağlar. 2000'li yıllardan itibaren kullanımı artan 4PLM ile ilgili yayınlar literatürde farklı veri setleri üzerinde yayınlanmıştır (Barton ve Lord, 1981; Erdemir ve Önen, 2019; Feuerstahler ve Waller, 2014; Liao, vd., 2012; Loken ve Rulison, 2010; Magis, 2013; Reise ve Waller, 2003; Rulison ve Loken, 2009; Rupp, 2003, Waller ve Reise, 2010, Yalçın, 2018; Yen vd., 2012). Tüm bu çalışmalar incelendiğinde 4PLM'nin diğer ikili modellere göre daha iyi model veri uyumuna sahip olduğu birçok çalışmada ortaya konulmuştur (Barton ve Lord, 1981; Erdemir ve Önen, 2019; Loken ve Rulison, 2010; Rulison ve Loken, 2009; Rupp, 2003; Waller ve Reise, 2010). Bilgisayar Ortamında Bireye Uyarlanmış Testler (CAT) çalışmalarında yetenek tahmininin 4PLM ile daha düşük standart hata ile elde edildiği görülmüştür (Magis, 2013; Yen vd., Chen, 2012). Türkiye'de ise 4PLM'nin uygulamaları sınırlı sayıdadır (örneğin; Erdemir ve Önen, 2019; Yalçın, 2018). 4PLM'yi dâhil ederek yapılan çoklu gruplarda madde ve yetenek parametrelerinin ve güvenilirliklerin farklılaşmasını inceleyen tamamen benzer bir araştırma bulunmamıştır.*

*Araştırmanın örneklemini TEOG 2017 Nisan sınavına giren ve A kitapçığı alan rastgele seçilmiş 7500 8. sınıf öğrencisi oluşturmaktadır. Alt gruplar arasında analiz yapabilmek için örneklemin tamamı rastgele 3750 öğrenciye ayrılmıştır.*

*Bu çalışmada gerçek verilere dayalı olarak 1PLM, 2PLM, 3PLM ve 4PLM karşılaştırması yapılmıştır. Bunun için öncelikle TEOG 2017 Nisan veri setinde matematik alt testini yanıtlayan rastgele seçilen 7500 öğrenciden ve veri seti rastgele ikiye bölünerek elde edilen ölçümlerin geçerlik ve güvenilirlik analizleri yapılmıştır. Daha sonra parametrik tek boyutlu MTK'nin temel varsayımları olan tek boyutluluk ve yerel bağımsızlık varsayımları incelenmiştir. Varsayımlar sağlandıktan sonra 1 PLM, 2PLM, 3PLM ve 4 PLM'deki model uyum indeksleri incelenmiş ve her bir model için madde ve yetenek parametreleri hesaplanmıştır. Ayrıca, her model için MTK'de güvenilirlik analizi yapılmış ve raporlanmıştır.*


*Elde edilen bulgular sonucunda kullanılan veri setinde en yüksek model-veri uyumunun 3 PLM'de olduğu görülmüştür. Bu çalışmada madde parametreleri örneklemden örnekleme farklılık*

*göstermezken, 3750 öğrenci örnekleminde yetenek parametreleri kısmen farklılık göstermiştir. Araştırmanın bir diğer sonucu, madde parametrelerinin (a ve b) kullanılan modele göre kısmen farklılaştığıdır. c parametresi ise örnekleme (3750-X, 3750-Y ve 7500) ve kullanılan modele (3 PLM ve 4 PLM) göre farklılık göstermemiştir. 7500 örnekte kullanılan yetenek parametresi kestirim yöntemine (EAP ve MAP) göre yetenek parametresi tahminleri farklılık göstermemiştir. 3750 örnekleme model (1PLM, 2PLM, 3PLM, 4PLM), yöntem (EAP-MAP), örneklem birlikte ve etkileşimleri dikkate alındığında farklılıklar ortaya çıkmıştır.*

*Bu araştırma, MTK Olabilirlik Oranı Testi gibi yöntemlerle alt gruplarda parametre değişmezliğini ele almamıştır. Araştırma kapsamında sadece parametrelerin farklı koşullarda farklılık gösterip göstermediği incelenmiştir. Gelecekteki araştırmalarda, parametre değişmezliği, MTK Olabilirlik Oranı Testi gibi yöntemler kullanılarak alt gruplarda test edilebilir. Bu araştırmanın bir sınırlılığı, TEOG 2017 sınavında matematik alt testinden elde edilen ikili puanlanan verilerin analizidir. Benzer çalışmalar farklı veri setleri üzerinde (çoklu veya ikili veya karma) gerçekleştirilebilir. Ayrıca, farklı simülasyon koşulları altında farklı modeller test edilebilir. Parametrik olmayan MTK modellerini ve Bayes modellerini karşılaştırmak için benzer çalışmalar yapılabilir. MTK karşılaştırmaları çok boyutlu veri setleri üzerinde yapılabilir.*

## ORCID

Serap BÜYÜKKIDIK  <https://orcid.org/0000-0003-4335-2949>

Hatice İNAL  <https://orcid.org/0000-0002-2813-0873>

## Contribution of Researchers

SB wrote abstract, introduction, method, findings, conclusion part. SB had roles in the conceptualization, resources, data analysis, reporting, drafting, reviewing and editing.

HI wrote some section of introduction, method, findings parts. HI had roles in data analysis, drafting, reviewing and editing.

## Acknowledgements

We would like to thank Ministry of National Education for giving permission 2017 April TPSEE data to use for scientific purposes.

**Conflict of Interest**

The researchers do not have any personal or financial conflicts of interest with other individuals or institutions related to the research.

**Ethics Committee Declaration**

Within the scope of the research, 2017 April TPSEE mathematics data, which is a data of students participating in a central exam system throughout Turkey, was used. The data was collected by the Ministry of National Education and the data was obtained by getting permission to use the data for scientific purposes from Ministry of National Education. Researchers did not collect the data themselves but performed analysis on the existing data. Therefore, this study does not require ethics committee approval.





