Araştırma Makalesi/Research Article

Pamukkale Üniversitesi Eğitim Fakültesi Dergisi (*PAU Journal of Education*) 50:50-69 [2020]
doi: 10.9779/pauefd.585774

Comparison of Different Ability Estimation Methods Based on 3 and 4PL Item Response Theory

3 ve 4PL Madde Tepki Kuramı Modellerine Göre Farklı Yetenek Kestirim Yöntemlerinin Karşılaştırılması

Ebru DOĞRUÖZ*, Çiğdem AKIN ARIKAN**

• ***Geliş Tarihi:*** *02.07.2019* • ***Kabul Tarihi:*** *05.02.2020* • ***Çevrimiçi Yayın Tarihi:*** *07.02.2020*

Abstract

This research analyzed the two-category Item Response Theory (IRT) models as part of different ability estimation methods. The research was carried out in consideration of responses to 20 items under the Mathematics subtest of TEOG (National Transition from Primary to Secondary Education) exam by the 8th-grade students in 2015-2016. The study group consisted of 400 students who were randomly selected from the students participated in the TEOG exam. Ability estimations and standard error values for these estimations were calculated based on the data. These estimations were compared by two-way analysis of variance (ANOVA) for repeated measurements According to the research findings; it was revealed that the four-parameter logistic (4PL) item model fit better. In terms of ability estimation methods, the accuracy of Weighted Likelihood Estimation (WLE) was higher than Maximum A Posteriori (MAP) and Expected A Posteriori (EAP). WLE and MAP ability estimation model gave lower standard error values compared to the 4PL and 3PL model, respectively. The highest marginal reliability coefficient value for the 3PL model was calculated using estimations made according to MAP while estimations made according to WLE were used for the 4PL model. According to the research findings, it was concluded that the accuracy of ability scores obtained by the WLE estimation method under the 4PL model was higher.

Keywords: ability estimation methods, item response theory, 3 PLM, 4PLM

Cited:

Doğruöz, E. ve Akın-Arıkan, Ç. (2020). Comparison of Different Ability Estimation Methods Based on 3 and 4PL Item Response Theory. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 50,* 50-69.doi: 10.9779/pauefd.585774

*Dr. Öğr. Üyesi, Çankırı Karatekin Üni. Edebiyat F. Eğitim Bil. Böl., 8706ebru@gmail.com, https://orcid.org/0000-0001-6572-274X

** Dr. Öğr. Üyesi, Ordu Üni. Eğitim F. Eğitim Bil. Böl., akincgdm@gmail.com, https://orcid.org/0000-0001-5255-8792

Öz

Bu araştırmada iki kategorili Madde Tepki Kuramı modelleri, farklı yetenek kestirim yöntemleri bağlamında incelenmiştir. Araştırma 2015-2016 yılında 8. Sınıf öğrencilerin TEOG sınavının matematik alt testinde yer alan 20 maddeye verdikleri yanıtlar ışığında gerçekleştirilmiştir. Bu verilerden seçkisiz olarak seçilen 4000 yanıtlayıcı, çalışma grubunu oluşturmaktadır. Veriler üzerinden yetenek kestirimleri ve bu kestirimlere ait standart hata değerleri hesaplanmıştır. Bu kestirimler tekrarlı ölçümler için iki faktörlü varyans analizi (ANOVA) kullanılarak karşılaştırılmıştır. Araştırma bulguları 4PL modelin daha iyi uyum gösterdiğini ortaya çıkartmıştır. WLE yetenek kestirim yönteminin doğruluğu MAP ve EAP yetenek kestirim yönteminin doğruluğundan daha yüksektir. 4PL modele göre WLE, 3PL modele göre MAP yetenek kestirim modelinin standart hata değeri daha düşüktür. En yüksek marjinal güvenirlik katsayı değeri 3PL model için MAP, 4PL model için WLE yöntemine göre gerçekleştirilen kestirimlerden hesaplanmıştır. Araştırma bulgularına dayalı olarak 4 PL model altında WLE kestirim yöntemine göre gerçekleştirilen yetenek puanlarının doğruluğunun yüksek olduğu sonucuna ulaşılmıştır.

Anahtar sözcükler: yetenek kestirim yöntemleri, madde tepki kuramı, 3PLM, 4PLM

Atıf:

Doğruöz, E. ve Akın Arıkan, Ç. (2020). 3 ve 4PL Madde Tepki Kuramı Modellerine Göre Farklı Yetenek Kestirim Yöntemlerinin Karşılaştırılması. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, . Pamukkale Üniversitesi Eğitim Fakültesi Dergisi, 50*, 50-69.doi: 10.9779/pauefd.585774

Introduction

Item Response Theory (IRT) is described as the relation between the level of the individual's ability and the item characteristics with responses of the individual to the item. The IRT is based on the assumption that individuals' abilities can be estimated independently of the items (Hambleton & Swaminathan, 1985). IRT models consist of the Rasch model, 1, 2 and 3 Parameter Logistic (PL) models for dichotomous responses. In addition to these models, there is the 4PL model within the scope of the literature of IRT models. Results of the analyses on the characteristics of the test items according to the IRT showed that the use of additional item parameters increased the accuracy and precision of the estimations of parameters characterizing the individuals. Kılıç (1999), found that the 3PL model was more compatible with ÖSS (National Student Selection Exam in Turkey) data of 1993 compared to 1 and 2 PL models. Similarly, it was shown that ability values estimated according to the 3 PL model in consideration of Turkish and Social Sciences subtests of OKS (National Secondary School Institutions Student Selection and Placement Test in Turkey) of 2002 had a more invariable characteristic compared to the values estimated according to 1 PL and 2 PL models (Can, 2003). These studies and some other studies made similar inferences by estimating item parameters according to the 1, 2 and 3PL models only (Barton & Lord, 1981; Baykul, 1979; Berberoğlu, 1988; Can, 2003; Kılıç, 1999; Reise & Waller, 2003; Yapar, 2003; Yeğin, 2003).

The 3PL model, which is quite popular among IRT models, is one of the unidimensional IRT models developed for dichotomous responses. In this model developed by Birnbaum (1968), the possibility of a correct response to item i for an individual j at θ ability level is calculated as follows:

$$Pi(\theta) = ci + (1 - ci)\frac{e^{Dai(\theta - bi)}}{1 + e^{Dai(\theta - bi)}}$$

Here, a_i, is determined as the discrimination parameter for item *i;* b_i, as the item difficulty for item *i* and c_i, as the correct response possibility for an individual at the lowest ability level or the success by chance. The b parameter takes a value usually between -2.00 and +2.00 and the parameter a is theoretically specified to be valued in the range of $-\infty$ and $+\infty$, it usually takes a value between 0 and 2 (Hambleton & Swaminatthan, 1985). Also, items with a negative parameter a should be omitted from the test (DeMars, 2010). In the 1 and 2PL models, when an individual at a low ability level response to difficult items correctly, the correct response possibility approaches 0. When an individual at a high ability level response to an easy item correctly, the correct response possibility approaches 1. Nevertheless, this hypothesis may not always be true. An individual knowing nothing could still select the correct answer by chance (Bar-Hillel, Budescu, & Attali, 2005; Gardner-Medwin & Gahan, 2003; Yen, Ho, Chen, Chou, & Chen, 2010). Besides, students at a high ability level may on occasion miss items that they should have answered correctly when they are anxious, careless, distracted by poor testing conditions, or even when they answered the item wrong (Hockemeyer, 2002; Rulison & Loken, 2009). Under these conditions, the 3PL model may lead to a low success level for a student at a high ability level who makes a careless mistake on an easy item (Barton & Lord, 1981; Rulison & Loken, 2009). More specifically, the low asymptote in the 3PL IRT model may accommodate a situation where a student at a low ability level makes a correct guess on a difficult item.

However, the upper asymptote of 1 in the 3PL model assigns a possibility of 0 when a student at a high ability level fails on an easy item.

Another IRT model is the 4PL model developed by adding an inattention parameter to the 3PL model (Barton & Lord, 1981). According to this model, the possibility of a correct response to item *i* is as follows:

$$Pi(\theta) = ci + (yi - ci)\frac{e^{Dai(\theta - bi)}}{1 + e^{Dai(\theta - bi)}}$$

In this equation, the upper asymptote shown as *di* is the inattention parameter. In addition to a, b, and c parameters, d parameter (the upper asymptote) allows values less than 1.00 and theoretically, it can be between 0.00 and 1.00. With the addition of the upper asymptote having a value less than 1.00, when a student at a high ability level answers an easy item wrong, its position in the ability scale does not change significantly. In other words, d parameter estimates possibilities where a student at a high ability level answers items with low-level difficulty wrong. To determine whether the upper asymptote is to be changed or not, standard tests can increase measurement precision, and Barton and Lord (1981) compared the 3PL model and 4PL model under two upper asymptote values d=0.99 and 0.98. When the test scores from Scholastic Aptitude Test (SAT) Verbal and Mathematics, Graduate Record Examination Verbal, and Advanced Placement Calculus AB Examination were evaluated, the results showed that changes in the ability estimation are very small in terms of significance (Barton and Lord, 1981).

Rulison and Loken (2009) showed that the 4PL model (with the upper asymptote d=0.98) may decrease estimation error for students at a high ability level who got off to a bad start. In this case, the 4PL IRT offers an opportunity for individuals to correct inattentive errors in the Computer Adaptive Test. To study the general implementation of the 4PL in detail, Loken and Rulison (2010) estimated item parameters for this model and evaluated model compliance and its performance in the IRT test which is not an experimental standard. In this research, the 4PL model was successfully applied to measure adolescent guilt experimentally.

The use of different models in IRT affects the accuracy of ability estimation significantly. However, the use of different ability estimation also provides significant information on the precision of ability estimation (Borgatto, Azevedo, Pinherio & Andrade, 2015; Ching-Fung, 2002; Rose, 2010; Wainer & Thissen, 1987; Wang & Vispoel, 1998). The ability parameter may be estimated with item parameters or pre-estimated, i.e. known item parameters. Maximum Likelihood Estimation (MLE) (Baker, 1992), Expected A Posteriori (EAP) (Bock & Aitkin, 1981; Bock & Mislevy, 1982), Weighted Likelihood Estimation (WLE) (Warm, 1989), Maximum a Posteriori (MAP) (Samejima, 1969), logistic regression (Reynolds, Perkins & Brutten, 1994), and Minimum Chi Quadrant (MCQ) (Zwinderman & van den Wollenberg, 1990) are only a few of the ability parameter estimation methods. This research benefited from the WLE method, EAP and MAP methods as part of Bayesian-based approaches. Information on these methods is given below.

Weighted Likelihood Estimation (WLE) method maximizes the likelihood function over the range of possible values of an ability. This method's function is also known as the bias correction term (Warm, 1989).

According to the *Maximum a Posteriori (MAP)* method, the ability estimation of an individual is the value that maximizes the posterior probability density function. This method enables lower standard error values to be achieved even when an individual answers all items correctly or wrong (Hambleton & Swaminathan, 1985).

Unlike the MAP method, the *Expected A Posteriori* (EAP) method is not an iterative method. Both the EAP and MAP methods use the posteriori distribution, but EAP uses the MAP mode when using the average of the posteriori distribution. According to this method, the assumption of normality and mixed iterative mathematical calculations are not required at every stage of the estimations. It also performs skill estimation in cases where the individual does not respond correctly or responds to all of the test items correctly. EAP estimation allows talent estimation of individuals with 0 and full scores (Embretson & Reise, 2000; Hambleton, Swaminathan & Rogers, 1991).

In the literature, there are views suggesting that the 3PL model is the one that fits the best according to the research estimating 1, 2, and 3PL models and testing model fit. One of the researches on this subject is by Çelik (2001) in which 1, 2, and 3PL model's level of fitness was analyzed in consideration of data obtained from Mathematics and Science subtests of National Secondary School Institutions Student Selection and Placement Test carried out by the Republic of Turkey Ministry of National Education (MEB). In this research, it was concluded that the model that fits the best in terms of the Mathematics subtest is the 3PL model. Another research on this subject is by Önder (2007) which explores the best-fit model of IRT-based models in consideration of data obtained from Science Test under Özdebir ÖSS 2004 D-II exam. Similarly, the research carried out by Taşdelen Teker, Kelecioğlu and Eroğlu (2013) found that, of the two category IRT models, the 3PL model is the one that fits the best in consideration of data obtained from Science subtest of 2009 Placement Test. Besides, only a few numbers of researches are incorporating the 4PL model as well for binary scored items within the framework of IRT. In one of these studies, items and ability parameters estimated according to the 1, 2, 3, and 4PL models were compared. The research has revealed that the estimation made under the 4PL model has a standard error lower than the other three models and the ability parameter was estimated more accurately in this model (Magic, 2013). Another study benefited from the Low Self-Esteem (LSE) scale under the Minnesota Multiphasic Personality Inventory Adult Form (MMPI-2) suggested that parameters were better estimated in the 4PL model (Reise & Waller, 2003). The ability estimation was advised to be applied according to the 4PL model for studies benefiting from Computerized Adaptive Test (CAT) since it provides lower standard error value (Rulison & Loken, 2009; Yen, Ho, Laio, Chen & Kuo, 2012).

Within this general framework in the present research, ability estimations methods compared based on the 3PL model, which assume the correct response possibility to an item for an individual at a low ability level, and 4PL model, which assume the wrong response possibility to an easy item due to inattention for an individual at a high ability level. Therefore, in this research we aimed to determine the best-fit IRT model and ability estimation method based on real data. In line with this, the research questions of the present study were given below:

1. What are the ability estimations made according to the ability estimation models and methods, and the standard error values to the ability estimations?

2. Which of the 3 and 4PL ability estimation models are best-fit to data?
3. Does the accuracy of ability estimations show significant variation according to the estimation models and methods?
4. Does the accuracy of standard errors to the ability estimations show significant variation according to the estimation models and methods?
5. Do marginal reliability coefficients differ?

Method

This research was descriptive which analyzed the model-data fit and the accuracy of item-parameter estimations comparatively based on the 3 and 4PL models.

Study Group

This research was carried out based on data obtained from the Mathematics subtest of the TEOG exam held in the 2015-2016 school year. TEOG is an exam which is held for the 8th-grade students in two semesters and consists of Turkish, Mathematics, Science, History of Turkish Revolution, Foreign Language, and Religion subtests. The analysis was carried out on a study group of 4000 students selected randomly after missing data and the full score was taken out for this research by the Directorate General for Measurement, Assessment, and Examination Services under the Republic of Turkey Ministry of National Education.

Data Collection Tool

As the data collection tool, this research used 20 items under the Mathematics subtest of the TEOG exam held for the 8th-grade students in the 2015-2016 school year fall semester. Test and item statistics to Mathematics subtest are given in Table 1 according to Classical Test Theory (CTT).

Table 1. Test Statistics of the Mathematics Subtest

Test	$\bar{x}_b$	$\bar{x}_a$	KR-20
Mathematics	0.42	0.61	0.82

(MEB, 2016) $\bar{x}_b$: mean of item difficulty; $\bar{x}_a$: mean of item discrimination index

As it is seen in Table 1, Mathematics subtest is a medium-difficulty and can distinguish between the lower and upper groups as desired. The test has high reliability.

Analysis of Data

Comparing EAP, WLE, and MAP ability estimation methods according to the 3 and 4PL models based on data obtained from the Mathematics subtest of TEOG held in 2015, this research tested whether IRT assumptions were to be met or not before the analysis. In this sense, the unidimensionality hypothesis was examined in Mplus 8 program using exploratory factor analysis (EFA), and its fitness values are shown in Table 2.

Table 2. Exploratory Factor Analysis Fit Indices

	χ^2	*df*	CFI	TLI	RMSEA	SRMR
Factor	1731.145*	170	.92	.91	.04	.03

Note: χ^2=x-square goodness of fit; *df*=degrees of freedom; CFI=Comparative Fit Index; TLI=Tucker Lewis Index; RMSEA=Root-Mean-Square Error of Approximation; SRMR=Standardized Square Root Mean Residual; AIC=Akaike Information Criterion*(p< 0.001).

As seen in Table 2, the TEOG Mathematics subtest had a unidimensional factor structure. Eigenvalues graph drawn as a result of EFA is shown in Figure 1.

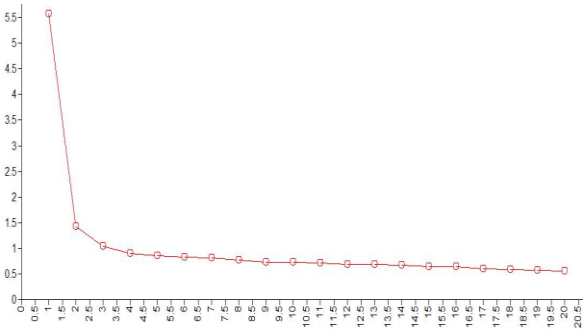

Figure 1. **Eigenvalues graph as a result of EFA**

As is seen in Figure 1, there was only one factor where the eigenvalue was greater than 1. A sharp drop, also seen in the graph, proves that the Mathematics subtest is unidimensional. To determine the validity of the unidimensional structure of the Mathematics subtest, confirmatory factor analysis was applied. The results are as follows: [$\chi 2$=1577.492*, *sd*=170, $\chi 2/df$=9.27, *RMSEA*=0.04, *CFI=0.*97, *TLI=0.*96]. Calculated goodness of fit values revealed that the unidimensional structure of the Mathematics subtest was valid for this research (Cole, 1987; Kline, 2005).

Q3 statistics were calculated to test the local independence hypothesis. Q3 statistics for the item pair formed for 20 items got values lower than 0.20 critical values ($Q3_{min}$= -0.13, $Q3_{max}$= 0.11, DeMars, 2010, p.50; De Ayala, 2009, p.134). These results prove that the items are statistically independent and the local independence hypothesis is met. Item parameters estimated according to the 3 and 4PL are shown in Table 3.

Table 3. Item Parameters Values Estimated According to the 3 and 4PL

	3PL			4PL			
Item	a	b	c	a	b	c	d
1	1.63	2.24	0.32	1.47	3.03	0.33	0.99
2	1.99	-1.17	0.21	1.55	-1.42	0.25	0.87
3	1.61	1.83	0.25	1.12	2.49	0.37	0.99
4	1.38	-0.37	0.08	1.75	-0.78	0.22	0.81
5	1.55	0.42	0.29	1.27	0.62	0.34	0.95
6	2.06	-2.36	0.36	1.43	-2.59	0.36	0.74
7	0.31	-1.14	0.04	0.29	-0.99	0.01	0.99
8	1.19	-2.06	0.23	1.45	-2.07	0.26	0.79
9	2.76	0.01	0.22	1.16	0.05	0.23	0.98
10	1.01	-2.68	0.28	1.53	-2.46	0.29	0.88
11	1.16	0.67	0.02	1.51	0.71	0.30	0.89
12	1.75	-0.01	0.18	1.84	-0.23	0.30	0.89
13	1.40	-2.79	0.19	1.60	-2.87	0.20	1.00
14	1.20	-2.97	0.24	1.75	-2.01	0.25	0.99
15	1.85	-1.73	0.18	1.81	-2.24	0.21	0.88
16	-0.51	-1.29	0.00	-0.47	-1.26	0.00	1.00
17	1.79	-2.29	0.14	0.33	-1.27	0.14	0.52
18	1.79	-1.54	0.30	1.80	-2.72	0.34	0.80
19	0.90	-2.85	0.18	1.17	-2.23	0.19	0.99
20	1.88	-1.56	0.15	1.84	-2.57	0.16	1.00

As seen in Table 3, evaluating the item parameters estimated according to the 3PL, it is seen that *a* parameter varied from -0.51 to 2.76, *b* parameter varied from -2.97 to 2.24, and *c* parameter varied from 0.00 to 0.36. According to the 4PL model, the item discrimination parameter varied from -0.47 to 1.84, difficulty parameter varied from -2.87 to 3.03, pseudo-chance parameter varied from 0.00 to 0.37, and *d* parameter varied from 0.52 to 1.00. It was seen that most of the items estimated according to both models had high discrimination and most of the items got values different than zero when *c* parameter values are considered. Although, item 16 must be excluded from the test since it had negative item discrimination value for two models. It was seen that item difficulty parameter estimations according to the 4PL model were lower than those estimated according to the 3PL model. *d* parameter estimations lower than 1.00 indicate the extent to which the students at a high ability level answered that item wrong. According to Table 3, it was seen that *d* parameter got values to differ than 1.00. The item with the highest wrong response possibility due to inattention was the 17th item with the value of di=0.52.

As part of the analysis of research data, ability estimations and their standard error values were obtained first according to the estimation methods. Subsequently, the amount of information and test information functions at each ability point were calculated and marginal reliability coefficients for each estimation method were obtained. Analyses were carried out using the Multidimensional Item Response Theory (MIRT) package in R studio program (Chalmers, 2013). Furthermore, SPSS 20 package program was utilized to test the differences between the estimation methods. Significance tests were carried out at the 0.001 level.

Findings

Findings on the Ability Estimations and Standard Error Values

Firstly, descriptive statistics of the research variables were calculated. In this sense, mean, minimum, maximum, skewness, and kurtosis coefficient values were calculated for the ability estimations made according to the 3 and 4PL models and their standard error values. The results are given in Table 4.

Table 4. Results of Descriptive Statistics to the Research Variables

	3PL						4PL					
	Ability Estimation			Standard Error			Ability Estimation			Standard Error		
	EAP	WLE	MAP	EAP	WLE	MAP	EAP	WLE	MAP	EAP	WLE	MAP
$\bar{x}$	0.08	0.10	0.08	0.45	0.47	0.41	0.00	0.13	0.08	0.45	0.33	0.34
Min	-1.85	-3.17	-1.72	0.31	0.27	0.28	-1.63	-1.66	-1.46	0.21	0.14	0.16
Max	2.16	2.49	2.08	0.73	6.17	0.81	2.03	3.73	1.75	1.19	1.85	1.65

As seen in Table 4, the estimation method having the highest average value in estimations made according to the 3 and 4PL models was WLE (0.10, 0.13). Similarly, for the 3PL model, the standard error value average for the ability estimation was calculated based on the highest WLE estimation method ($\bar{x}$=0.47). For the 4PL model, the standard error value average for the highest ability estimation was calculated according to the EAP estimation method ($\bar{x}$=0.45). It was seen that ranges of the ability estimations made according to the 3 and 4PL models were close to each other in terms of each ability estimation model.

Findings on the Fitness of the 3 and 4PL Models to Data

For the 3 and 4PL models, to find out which one was more compatible with the data, the models were examined using paired comparison with the calculated -2loglik, AIC, BIC, and RMSEA values. The results are given in Table 5.

Table 5. Model-Data Fit Comparison for the 3 and 4PL Models

Models	-2loglik	df	AIC	BIC	RMSEA
3PL	-45561.86	60	91243.71	91621.36	0.0001
4PL	-45434.32	80	91028.63	91532.16	0.0001

As seen in Table 5, -2loglik, AIC, and BIC values calculated for the 4PL model were lower compared to those for the 3PL model. This indicated that the 4PL model fit better than the 3 PL model. The difference between the 4PL and the 3 PL models is evaluated ($X^2_{(20)}$=127.54, p<.05). So, the 4PL model fits better than the 3PL model.

Findings on the Accuracy of the Ability Estimations Values

In the 3 and 4PL models, it was examined whether the accuracy of the ability estimations made according to EAP, WLE and MAP estimation methods differed significantly. Examining the descriptive statistics values given in Table 3 by taking into consideration that the range of the study group was substantially wide, it was seen that the ability estimation and standard error values to the scores obtained from the data set showed normal distributions. The results

obtained from the two-way analysis of variance (ANOVA) for iterative measurements are given in Table 6.

Table 6. Results of Two-Way Analysis of Variance (ANOVA) to the Ability Estimations Made According to the 3 and 4PL Model

Source	Sum of squares	df	Mean square	F
Between Groups	18068.43	7999		
Models	0.282	1	0.28	0.12
Error	18068.15	7998	2.25	
Within Groups	360.37	16000		
Ability (WLE-MAP-EAP)	60.52	2	30.26	1621.26*
Ability*Model	1.28	2	0.64	34.42*
Error	298.56	15996	0.01	

*p < 0.001

As seen in Table 6, it was found that estimation of the responses of the individuals analyzed according to two different IRT models by different ability estimation methods showed significant differences [F(2, 15996)=34.42, p<0.001]. This indicated that factors of different ability estimation methods had significant mutual effects on individuals' ability scores when the estimation was made according to different IRT models. Accordingly, using different IRT models had different effects on obtaining individuals' ability scores. Another finding indicated that there was a significant difference [F(2, 15996)=1621.26, p<0.001] between the average scores as a result of different ability estimation methods applied to the individuals analyzed according to the 3 and 4PL models. In other words, it can be argued that there was a significant change at the ability estimation level according to EAP, WLE, and MAP ability estimation methods. This means that individuals' ability estimations varied based on the applied estimation methods (EAP, WLE, and MAP) unless IRT models are distinguished. Bonferroni Test -one of the multiple comparison tests in statistics- was applied to determine which ability estimation methods had differences between each other. Evaluating the average scores of individuals' abilities estimated according to the ability estimation methods, it was found that all estimation methods were statistically different from each other. According to the results of this test, evaluating the average scores of the individuals according to the ability estimation methods, it was seen that WLE ability estimation method ($\bar{x}$=0.10) according to the 3PL model was higher than the averages of ability estimations made according to EAP ability estimation ($\bar{x}$=0.00) and MAP ability estimation ($\bar{x}$=0.08) methods. For the highest ability estimation value average made according to the 4PL model, it was seen that the WLE ability estimation method ($\bar{x}$=0.13) is higher than the averages of ability estimations made according to EAP ability estimation ($\bar{x}$=0.002) and MAP ability estimation ($\bar{x}$=0.08) methods. Furthermore, it was found that the model variable had no significant effect on the ability estimation scores [F(1, 7998)=0.12, p>0.001]. According to this finding, the ability estimations made according to the 3 or 4PL model showed that there were no significant changes in ability estimation scores of individuals.

Findings on the Accuracy of Ability Estimations and Their Standard Error Values

Two-way analysis of variance (ANOVA) was applied to determine the differences between the standard error values of MAP, EAP, and WLE ability estimation methods according to the 3 and 4PL models. The obtained results are given in Table 7.

Table 7. Two-Way Analysis of Variance (ANOVA) Results of the Standard Error Values of the Ability Estimations Made According to the 3 and 4PL Model

Source	Sum of squares	df	Mean square	F
Between Groups	18098.16	7999		
Models	30.022	1	30.02	568.09*
Error	18068.14	7998	2.25	
Within Groups	265.72	16000		
Ability (WLE-MAP-EAP)	22.98	2	11.49	833.02*
Ability*Model	22.08	2	11.04	800.33*
Error	220.66	15996	.01	

*p<0.001

As seen in Table 7, it can be argued that ability estimation models, ability estimation methods, and ability estimation model-ability estimation method interaction had significant effects on the standard error values of ability estimations [$F(1, 23994)_{model}=1121.17$, $p<0.001$; $F(2, 23994)_{estimation\ method}=429.27$, $p<0.001$; $F(2, 23994)_{model\text{-}estimation\ method}=412.00$, $p<0.001$]. Bonferroni Test -one of the multiple comparison tests in statistics- was applied to determine the differences between the ability estimation models, ability estimation methods, and ability estimation model-ability estimation method interaction. Evaluating the average scores of individuals' abilities estimated according to the ability estimation methods, it was found that all estimation methods were statistically different from each other. According to the results of this test, evaluating the average scores of the standard errors of the abilities of individuals estimated according to the ability estimation models, it was seen that the standard error value ($\bar{x}$=0.44) of the ability estimation made according to the 3PL was higher than the standard error value ($\bar{x}$=0.37) of the ability estimation made according to the 4PL model. Secondly, evaluating the ability estimation methods affecting the standard error values of ability estimations, it was seen that Finally, evaluating the effect of the ability estimation model-ability estimation methods interaction on the standard error values of ability estimations, for the 3PL model, the highest ability estimation value average was obtained by WLE ability estimation ($\bar{x}$=0.47) while the lowest ability estimation value average was obtained by MAP ability estimation method ($\bar{x}$=0.41). Secondly, evaluating the ability estimation methods affecting the standard error values of ability estimations, it was seen that the standard error values obtained by the EPA ability estimation method ($\bar{x}$=0.45) was higher than those obtained by WLE ability estimation method ($\bar{x}$=0.40) and MAP ability estimation method ($\bar{x}$=0.37). For the 4PL model, the highest ability estimation value average was obtained by the EAP ability estimation method ($\bar{x}$=0.45) while the lowest ability estimation value average was obtained by the WLE ability estimation method ($\bar{x}$=0.33). Marginal reliability coefficients calculated based on the estimation methods according to the 3 and 4PL models are given in Table 8.

Examining Table 8, it was seen that, for the 3PL model, the highest and lowest marginal reliability coefficient values were calculated with the ability scores estimated by MAP and WLE, respectively, whereas, for the 4PL model, the highest and lowest marginal reliability coefficient values were calculated with the ability scores estimated by WLE and EAP, respectively. For the estimations made according to IRT models, marginal reliability coefficients of ability scores estimated by MAP and EAP estimation methods were very close to each other.

Table 8. Findings on the Marginal Reliability Coefficients for the Ability Estimation Methods

Ability Estimation Methods	Marginal Reliability Coefficient	
	3PL	4PL
MAP	0.79	0.80
EAP	0.78	0.78
WLE	0.75	0.84

Discussion

In this research, based on data consisting of the answers of 4000 students to the Mathematics subtest of TEOG exam in 2015-2016 school year, which of the 3 and 4PL models the data was more compatible with, MAP, EAP, and WLE estimation methods under the 3 and 4PL models, the ability estimations, standard error values of the ability estimations and their marginal reliability coefficients were analyzed.

Model-data fit was compared by -2loglik, AIC, BIC, and RMSEA methods. According to the comparisons, three of these methods (2loglik, AIC, and BIC) indicated that the 4PL model fit better than the 3PL model. The same value was calculated for the 3 and 4PL models according to the RMSEA method. This finding was in line with those reported in the previous studies. Loken and Rulison (2010) also carried out parameter estimation utilizing the 4PL model and found that the 4PL model fit better than the 3PL model. Similarly, Erdemir (2015) has reported that the best-fit model was the 4PL model in terms of model-data fit. However, unlike this result, Barton and Lord (1981) and Yalçın (2018) suggested that the 3PL model fit better than the 4PL model. Barton and Lord (1981) discussed that this was since *d* parameter cannot be estimated freely and therefore, it was calculated by fixing one *d* parameter estimation for all items. Furthermore, Yalçın (2018) carried out parameter estimations by a different model type, the MixIRT model.

The research also analyzed the accuracy of individuals' ability estimations in consideration of scores obtained by MAP, EAP, and WLE ability estimation methods by the 3 and 4PL models. The results showed that the accuracy of ability estimation scores was significantly different based on the estimation methods. This difference indicated that the accuracy of scores obtained by the WLE ability estimation method was higher than those obtained by the MAP ability estimation method while the accuracy of scores obtained by the MAP ability estimation method was higher than those obtained by the EAP ability estimation method. There are contradictory findings in the relevant literature. Çetin and Çelikten (2016) have reported that the methods making the most accurate estimations were MAP, EAP, WLE, and ML estimation methods, respectively. The present study and the cited study showed that MAP made more accurate estimations than the EAP estimation method. This finding was also supported by various studies (Wang & Vispoel, 1998; Wang & Wang, 2001; Finch & French, 2012; Seong, Kim & Cohen, 1997). On the other hand, Borgatto, et al., (2015) have reported that the WLE method gave the best results for the estimation of abilities of the individuals at a high ability level for low-difficulty tests. According to the findings of the item analysis performed for the TEOG exam used in the present research, the item difficulty values were at a

low level. This finding was parallel with the findings by Wang and Wang (2001) who argued that the WLE method made estimations with lower bias compared to EAP and MAP estimation methods for fixed-length tests based on CAT application.

Within the scope of the present research, it was found that IRT models, ability estimation methods and ability estimation model-ability estimation method interaction had a significant effect on the standard error values of ability estimations. In this sense, evaluating the standard error values of ability estimations, it was found that the standard error value of the ability estimation made according to the 3PL was higher than the standard error value of the ability estimation made according to the 4PL model. In other words, in consideration of ability estimation standard error values, the lowest standard error value was obtained by the ability estimation made according to the 4PL model. This finding was in line with those obtained in similar studies (Liao, Ho, Yen, & Cheng, 2012; Rulison & Loken, 2009; Yen, Ho, Liao, & Chen, 2012; Yen, et al., 2012). For instance, when Erdemir (2015) used the 4PL model instead of the 3PL model, the standard error value of the ability became lower. Accordingly, it can be inferred that the accuracy of the estimation increased. Another finding from the study was that the most accurate ability estimation on the standard error values of estimations was the score points obtained by EAP, WLE, and MAP estimation methods, respectively. This finding supports the view that the systematic error of the MAP estimation method was higher than the systematic error of the EAP estimation method (Çetin & Çelikten, 2016). In this sense, it can be inferred that the accuracy of estimation increased as the ability range increased. Moreover, according to the ability estimation model-ability estimation methods interaction affecting the standard error values of the ability estimations analyzed as part of this research, the highest ability estimation value average according to the 3PL model was obtained by WLE estimation while the highest ability estimation value average according to the 4PL model was obtained by EAP ability estimation.

Lastly, the highest marginal reliability coefficient value according to the 3PL model was calculated with the ability values obtained by the MAP estimation method while the highest marginal reliability coefficient value according to the 4PL model was calculated with the ability values obtained using WLE estimation method. In this sense, marginal reliability coefficients showed similarity with the order of accuracy of standard errors of the ability estimations. This might be caused by the mean reversion of the marginal reliability coefficient.

Conclusion and Implication

In this research, based on answers of 8th-grade students taking the TEOG exam in the 2015-2016 school year to 20 items under the Mathematics subtest of the TEOG exam, IRT based model-data fit, ability estimations, and their standard values, and marginal reliability coefficient of the test were analyzed. In an overall evaluation of the findings were evaluated, it was found that the 4PL model fit better, and the standard error value of WLE and MAP ability estimation models were low according to the 4 and 3PL model, respectively. Furthermore, in this research, it was observed that the reliability coefficient obtained based on these estimation methods under both ability estimation models was higher.

Individuals' estimated ability scores were used in the evaluation stage of large-scale exams such as TEOG which is very important for determining success and competence and

performing selection and placement. Accordingly, it can be suggested that calculating these scores according to the 4PL model and by the WLE ability estimation method may provide more accurate results. Carrying out similar research for large-scale exams held as of 2016 may contribute to the precision of results. Moreover, EAP, MAP, and WLE ability estimation methods were analyzed as part of this research. Research results may be expanded by testing other types of Bayesian methods.

References

Baker, F. B. (1992). *Item Response Theory: Parameter Estimation Technique*. New York: Marcel Dekker.

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple-choice tests: A case study in irrationality. *Mind & Society, 4*, 3-12. http://doi.org/cp7ddc

Barton, M. A., & Lord, F. M. (1981). An upper asymptote for the three-parameter logistic item-response model. *Research Bulletin*, 81-20. Princeton, NJ: Educational Testing Service.

Baykul, Y. (1979). *Örtük özellikler ve klasik test kuramları üzerine bir karşılaştırma* (Unpublished Doctoral thesis). Hacettepe University, Graduate School of Social Sciences, Ankara.

Berberoğlu, G. (1988). *Seçme amacıyla kullanılan testlerde Rasch modelinin katkıları* (Unpublished Doctoral thesis). Hacettepe University, Graduate School of Social Sciences, Ankara.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. F. M. Lord & M. R. Novick (Ed), In *Statistical theories of mental test scores* (pp. 397-472). Reading MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 433-459.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in microcomputer environment. *Applied Pyschological Measurement, 6*, 431-444. Doi: 10.1177/014662168200600405.

Borgatto, A. F., Azevedo, C. L. N., Pinheiro, A., & Andrade, D. F. (2015). Comparison of ability estimation methods using irt for test with different degrees of difficulty. *Communications in Statistics-Simulation and Computation*, *44*(2), 474-488.

Ching-Fung, B. S. (2002). *Ability estimation under different item parametrization and scoring models* (Unpublished Doctoral thesis). North Teksas University, Teksas.

Can, S. (2003). *The analyses of secondary education institutions student selection and placement test's verbal section with respect to item response theory models* (Unpublished Master's thesis). Middle East Technical University, Graduate School of Social Sciences, Ankara.

Chalmers R. P. (2013). mirt: Multidimensional Item Response Theory. R package version 0.9.0, [Çevirim içi: http://CRAN.R-project.org/package=mirt].

Cole, D. A. (1987). Utility of confirmatory factor analysis in test validation research. *Journal of Consulting and Clinical Psychology, 55*, 584-594.

Çelik, D. (2001). *The Fit of one, two and three-parameter models of item response theory (IRT) to the ministry of National Education secondary school institutions student selection and placement test data* (Unpublished Master's thesis). Middle East Technical University, Graduate School of Social Sciences, Ankara.

Çetin, B. ve Çelikten, S. (2016). Nominal response model altında yetenek kestirim yöntemlerinin karşılaştırılması. *International Engineering, Science and Education Conference*, 01-03 December 2016, Diyarbakır.

DeMars, C. (2010). *Item response theory.* New York: Oxford University Press.

De Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. U. S. A.

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

Erdemir, A. (2015). *Bir, iki, üç ve dört parametreli lojistik madde tepki kuramı modellerinin karşılaştırılması (Comparison of 1PL, 2PL, 3PL and 4PL item response theory models)* (Unpublished Master's thesis). Gazi University, Graduate School of Educational Sciences, Ankara.

Finch, W. H., & French, B., F. (2012). Parameter Estimation with Mixture Item Response Theory Models: A Monte Carlo Comparison of Maximum Likelihood and Bayesian Methods. *Journal of Modern Applied Statistical Methods*, *11*(1), Article 14. DOI: 10.22237/jmasm/1335845580.

Gardner-Medwin, A. R., & Gahan, M. (2003). Formative and summative confidence-based assessment. In J. Christie (Ed.), *Proceedings of the 7th International Computer-Aided Assessment Conference* (pp.147-155). Loughborough, UK: Loughborough University.

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*. Newbury Park, CA: Sage.

Hockemeyer, C. (2002). A comparison of non-deterministic procedures for the adaptive assessment of knowledge. *Psychologische Beiträge*, *44*, 495-503.

Kılıç, İ. (1999). *The fit of one- two- and three- parameter models of item response theory to the student selection test of the student selection and placement center* (Unpublished Doctoral thesis). Middle East Technical University, Graduate School of Social Sciences, Ankara.

Kline, R. B. (2005). *Principles and practices of structural equation modeling.* New York: The Guildord.

Liao, W., Ho, R., Yen, Y. & Cheng,, H. C.(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.

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*, 509-525.

Magic, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, 37(4) 304–315.

Önder, İ. (2007). An investigation of goodness of model data fit. *Hacettepe University Journal of Education*, *32*, 210-220.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods 8*(2), 164-184.

Reynolds, T., Perkins, K., & Brutten, S. (1994). A comparative item analysis study of a language testing instrument. *Language Testing*, *11*, 1-14.

Rose, N. (2010). *Maximum likelihood and Bayes modal ability estimation in two-parametric IRT models: Derivations and implementation*. Jena, Germany: Schriften zur Bildungsf. Retrieved September 9, 2017, from https://www.kompetenztest.de/downloads/schriften.

Rulison, K., & 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*, 83-101. http://doi.org/dtqjq8

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97.

Seong, T. J., Kim, S. H., & Cohen, A. S. (1997). A comparison of procedures for ability estimation under the graded response model. *Paper presented at the Annual Meeting of the American Educational Research Association*, Chicago.

Taşdelen Teker, G., Kelecioğlu, H. ve Eroğlu, M. G. (2013). An investigation of goodness of model data fit. 4. *International Conference on New Horizons in Education*, June, 25-27, 2013, Roma, Italia.

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, *12*, 339-368.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Education Measurement*, *35*, 109-135.

Wang, S., & Wang, T. (2001). Precision of Warm's weighted likelihood estimates for a polytomous model in computarized adaptive testing. *Applied Psychological Measurement, 25,* 317-331.

Warm, T. A. (1989).Weighted likelihood estimation of ability in item response theory. *Psychometrika* 54, 427-450.

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 DOI:10.21449/ijate.402806.

Yapar, T. (2003). *A study of the predictive validity of the Başkent a study of the predictive validity of the Başkent University English proficiency exam through the use of the two-parameter irt model's abiliıty estimates* (Unpublished Master's thesis). Middle East Technical University, Graduate School of Social Sciences, Ankara.

Yeğin, O. P. (2003). *The predictive validity of Başkent University proficiency exam (buepe) through the use of the three-parameter irt model's ability estimates* (Unpublished Master's thesis). Middle East Technical University, Graduate School of Social Sciences, Ankara.

Yen, Y.C., Ho, R.G., Chen, L.J., Chou, K.Y., & Chen, Y. L. (2010). Development and evaluation of a confidence-weighting computerized adaptive testing. *Educational Technology & Society*, *13*, 163-176.

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.

Yen, Y.C., Ho, R.G., Liao, W.W., Chen, L.J., & Kuo, C.-C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement*, *36*(2), 75-87. doi:10.1177/0146621611432862

Zwinderman, A. H., & van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the rasch model. *Applied Psychological Measurement, 14*(1), 73-81.

Genişletilmiş Özet

Giriş

Son yıllarda Madde Tepki Kuramına (MTK) göre gerçekleştirilen test maddelerinin niteliğinin araştırıldığı çalışmaların sonuçları, ek madde parametreleri kullanılmasının bireyleri karakterize eden parametrelerin kestirimlerinin doğruluğunu ve kesinliğini artırdığını ortaya koymaktadır. Örneğin 3PL modelin 1993 yılına ait Öğrenci Seçme Sınavı (ÖSS) verilerine uyumunun diğer modellere göre daha iyi olduğu saptanmıştır (Kılıç, 1999). Bu çalışmalar ve benzerleri yalnızca 1, 2 ve 3PL model altında madde parametrelerini kestirerek, benzer çıkarımlarda bulunmuşlardır (Barton & Lord, 1981; Baykul, 1979; Berberoğlu, 1988; Can, 2003; Kılıç, 1999; Reise & Waller, 2003; Yapar, 2003; Yeğin, 2003). Bu üç modelin yanı sıra 4PL modeli de MTK modeli olarak literatürde yer almaktadır. Bir başka deyişle, kişinin yetenek düzeyi ve madde niteliklerinin kişinin maddeye verdiği yanıtlarla olan ilişkisi olarak tanımlanan 'MTK Modelleri' 1, 2, 3 ve 4PL modellerden oluşmaktadır.

Alanyazında 1, 2 ve 3PL modellerin kestirilmesi ve model uyumunun test edilmesi konusunda yapılan çalışmalarda en iyi uyum sergileyen modelin 3PL model olduğu görüşleri bulunmaktadır. Bu kapsamdaki görüşlerden birisi Çelik'in (2001) 1, 2 ve 3PL modelin Milli Eğitim Bakanlığı ortaöğretim kurumları öğrenci seçme ve yerleştirme sınavında uygulanan matematik ve fen bilgisi alt testlerinden elde edilen verilere uyum düzeylerini incelediği çalışmadır. Bir diğeri de Önder'in (2007) Özdebir ÖSS 2004 yılında uygulanan D-II sınavının Fen Testinden elde edilen veriye MTK'ya dayalı modellerden hangisinin en iyi uyum sergilediğini incelediği çalışmadır. Bununla birlikte MTK çerçevesinde ikili puanlanan maddeler için 4PL modelin de araştırma kapsamına dahil edildiği oldukça az çalışmaya rastlanmıştır. Bu çalışmalardan birisinde 1, 2, 3 ve 4PL modele göre kestirilen madde ve yetenek parametreleri karşılaştırılmıştır. Çalışmanın sonucunda 4PL model altında yapılan kestirimin, diğer üç modelden daha düşük standart hataya sahip olduğu ve yetenek parametresinin bu model altında daha doğru kestirildiği bulgusuna ulaşılmıştır (Magic, 2013). BBT uygulamalarının gerçekleştirildiği çalışmalarda daha düşük standart hata değeri elde edildiği için 4PL modele göre yetenek kestirimi yapılması önerilmektedir (Rulison & Loken, 2009; Yen ve ark., 2012).

Çalışmanın Amacı

Bu araştırmanın problemi, iki kategorili MTK modelleri altında yetenek kestirim yöntemlerinin karşılaştırılmasıdır. Araştırmanın gerekçesi ise, yetenek kestirimlerini düşük yetenek düzeyindeki yanıtlayıcıların maddeyi doğru yanıtlama olasılıklarını dikkate alan 3PL modele ve yüksek yetenek düzeyindeki yanıtlayıcıların kolay bir maddeyi dikkatsizlik nedeniyle yanlış yanıtlama olasılıklarını da dikkate alan 4PL modele göre kestirmektir. Böylece bu araştırmayla gerçek veriye dayalı olarak en uygun MTK modeli türü ve yetenek kestirim yönteminin belirlenmesi amaçlanmıştır. Bu genel amaçla uyumlu olarak araştırmanın alt-problemleri aşağıda sunulmuştur. 3 ve 4PL modele göre kestirilen EAP, WLE ve MAP yetenek kestirim yöntemlerinden elde edilen;

1. Yetenek kestirim modelleri ve yöntemlerine göre kestirilen yetenek kestirimleri ve yetenek kestirimlerine ait standart hata değerleri nasıldır?
2. 3 ve 4PL yetenek kestirim modellerinden hangisi veriye daha fazla uyum sergilemektedir?
3. Yetenek kestirimlerinin doğruluğu kestirim modelleri ve yöntemlerine göre anlamlı farklılık göstermekte midir?
4. Yetenek kestirimlerine ait standart hataların doğruluğu kestirim modelleri ve yöntemlerine göre anlamlı farklılık göstermekte midir?
5. Marjinal güvenirlik katsayıları farklılaşmakta mıdır?

Yöntem

Bu araştırma 3 ve 4PL modele dayalı olarak model veri uyumu, madde parametre kestirimlerinin doğruluğunu karşılaştırmalı olarak inceleyen betimsel bir araştırmadır.
2015-2016 eğitim öğretim yılında uygulanan temel eğitimden orta öğretime geçiş için yürütülen ulusal geçiş sınavının (TEOG) Matematik alt testinden elde edilen veriye dayalı olarak gerçekleştirilmiştir. Milli Eğitim Bakanlığı Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü tarafından bu araştırma için alınan kayıp veri ve tam puan içermeyen seçkisiz olarak seçilen 4000 kişilik çalışma grubu üzerinden analizler gerçekleştirilmiştir.

Bulgular

Veri analizinde ilk olarak araştırma değişkenlerine ait betimsel istatistikler hesaplanmıştır. Bu sonuçlara göre 3 ve 4PL modele göre yapılan kestirimlerde en yüksek ortalama değere sahip kestirim yöntemi WLE (0.10, 0.13)'dir. Ayrıca, 4PL modelin 3PL modele göre daha iyi uyum gösterdiği sonucuna ulaşılmıştır.

3 ve 4PL modelde EAP, WLE ve MAP kestirim yöntemlerine göre yapılan yetenek kestirimlerinin doğruluğunun anlamlı olarak birbirinden farklılık gösterdiği bulunmuştur. Bireylerin yetenek kestirim modellerine göre ortalama puanları incelendiğinde, 3PL modele göre yapılan WLE yetenek kestirim yönteminin ($\bar{x}$=0.10), EAP yetenek kestirim ($\bar{x}$=0.00) ve MAP yetenek kestirim ($\bar{x}$=0.08) yöntemlerine göre gerçekleştirilen yetenek kestirimleri ortalamalarından daha yüksek olduğu görülmektedir. 4PL modele göre ise en yüksek yetenek kestirim değeri ortalaması WLE yetenek kestirimine ($\bar{x}$=0.13), EAP yetenek kestirim ($\bar{x}$=0.002) ve MAP yetenek kestirim ($\bar{x}$=0.08) yöntemlerine göre gerçekleştirilen yetenek kestirimleri ortalamalarından daha yüksek olduğu görülmektedir.

Araştırmanın bir diğer bulgusu ise yetenek kestirim modelleri, yetenek kestirim yöntemleri ve yetenek kestirim modeli-yetenek kestirim yöntemi etkileşiminin yetenek kestirimlerinin standart hata değerleri üzerinde anlamlı bir etkiye sahip olmasıdır. 3PL modele göre yapılan yetenek kestiriminin standart hata değerinin ($\bar{x}$=0.44) 4PL modele göre yapılan yetenek kestiriminin standart hata değerinden ($\bar{x}$=0.37) daha yüksek olduğu gözlenmiştir. Yetenek kestirim modeli-yetenek kestirim yöntemleri etkileşiminin yetenek kestirimlerinin standart hata değerleri üzerindeki etkisi incelendiğinde de 3PL modele göre en yüksek yetenek kestirim değeri ortalaması WLE yetenek kestirimine ($\bar{x}$=0.47), en düşük yetenek kestirim değeri ortalaması MAP yetenek kestirim yöntemine ($\bar{x}$=0.41) göre geçekleştirildiğinde hesaplanmıştır.

Son olarak 3PL modele göre en yüksek marjinal güvenirlik katsayı değerinin MAP, en düşük marjinal güvenirlik katsayı değerinin WLE; 4PL modele göre en yüksek marjinal güvenirlik katsayı değerinin WLE, en düşük marjinal güvenirlik katsayı değerinin EAP kestirim yöntemine göre kestirilen yetenek puanlarından hesaplandığı gözlenmektedir.

Sonuç ve Öneriler

Araştırmada 2015-2016 yılında TEOG sınavına katılan 8. Sınıf öğrenciler TEOG sınavının matematik alt testinde yer alan 20 maddeye verdikleri yanıtlara göre gerçekleştirilen MTK'ya dayalı model-veri uyumu, yetenek kestirimleri ve yetenek kestirimlerine ait standart değerleri, testin marjinal güvenirlik katsayısı incelenmiştir. Bulgular bir bütün olarak değerlendirildiğinde 4PL modelin daha iyi uyum gösterdiği, 4PL modele göre WLE, 3PL modele göre MAP yetenek kestirim modelinin standart hata değerinin de düşük olduğu sonucuna ulaşılmıştır. Başarı veya yeterliliklerinin belirlenmesi, seçme ve yerleştirmelerin yapılması açısından oldukça önemli olan TEOG gibi geniş ölçekli sınavların değerlendirilmesinde yetenek puanlarının 4PL modele ve WLE yetenek kestirim yöntemine göre hesaplanmasının daha doğru sonuçlar üretebileceği söylenebilir. 2016 yılından itibaren yapılan geniş ölçekli sınavlar için de benzer araştırmanın yürütülmesi sonuçların kesinliğine katkı sağlayabilir. Ayrıca, araştırma kapsamında yetenek kestirim yöntemlerinden EAP, MAP ve WLE yöntemleri incelenmiştir. Bayesian yöntemlerinin diğer türleri de sınanarak araştırma sonuçları genişletilebilir.