International Journal of Assessment Tools in Education
2018, Vol. 5, No. 2, 301–313

DOI: 10.21449/ijate.402806

Published at http://www.ijate.net http://dergipark.gov.tr/ijate Research Article

Data Fit Comparison of Mixture Item Response Theory Models and Traditional Models

Seher Yalçın [1*]

[1] Ankara University, Faculty of Educational Sciences, Department of Measurement and Assessment, Ankara, Turkey

Abstract: The purpose of this study is to determine the best IRT model [Rasch, 2PL, 3PL, 4PL and mixed IRT (2 and 3PL)] for the science and technology subtest of the Transition from Basic Education to Secondary Education (TEOG) exam, which is carried out at national level, it is also aimed to predict the item parameters under the best model. This study is a basic research as it contributes to the information production which is fundamental for test development theories. The study group of the research is composed of 5000 students who were randomly selected from students who participated in TEOG exam in 2015. The analyses were carried out on 17 multiple choice items in TEOG science and technology subtest. When model fit indices were evaluated, the MixIRT model with two parameters and three latent classes was found to fit the data best. According to this model, when the difficulties and discrimination averages of the items are taken into account, it can be expressed that items are moderately difficult and discriminative for students in latent class-1; the items are considerably easy and able to slightly distinguish the students in latent class-2; the items are difficult to the students in the third latent class and they can slightly distinguish the students in this group.

ARTICLE HISTORY

Received: 10 February 2018
Revised: 24 March 2018
Accepted: 29 March 2018

KEYWORDS

Item response theory,
Latent class models,
Mixed item response theory,
Model-data fit

1. INTRODUCTION

The purpose in educational and psychological measurements is to ensure that the decisions made about the individual are valid and reliable. To this end, models and theories which try to better demonstrate the state of individual's having the measured characteristics are being developed. Within the scope of the models known as latent variable models; structural equation models, latent class models, latent profile models, and latent trait models (item response theory) are discussed (Skrondal & Rabe-Hesketh, 2007). Commonly used theories in the literature are: Classic Test Theory (CTT) and Item Response Theory (IRT). If the assumptions are met, IRT models are often preferred over CTT because CTT fails to provide as much information as IRT due to the limitations of the theory [e.g. individuals' ability levels depend on the item they receive, the item properties depend on the respondent group; it is difficult to compare individuals who take different tests and the need for parallel tests for

CONTACT: Seher Yalçın ✉ yalcins@ankara.edu.tr Ankara University, Faculty of Educational Sciences, Department of Measurement and Assessment, Ankara, Turkey

ISSN-e: 2148-7456 /© IJATE 2018

reliability prediction (Hambleton, Swaminathan & Rogers, 1991)]. Some of the reasons for preferring IRT models are; obtaining more reliable results thanks to error prediction on individual level, invariant item parameters across groups, making item independent ability predictions (De Ayala & Santiago, 2017; De-Mars, 2010; Embretson & Reise, 2000). The Item Response Theory (IRT) allows individuals' ability (θ) and item parameters to be predicted by associating the individual's response to the item with the individual's level of ability and item traits (Embretson & Reise, 2000). Since trait or ability cannot be measured directly, item response theory identifies the relationship between individuals' observed performances for items and the unobservable traits or abilities that are assumed to underlie this related performance (Hambleton & Swaminathan, 1985).

Predictions in IRT can be conducted by different models. IRT models are grouped as unidimensional and multidimensional models. The unidimensional models are composed of different models based on item scoring (dichotomous and polytomous items). Models used for dichotomous scoring items are; 1, 2-, 3-, 4- parameter logistic (PL) models. These models are named according to the number of item parameters used in the function which models the relationship between the item response and individual's ability (De Mars, 2010). The possibility of a correct response to the item j for 4PLM is given in Equation 1 (Barton & Lord, 1981):

$$P(\theta_j)=c_j+(d_j-c_j)\frac{e^{Daj(\theta-bj)}}{1+e^{Daj(\theta-bj)}} \qquad \text{(Equation 1)}$$

P(θj) is the correct response possibility to item j for a randomly selected individual at θ ability level. "c_j" is the correct response possibility by chance, while "d_j" is the possibility of high-ability individuals' responding wrong to an easy item due to the lack of attention. As a constant, value of *e* is 2.718 while *D* is usually taken as 1.7. Item discrimination parameter of item j is a_j; and b_j is the difficulty parameter of the item j. When "1" is written instead of the d_j parameter in Equation 1, the formula of 3PLM is obtained. In this formula, if the c_j parameter is taken as "0", the formula of 2PLM is obtained. In the formula of 2PLM, when the a_j parameter is taken as "same value for all items (i.e., usually with 1 at Rasch model)" and when D parameter is subtracted from the formula, the formula for 1PLM is reached.

The latent variable is assumed to be categorical in the latent class analysis (LCA), which is one of the latent class models, while there is a constant latent variable assumption in IRT (De Ayala, 2009). That is, when the observed variable is discontinuous and the latent variable is also discontinuous, LCA is used. LCA is utilized to generate homogeneous subclasses from heterogeneous latent traits (Vermunt & Magidson, 2002). In latent class analysis, it is accepted that all observed variables are the cause of a latent variable. If the latent variable is set as a control variable, the relationship between the observed variables is concluded to be conditionally independent. Under this condition, LCA is conducted to determine the latent variable which is also the control variable (Vermunt & Magidson, 2004).

The use of item response theory and latent class analysis combination brings Mixture item response theory (MixIRT) into light (Cohen & Bolt, 2005). MixIRT model is a powerful statistical method combining the LCA and IRT. Even though the concept of MixIRT has emerged with Rost in the 1980s, it is in the 2000s that it has begun to have a widespread use. The article, in which De Ayala and Santiago (2017) introduced the MixIRT and its applications, was published in 2017. It can be said that models based on MixIRT have become more widespread recently in the literature. MixIRT models (Kelderman & Macready, 1990; Maij-de Meij, Kelderman & van der Flier, 2010; Rost, 1990) have no assumptions about the type or cause of the qualitative differences in participants' responses. In the MixIRT models, latent

classes (homogeneous subgroups) are defined and different parameter estimates are made between the latent classes. The MixIRT model assumes that the population consists of a limited number of latent individual, and these classes can be differentiated based on item response patterns (von Davier & Rost, 2017). On the contrary, these different response patterns will indicate themselves as differences in the parameters of the item response model related to each group. The formula for two parameter MixIRT model is as follows (Finch & French, 2012):

$$P(U = 1|g,\ \theta_{ig}) = \frac{e^{(a_{jg}(\theta_{ig} - b_{jg}))}}{1+e^{(a_{jg}(\theta_{ig} - b_{jg}))}} \qquad \text{(Equation 2)}$$

In the formula, "g: 1, 2, ..., G" indicates the latent class membership, " b_{jg}" indicates the intraclass difficulty for the item j, "a_{jg}" shows the intraclass discrimination for the item j, and " θ_{ig}" shows the level of latent trait which is measured in classroom for the individual i.

When the literature is reviewed, many studies comparing the traditional models of IRT (Rasch, 1PL, 2PL, 3PL and 4PL) have been found (Barton & Lord, 1981; Can, 2003; Erdemir, 2015; Kılıç, 1999; Loken & Rulison, 2010; Waller & Reise, 2010). Some studies (Can, 2003; Erdemir, 2015; Kılıç, 1999) indicated that 3PL or 4PL models generally fit better to data. However, it is seen in the other studies (Barton & Lord, 1981; Loken & Rulison, 2010; Waller & Reise, 2010) they are generally conducted in the field of psychology, and 4PLM has fitted better to the data in the studies conducted in recent years. Upon looking at the studies conducted for the purpose of scaling with MixIRT models; it is observed that they are employed in various studies in different subject fields such as evaluating the cognitive abilities of students (De Ayala & Santiago, 2017), analysing individual differences according to the response categories they choose in multiple choice items (Bolt, Cohen, & Wollack, 2001), interpretation of response behaviours in personality questionnaires (Maij-de Meij, Kelderman & Van der Flier, 2008), analysis of tobacco dependence in a general population survey (Muthen & Asparouhov, 2006), and scaling of a conscience scale in the context of career development (Egberink, Meijer & Veldkamp, 2010).

This study is important as it provides an application example for an exam conducted at national level regarding the use of MixIRT models. In addition, the validity and reliability of the decisions made in the exams conducted at national level are also important. Different statistical models and theories have been developed to make the most accurate predictions about the individuals' scores. In this study, results according to MixIRT are presented by trying out these models and theories. The MixIRT models allow researchers to obtain more reliable, thus more valid information about the traits of the item and group by dividing the ability of students into latent classes.

1.1. Purpose of Research

The purpose of this study is to determine the best IRT model [Rasch, 2PL, 3PL, 4PL and mixed IRT (2 and 3PL)] for the science and technology subtest of the national transition examination which is conducted for transition from basic to secondary education. It is also aimed to predict the item parameters under the best model. In this context, the questions that are sought to be answered in the study are:

1. Which IRT model (Rasch, 2PL, 3PL, 4PL and MixIRT) do TEOG 2015 science and technology subtest items fit better to?
2. What are the item parameters based on the model that fits best to data?

2. METHOD

2.1. Research Model

This study is a basic research as it aims to determine the model which fits best to the data by trying different IRT models, in other words, it contributes to the production of information necessary for test development theories.

2.2. Study Group

The study group of the research is composed of 5000 students who were randomly selected from the students participated in the Transition from Basic Education to Secondary Education (TEOG) exam in 2015. When the students' gender distribution is examined, it is seen that 48.5% (N: 2425) of the students were female and 51.5% (N: 2575) of them were male. It can be expressed that the gender rates are rather close to each other.

2.3. Data Collection Tools

The data used in this study are obtained from the application that is carried out according to the curriculum which is taught in the lessons with centralized joint exams of six core curriculum (Turkish, Mathematics, Science and Technology, Religion and Ethics, History of Revolution and Kemalism, Foreign Language). It was applied at the end of the first semester in 2015 by the Ministry of National Education. The TEOG exams, which started to be implemented in 2013, gave its place to another exam in the 2017-2018 academic year. Science and Technology subtest data consisting of 20 multiple choice items were used in this study. Because an item (item 13) was cancelled, analyses were conducted on 19 items. The data was obtained with a written permission from the Ministry of National Education (MoNE) General Directorate of Measurement and Evaluation Examination Services with the request of the researcher.

2.4. Data Analysis Procedures

Before analyzing the data, the data missing rates of the items were analysed. It was observed that they varied between 0.1% and 0.2%. The state of having extreme value of the data is examined and no extreme value is encountered. In addition, the normality of the distribution was tested, the skewness coefficient was found to be .06, and the kurtosis coefficient was -1.00. The average score of students' science and technology scores were found to be 10.62 and the standard deviations 4.51. The histogram of students' science and technology scores was also examined and seen to be in line with the normal distribution assumptions. Then, the assumptions of IRT (unidimensionality, local independence, monotone increase of the item characteristics curve and whether the test is a speed test or not) were tested (Hambleton & Swaminathan, 1985).

Confirmatory Factor Analysis (CFA) was carried out with Mplus 8 program to examine whether the nature of data-meets unidimensionality assumptions. As a result of the analyses conducted for 19 items, two items, which are items 16 and 18, were subtracted from the analysis because their factor loading values were below .30, and the CFA analysis was repeated for 17 items. As a result of the analyses, factor loadings of all items are above .30 and are statistically significant. Table 1 demonstrates the results of the analysis. When the fit indices obtained from the unidimensional model are examined, it can be expressed that the data has a good level of fit to the model ($\chi^2_{(119)}$=504.198, p<.01, χ^2/sd=4.24; RMSEA: 0.025, CFI: 0.988, TLI: 0.987).

Table 1. Results of the analysis of the unidimensional model for science and technology subtest

Items	Estimate	Standard error	Estimate/standard error
i1	0.693	0.014	50.996*
i2	0.619	0.014	43.977*
i3	0.555	0.017	32.673*
i4	0.556	0.015	36.628*
i5	0.636	0.014	46.973*
i6	0.504	0.018	28.017*
i7	0.667	0.015	45.340*
i8	0.705	0.012	57.694*
i9	0.488	0.017	29.403*
i10	0.625	0.014	44.942*
i11	0.742	0.012	59.400*
i12	0.723	0.012	59.714*
i13	0.658	0.014	47.823*
i14	0.674	0.013	51.330*
i15	0.672	0.013	50.120*
i16	0.571	0.015	37.171*
i17	0.383	0.018	21.716*

*p< .05

As it can be seen in Table 1, the factor loadings of the items vary between .383 [item17 (i17)] and .742 (i11), and all items appear to make significant contribution to the unidimensional model.

Yen's Q3 statistics was used to examine whether the data validate the local independence assumptions. Although the local independence assumption is stated to be met as well in the case of the unidimensionality assumption (Hambleton & Swaminathan, 1985), the Q3 statistics which is frequently used in testing the local independence is also calculated. The calculations are carried out based on examining the correlations between items under the four different models (Rasch, 2PL, 3PL and 4PL). Q3 statistics are calculated for each model in R with the help of "sirt" package (Robitzsch, 2015). In all models, the correlations between the items were found to be -0.127 (the lowest) and 0.042 (the highest). It can be stated that the local independence assumption is met as the values calculated are less than .20 (DeMars, 2010). Item characteristic curves (ICCs), were examined for four models to see the monotonic increase of the item characteristic curve. The ICCs are drawn for each model in R with the help of the "sirt" package (Robitzsch, 2015), and are presented in Figures 1 and 2.

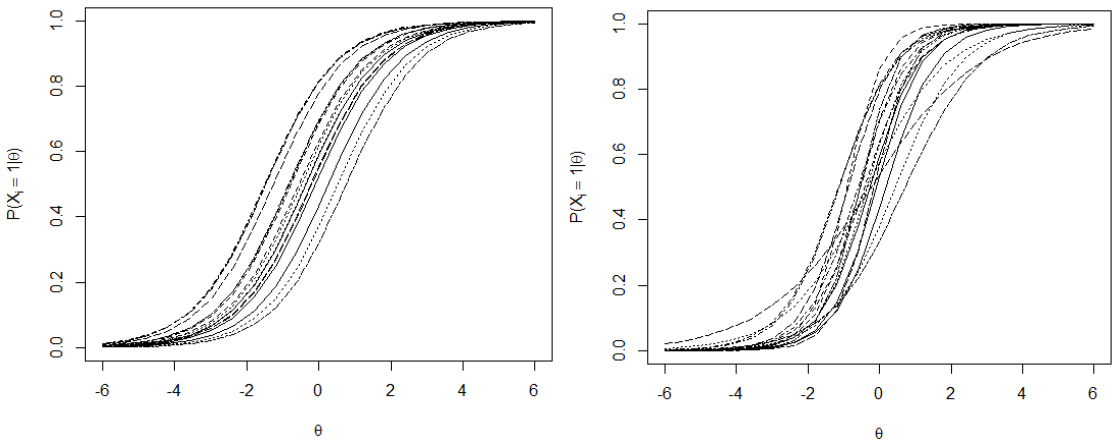


Figure 1. ICCs according to 1PL and 2PL models, respectively

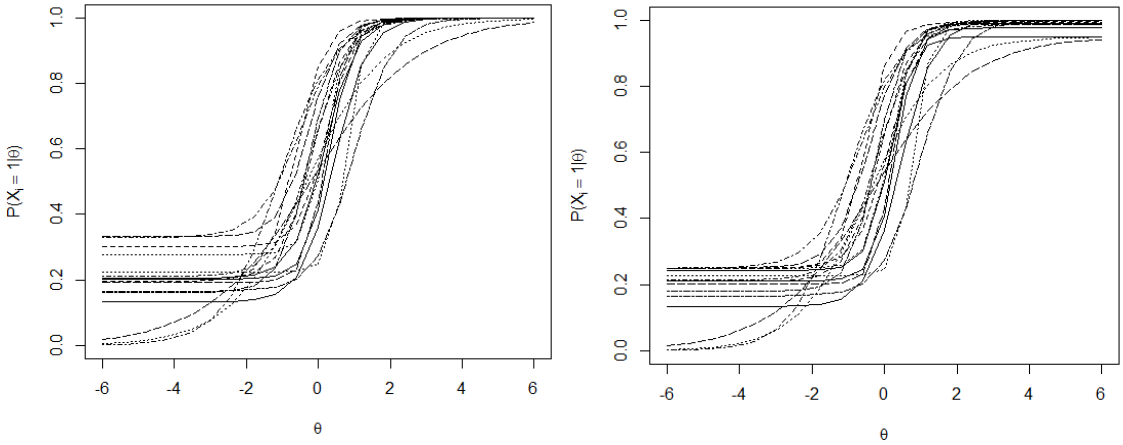
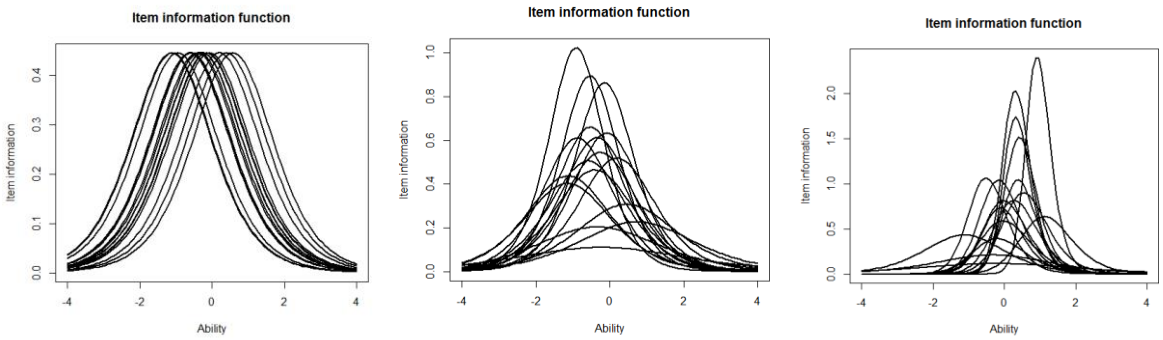


Figure 2. ICCs according to 3PL and 4PL models, respectively

As can be seen in Figures 1 and 2, the probability of correct response to an item increases as the level of the individual's ability increases in the four models, that is, the item characteristic curves increase monotonically.

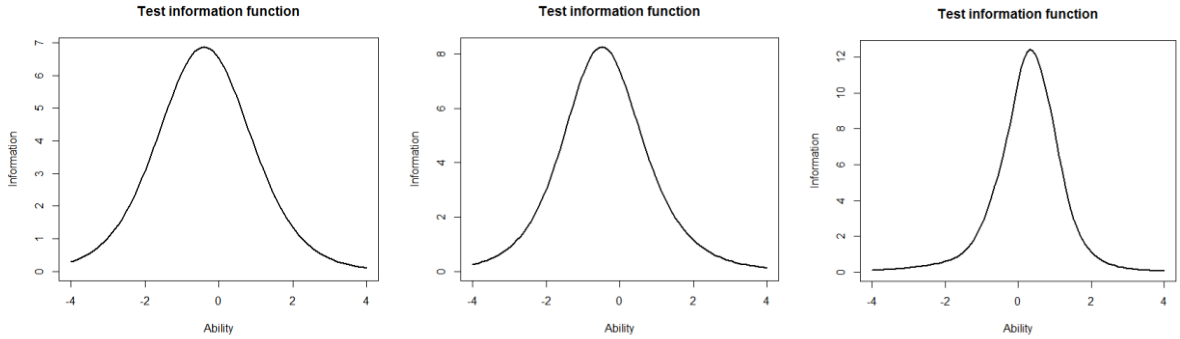
In order to determine whether the test is a speed test, the variance of number omitted items was divided by the variance of the number of incorrectly answered items. The rate found was near zero, and the test is accepted not to be a speed test (Hambleton & Swaminathan, 1985). Moreover, the rate of responding the items correctly is also examined and it is seen that it varied between .36 (item6) and .75 (item17), and that the rates of responding to the final items correctly are similar to those of other items.

Item and test information graphics based on 1, 2 and 3 PL models related to reliability were created. The graphs for items, test information values and functions are calculated and drew in R with the help of the "irtoys" package (Partchev, 2017). Item and test information functions according to three models are given in Figures 3 and 4. Since there is no package which calculates the information function according to the 4PL model, it could not be drew.



1PL model 2PL model 3PL model

Figure 3. Item information functions for three models



1PL model 2PL model 3PL model

Figure 4. Test information functions for three models

As it can be seen in the Figures 3 and 4, predictions under 3 PLM provided the most information for a higher ability group than other models. The model that provides information for the largest ability level is the 1PL model, which is also the one with the least information.

As a results of the examinations, it is concluded that the 17-item science and technology sub-test meets the IRT's assumption. Analyses were conducted according to four models (2-, 3-, 4 PLs and mixture-IRT) to determine which model fits the data better to, in other words to find an answer to the first research question presented above. Estimates were made for 2-, 3-, 4 PL and mixture-IRT in Mplus 8 program (Muthén & Muthén, 2017). The Bayesian information criterion (BIC) value which is recommended in the literature to determine the model data fit (Li, Cohen, Kim & Cho, 2009) and -2 log $\chi 2$ values of the models (Hambleton et al., 1991) is used for comparisons. Then, for the second research question, the parameter values of the fitting model are presented and interpreted.

3. FINDINGS

Analysis which were conducted to determine the most appropriate IRT model for TEOG 2015 science and technology subtest data resulted some model fit indices to be discussed. Some indices such as likelihood- (LL), the degree of freedom (df), BIC and Akaike Information Criterion (AIC) are presented in Table 2.

Table 2. Model data fit results based on models

Models	LL	df	BIC	AIC
2PL	-48112.103	34	96513.790	96292.206
3PL	-47744.809	51	95923.994	95591.617
4PL	-47773.491	68	96126.150	95682.981
MixIRT (2PL) 1-Latent Class	-48112.110	34	96513.804	96292.220
MixIRT (2PL) 2- Latent Class	-47757.030	53	95965.471	95620.060
MixIRT (2PL) 3- Latent Class	**-47649.129**	**72**	**95911.496**	**95442.258**
MixIRT (2PL) 4- Latent Class	-47599.948	91	95974.961	95381.896
MixIRT (3PL) 2- Latent Class	-47643.375	86	96019.228	95458.749
MixIRT (3PL) 3- Latent Class	-47588.756	121	96208.093	95419.512

As it can be seen in Table 2, when traditional IRT models (2, 3 and 4PL) are examined solely with the LL, BIC and AIC values, the model that fits best is the three-parameter model. When predictions are made with MixIRT models, the model that best fits the data according to the BIC value, which is the best indicator of model data fit, is the model predicted according to MixIRT with three latent classes (3LC) and two parameters. When deciding on the model data fit, together with taking the BIC value under consideration, -2 log $\chi 2$ values can be compared. In this context, Chi-Square statistics, the degree of freedom and the difference between the values of -2 log $\chi 2$ belonging to the 2- and 3PL models were evaluated at first. Since the calculated value ($\chi 2$ = 48112.103-47744.809 = 367.294) is greater than the table value ($\chi 2_{(17;\ 0.05)}$ = 27.857), the difference between -2 log $\chi 2$ values is significant. In this case, it can be said that the three-parameter model is more suitable for data. Then, the Chi-Square statistics, the degree of freedom and the difference of the -2 log $\chi 2$ values belonging to the 4PL and 3PL models are evaluated. Since the calculated value ($\chi 2$ = 47773.491- 47744.809 = 28.682) is greater than the table value ($\chi 2_{(17;\ 0.05)}$ = 27.857), the difference between -2 log $\chi 2$ values is significant. In this case, it can be stated that the three parameter model for the data is more suitable than the four parameter model. When compared to the model with the lowest BIC value among MixIRT models, since the calculated value ($\chi 2$ = 47744.809-47649.129 = 95.68) is greater than the table value ($\chi 2_{(21;\ 0.05)}$ = 32.671), the difference between the values of -2 log $\chi 2$ is significant. In this case, it is stated that the two parameter MixIRT model with three latent classes is more suitable for the data. The results of the two-parameter MixIRT model with three latent classes are given in Table 3 in order to present the item parameters [(item discrimination (a) and item difficulty (b)] according to the model which fits best to the data.

As shown in Table 3, 37% (N: 1868) of the students are in the first latent class, 37% (N: 1848) of the students are in the second latent class and 26% (N: 1284) of the students are in the third latent class. When the gender distribution of the students in latent classes is examined, it is seen that the ratio of the students in terms of gender in all the latent classes is very close. When item-model fit is evaluated, it is indicated that the difficulty values of one item (i6) in the first latent class, three items (i2, i11 and i13) in the second latent class and two items (i4 and i16) in the third latent class do not fit to the model. It is thought that the reason why different items in different latent classes do not fit the data is resulted from the different traits individuals carry in the latent classes. Within the scope of this research, the emerged latent classes could not be interpreted in more details because information obtained from MoNE is limited to individual responses for items and their gender.

In latent classes, the item discrimination averages are respectively; 1.70, 0.77 and 0.27. It is observed that discrimination decreases from the latent class-1 to the latent class-3. Item difficulty averages in latent classes are respectively; 1.33, -0.79 and 4.20. In this context, it can

be expressed that items are moderately difficult and discriminative for students in latent class-1; the items are considerably easy and able to slightly distinguish the students in latent class-2; the items are difficult to the students in the third latent class and they can slightly distinguish the students in this group.

Table 3. Item parameters in each model for 2PLM with three latent classes

	LC1		LC2		LC3	
Gender	Frequency	%	Frequency	%	Frequency	%
Female	872	49	879	48	674	49
Male	921	51	953	52	701	51
Total	1868	37	1848	37	1284	26
Items	a	b	a	b	a	b
i1	1.946	1.010	0.930	0.366	0.306	3.628
i2	1.454	1.732	0.695	-0.029*	0.229	3.149
i3	1.285	0.757	0.614	2.315	0.202	5.232
i4	0.922	3.512	0.441	-2.675	0.145	1.093*
i5	1.574	1.949	0.752	-1.632	0.247	1.612
i6	1.070	0.136*	0.511	1.854	0.168	8.284
i7	1.512	0.734	0.722	0.736	0.238	6.439
i8	2.116	1.753	1.011	-1.939	0.332	1.427
i9	0.731	1.151	0.349	-1.973	0.115	9.767
i10	1.296	1.753	0.619	-1.129	0.204	3.878
i11	2.328	1.062	1.112	0.186*	0.366	3.278
i12	2.675	0.968	1.278	-0.991	0.420	1.899
i13	1.840	0.969	0.879	-0.199*	0.289	2.978
i14	2.059	1.071	0.984	-0.947	0.324	2.277
i15	3.306	0.660	1.579	-0.327	0.519	1.027
i16	1.471	1.490	0.703	-3.027	0.231	0.630*
i17	0.665	0.669	0.318	-1.716	0.105	8.131

*p>.05

Item discrimination values of the items in the first latent class vary between 0.665 (i17) and 3.306 (i15), and the item difficulty values range from 0.660 (i15) to 3.512 (i4). The item discrimination values of the items in the second latent class are between 0.318 (i17) and 1.579 (i15), and the item difficulty values range from -0.327 (i16) to 2.315 (i3). Item discrimination values of the items in the third latent class are between 0.105 (i17) and 0.519 (i15), and the item difficulty values range from 1.027 (i15) to 9.767 (i9). When the difficulty range of items is examined in three latent classes, it is seen that the vast majority of the items in the second latent class have negative difficulty value. In this context, it can be expressed that the items are easier for the students in this group. Yet, in the third latent class, it is seen that the difficulty values of the items increase. This situation makes it possible to state that the items are difficult for the students in this group.

The item with the lowest discrimination value in all three latent classes is item-17, which is the last item in the test. This item is a question asking the relationship between the weight of the objects and the lifting force applied to objects which are in status of swimming and sinking. When the students' response distribution to the choices for this item is examined, 75% of the students have marked the wrong "C" option. Only 6% of students responded correctly to this item. However, when the difficulty levels of the item in the latent classes are examined, it is seen that this is an easy item for the students in the second latent class. This is also constituting as an example of the change of item parameters according to MixIRT in latent groups.

Furthermore, item-15 is the item that has the highest discriminate value in all three latent classes. At the same time, this item has the lowest difficulty value in three latent classes. This item corresponds to the item-17 in the original test (since item-13 is cancelled, and item-16 is excluded from the analysis). When the students' response distribution for this item is examined, it is observed that 60% of the students marked the option "B" which is one of the wrong choices. When the relevant question and the "B" option are examined, it is seen that 60% of the students turn towards the wrong conceptual knowledge that the intensity of an object in swimming state is equal to that of the liquid it is in. This has led to the increase in the item discrimination, and for the item to have a difficult trait.

In the first latent class, the fourth item which was observed to have the highest item difficulty value (b=3.512) was determined to be considerably easy. However, in the second latent class, it had the lowest difficulty value (b=-2.675), which means it was a difficult item. In the third latent class, on the other hand, this item showed no significant fit with the model tested. According to the CTT, the item difficulty of this item is .75. In this case, it can be stated that an item which seems quite easy according to the CTT could be a difficult question for students in some latent groups. More detailed studies should be conducted as to why this problem prepared on the subject of "genetic crossing" has been identified as difficult in the second latent class. Findings can be interpreted for all items similarly. In this study, only a few remarkable items have been interpreted.

4. DISCUSSION AND CONCLUSION

The aim of this study is to determine to which IRT models [(Rasch, 2PL, 3PL, 4PL and mixed-IRT (2 and 3PL)] TEOG 2015 science and technology sub-test conducted at national level fits best. In addition, it is also aimed to predict item parameters for the model that fits best. For this purpose, before analysing the data, the assumptions of IRT (unidimensionality, local independence, monotone increase of the item characteristics curve and whether the test is a speed test) are tested and all assumptions were seen to be met. Predictions are made for the 2-, 3- and 4-PL and the 2- and 3-PL models according to the MixIRT in order to determine the model that fits data best. Then the item parameters are predicted for three latent classes separately according to MixIRT model with two parameters and three latent classes which is also fits the model the best. When the gender distribution of the students in latent classes is examined, it is seen that the ratio of the students based on gender in all the latent classes is very close. When items' fitting to the model is evaluated, difficulty value of one item in the first latent class, three items in the second latent class, and two items in the third latent class do not fit to the data. When the difficulty range of the items and the difficulty averages in all three latent classes are examined, it is seen that the vast majority of the items in the second latent class have negative difficulty value. In this context, it can be stated that the items are easier for the students in this latent class than it is for the first and third class. The difficulty values of the items are seen to increase in the third latent class. For this case, it can be stated that the items are difficult for the students in this group. This finding is consistent with findings obtained from the study of De Ayala and Santiago (2017) in which the MixIRT models are tested with the mathematical abilities of students in the 1-3 latent class according to the 1PL model. According to the fitting model, it is determined that some of the items have been found easier by those in a latent class while harder for some others in this study as well.

When the discriminate values of the items are examined, it is seen that the highest discriminate values are in the latent class one. Considering the difficulty and discrimination averages of the items in the latent classes, it can be expressed that items are of moderate difficulty and discriminative for students in latent class-1; the items are considerably easy and able to distinguish the individuals a little for the students in the latent class-2; the items are

difficult to the students in the third latent class and they can distinguish the students in this group a little. When the results are evaluated in general, students who have the lowest science literacy are most probably in the LC-3 (students with the lowest science achievements). Students who have science literacies at the highest level are most probably in LC-2 (students with the highest science achievements). Furthermore, students who have science literacies at the moderate level are most probably in LC-1 (students with the intermediate science achievements). In this context, it is recommended to carry out studies in which many variables such as school, teacher and student characteristics are discussed together in order to be able to put forward the profiles of the students in the emerged latent classes.

The item with the lowest discrimination value in all three latent classes is item 17, which is the last item in the test. This item seems to be an easy item for students in the second latent class when the difficulty values of this item in the latent classes are examined. This is also an example of the change of the item parameters according to MixIRT in latent groups. The item 15, is the item that has the highest discrimination value in all three latent classes. This item also has the lowest difficulty value in all three latent classes. It is observed that this item, which seems quite easy according to the CTT, could be a difficult item for students in some latent groups.

MixIRT is based on the assumption that the sample consists of latent subclasses. Different from IRT, MixIRT does not assume that the item parameters remain invariant among the groups. It is flexible on this subject and it allows the change of item parameters between the latent classes (De Ayala & Santiago, 2017). Separating students' ability into latent classes allows researchers to obtain more reliable thus more valid information about item and group characteristics. In addition, MixIRT approach also enables modelling both continuous and categorical data at the same time and this makes it possible to gather more information (De Ayala & Santiago, 2017).

In order for the estimates made to be less inaccurate, different models based on theories must be discussed in the analysis of the data in the exams that are conducted at the national level and for the purpose of selection and placing of the students to a secondary education institution. This research is the first study to compare the model fit data according to MixIRT models for a national test in Turkey. In this context, it is important to support the results obtained with the studies to be made on this subject with different subtests in different years.

The conducted study also has some limitations. First of all, only the TEOG 2015 application science and technology subtest has been dealt with in the study. Interested researchers can test the model-data fit for the data of different subtests in different years. Moreover, dichotomous data were studied in this study. Interested researchers can compare the results by using MixIRT in polytomous items with traditional models. Finally, no constraints have been identified while making parameter predictions for the IRT. Researchers who are interested in studying on this subject can examine the model data fit by setting constraints for item parameters.

ORCID

Seher Yalçın https://orcid.org/0000-0003-0177-6727

5. REFERENCES

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

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2001). A mixture item response for multiple-choice data. *Journal of Educational and Behavioral Statistics*, 26, 381–409.

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.

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148. doi: 10.1111/j.1745-3984.2005.00007

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

De Ayala, R. J. & Santiago, S. Y. (2017). An introduction to mixture item response theory models. *Journal of School Psychology*, 60, 25-40. doi: 10.1016/j.jsp.2016.01.002

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

Egberink, I. J., Meijer, R. R. & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality*, 44, 232–244.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: 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), 167-178. doi: 10.22237/jmasm/1335845580

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and application.* Boston: Kluwer Academic Publishers Group.

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

Kelderman, H., & Macready, G. B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement*, *27*(4), 307–327.

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 dissertation). Middle East Technical University, Graduate School of Social Sciences, Ankara.

Li, F., Cohen, A. S., Kim, S., & Cho, S. (2009). Model selection methods for mixture dichotomous IRT models. *Applied Psychological Measurement*, *33*(5), 353–373. doi: 10.1177/0146621608326422.

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), 509-525. doi:10.1348/000711009X474502

Maij-de Meij, A. M., Kelderman, H., & Van der Flier, H. (2008). Fitting a mixture item response theory model to personality questionnaire data: Characterizing latent classes and investigating possibilities for improving prediction. *Applied Psychological Measurement*, 32, 611–631.

Maij-de Meij, A. M., Kelderman, H. & van der Flier, H. (2010). Improvement in detection of differential item functioning using a mixture item response theory model. *Multivariate Behavioral Research, 45*(6), 975-999. doi:10.1080/00273171.2010.533047

Muthen, B. & Asparouhov, T. (2006). Item response mixture modeling: Application to tobacco dependence criteria. *Addictive Behaviors*, 31, 1050-1066.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (Eighth Edition). Los Angeles, CA: Muthén & Muthén.

Partchev, I. (2017). *Simple interface to the estimation and plotting of IRT Models*. R-project, Package 'irtoys' manual. Retrieved from https://cran.r-project.org/web/packages/irtoys/irtoys.pdf

Robitzsch, A. (2018). *Supplementary item response theory models*. Retrieved from https://cran.r-project.org/web/packages/sirt/sirt.pdf

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, 14, 271–282.

Skrondal, A. & Rabe-Hesketh, S. (2007). Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, *34*(4), 712–745. doi: 10.1111/j.1467-9469.2007.00573.x

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars, & A. L. McCutcheon, *Applied latent class analysis* (p. 89-107). New York: Cambridge University Press.

Vermunt, J. K., & Madigson, J. (2004). Local independence. In A. B. M. S. Lewis Beck (Ed.), *Encyclopedia of social sciences research methods* (pp. 732-733). Thousand Oaks: Sage Publications.

von Davier, M. & Rost, J. (2017). Logistic mixture-distribution response models. In W. J. van der Linden (Ed.), *Handbook of item response theory, volume one: Models* (p. 393-406). Boca Raton: Chapman and Hall/CRC.

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with non-standard item response theory models: Fitting the four-parameter model to the minnesota multiphasic personality inventory. In Embretson, S. (Ed), *New directions in psychological measurement with model-based approaches* (p. 147-173). Washington DC: American Psychological Association.