Received Date: 30 December 2022 | Accepted Date: 28 May 2023

OPUS Journal of Society Research

RESEARCH ARTICLE

Examining Group Differences in Mathematics Achievement: Explanatory Item Response Model Application

Erdem Boduroğlu[1] | Duygu Anıl[2]

[1] PhD., Candidate Hacettepe University, Ankara /Türkiye
ORCID: 0000-0001-8318-4914
E-Mail: erdemboduroglu@gmail.com

[2] Prof., Dr., Hacettepe University, Ankara / Türkiye
ORCID: 0000-0002-1745-4071
E-Mail: duygu.anil73@gmail.com

Corresponding Author: Erdem Boduroğlu

May 2023
Volume:20
Issue:53
DOI: 10.26466//opusjsr.1226914

Citation: Boduroğlu, E.& Anıl, D. (2023). Examining group differences in mathematics achievement: Explanatory item response model application. *OPUS– Journal of Society Research, 20*(53), 386-397.

Abstract

Students take many different exams throughout their educational lives. In these exams, various individual and item characteristics can affect the responses of individuals to the items. In this study, it was aimed to examine the effects of person and item predictors on the mathematics common exam results of 365 9th grade students with explanatory item response models. Gender and school type as person variables and cognitive domain, content domain and booklet type as item variables were added to the models due to their widespread inclusion in the literature. When the predicted item parameters were examined, it was seen that the smallest parameter values were obtained for all items with the Rasch model. When the model data fit values of four different models were examined, it was concluded that the latent regression and latent regression linear logistic test models showed better fit than the Rasch model. By adding person and item predictors to the model, the parameters obtained for each variable group were compared, and differences were observed between the groups for school type, cognitive domain, and content domain variables. It was concluded that the item parameters did not differ for the variables of gender and booklet type. It is thought that it would be beneficial to use these models more widely in studies to be conducted in the field of education and psychology, since they provide more detailed information about the reasons for the differences in the estimated parameters.

Keywords: *Explanatory item response models, rasch model, math success*

Öz

Öğrenciler eğitim hayatları boyunca birçok farklı sınava katılmaktadır. Bu sınavlarda, çeşitli birey ve madde özellikleri öğrencilerin maddelere verdikleri yanıtları etkileyebilmektedir. Bu çalışmada 9.sınıflarda öğrenim gören 365 öğrencinin matematik dersi ortak sınav sonuçları üzerinde birey ve madde yordayıcılarının etkisinin açıklayıcı madde tepki modelleri ile incelenmesi amaçlanmıştır. Alanyazında araştırmalara yaygın olarak dahil edilmesi sebebiyle; birey değişkeni olarak cinsiyet ve okul türü, madde değişkenleri olarak ise bilişsel alan, içerik alanı ve kitapçık türü değişkenleri modellere eklenmiştir. Kestirilen madde parametreleri incelendiğinde, Rasch modeli ile tüm maddeler için en küçük parametre değerlerinin elde edildiği görülmüştür. Dört farklı modelin model veri uyumu değerleri incelendiğinde ise örtük regresyon ve örtük regresyon doğrusal lojistik test modellerinin Rasch modeline göre daha iyi uyum gösterdiği sonucuna ulaşılmıştır. Birey ve madde yordayıcıları modele eklenerek her bir değişken grubu için elde edilen parametreler karşılaştırılmış ve okul türü, bilişsel alan, içerik alanı değişkenleri için gruplar arasında farklılıklar gözlenmiştir. Cinsiyet ve kitapçık türü değişkenleri için ise madde parametrelerinin farklılaşmadığı sonucuna ulaşılmıştır. Bu modellerin, kestirilen parametrelerdeki farklılıkların nedenlerine ilişkin daha detaylı bilgiler sunması sebebiyle eğitim ve psikoloji alanında yapılacak çalışmalarda daha yaygın kullanılmasının faydalı olacağı düşünülmektedir.

Anahtar Kelimeler: *Açıklayıcı madde tepki modelleri, rasch modeli, matematik başarısı*

Introduction

Many research areas cannot produce valid information without adequate measurement of various psychological qualities such as intelligence or personality traits of individuals (Sijtsma, 2020). The main purpose of measurement and evaluation studies carried out in the fields of education and psychology is to reach comprehensive and reliable information about individuals based on their responses to scale items. It is not possible to say that the individual characteristics that are the subject of the research are always directly observable. Various measurement tools are needed to reveal the degree to which individuals have the variable that is aimed to be measured. Measurement tools such as scales, questionnaires and tests are typical instruments used by researchers to measure a construct or trait (Desjardins & Bulut, 2018). Many theories have been developed about the process of developing these instruments and interpreting test scores. Among these theories, Classical Test Theory (CTT) and Item Response Theory (IRT) are the most widely used (Crocker & Algina, 1986; Embretson & Reise, 2000).

Various problems arise with test development and analysis of scores, as the assumptions of the CTT are weak assumptions that can be met by most test data. For example, item difficulty and item discrimination index depend on the skill level and range of ability scores of the individuals taking the exam. Items will be interpreted more easily in exams attended by individuals with high ability levels. Item discrimination tends to be higher in heterogeneous groups than in homogeneous groups. The reliability of the test is also directly related to the test scores of the individual sample taking the test. Similarly, the ability levels of individuals vary according to the difficulty of the test, and various problems arise in the comparison of the ability levels of individuals who take the test with different difficulties. Because of these and similar problems, psychometrists needed to develop more suitable measurement models (Hambleton, Swaminathan, & Rogers, 1991).

IRT is one of the theories developed to reveal the extent to which individuals have features that cannot be directly measured, which are called latent features. While the raw score of the individual in CTT is obtained by the sum of the scores obtained from each scale item, the IRT is concerned with whether the answers given to the items are correct or wrong rather than the total score. When responding to an item, it is assumed that the participants have a certain amount of the underlying feature (ability) and a corresponding score on the skill scale is assigned to each participant. In this sense, IRT reveals the relationship between the ability levels of individuals and their probability of correctly responding to the items. While the probability of answering the item correctly for individuals with high ability level is close to 1, the value of this probability approaches 0 as the skill level decreases (Baker, 2001). One of the important advantages of the theory is that individual abilities are independent from the test applied and the item sample, and that the item parameters are independent from the group. Test scores obtained from the CTT can vary significantly across tests. Therefore, it is easier to compare individuals' performances in different tests within the framework of the IRT than the CTT (Desjardins & Bulut, 2018). Another advantage is the use of information functions that can be defined at the item level to calculate reliability in the model. Higher level of knowledge indicates lower standard error and higher reliability. Thanks to these advantages, IRT has recently; its use in computerized adaptive testing, item and test bias determination, and test equating studies is becoming increasingly common (DeMars, 2010).

The main factor in the recognition of IRT as a powerful modeling method is the necessity of meeting strong assumptions (Embretson & Reise, 2000). There are different classifications for these assumptions in the literature. DeMars (2010) discussed the assumptions under three headings as unidimensionality, local independence and appropriate model properties. Unidimensionality means that test items are associated with only one latent feature. It is the situation where there is only one dominant factor that affects the reactions of individuals to the items and the probability of answering that item correctly. In some cases, psychometrists state that tests measure factors such as speed, motivation, and excitement apart

from the individual's ability to be measured. However, this does not always mean that unidimensionality is violated. For example, if the motivation level of all participants is high, this variable may not be interpreted as a separate dimension. The main thing is to determine whether there is a single dominant factor by various statistical methods. Violation of the unidimensionality assumption can lead to incorrect parameter and standard error estimations. Another assumption of the IRT is local independence. In order to ensure this assumption, when the ability level to be measured in the test is kept constant, the answers given by the individuals to the items should be statistically unrelated. In local independence, the relationship between item pairs is examined, but even if the inter-item relationship is not observed, any subset of the test can create a new dimension. Therefore, unidimensionality and local independence assumptions should be tested separately. The last assumption is to examine the model data fit to determine the appropriate model. The emergence of the advantages provided by IRT is possible with the model data fit (Orlando & Thissen, 2000). If the appropriate model is not selected, the estimated parameters will be incorrect. For example, using the 1 PL model when the items have different slope values or the horizontal asymptote is different from zero will lead to erroneous results. There are many statistical methods used to determine model data fit.

The main purpose of many item response models developed is to measure the latent features underlying human behavior based on individual performances or test responses. In standard item response models, items and individuals are represented by one or more parameters. The estimated individual parameters provide a reference for the measurement of latent traits. This general approach falls short of explaining the differences in individual and item parameters and their reasons. Since the cognitive processes that individuals use when answering items cannot be modeled with traditional IRT, alternative approaches are needed. For this purpose, Explanatory Item Response Models, which reveal the effect of item and individual characteristics on responses to items and have a broader statistical approach than standard item response models, have been developed (De Ayala, 2022; De Boeck & Wilson, 2004).

Many of the current item response models are more specific and stretched versions of generalized linear or nonlinear mixed models (GLMM and NLMM). Explanatory item response models also appear as a flexible model that provides access to a wider knowledge base by strongly linking psychometry to the field of statistics. In item response models used in educational research, individuals are typically viewed as a unit of analysis. When individual covariates are added to the model to describe or explain the differences between individuals, the model determined within the framework of GLMM turns into an explanatory item response model. In psychological research, on the other hand, the items themselves are usually units of analysis, and by adding item covariates to the model, the existing model turns into an explanatory response model (Briggs, 2008). Explanatory item response models were handled under four main headings as descriptive and explanatory models in terms of their individual and item characteristics, and these models are given in Table 1 (De Boeck & Wilson, 2004).

Table 1. Explanatory item response models

	Person Predictors	
Item Predictors	Absence	Inclusion
Absence	Doubly Descriptive	Person Explanatory
Inclusion	Item Explanatory	Doubly Explanatory

The models given in Table 1 represent only a small subset of the set of possible models. Situations where person and item characteristics are not added are called the doubly descriptive model (Rasch model). This model is the basic version of the explanatory item response models. If only person characteristics are added to the model, it is called person explanatory item response model (latent regression model), if only item properties are added, it is called an item explanatory item response model (linear logistic test model [LLTM]). In cases where both are added, it is called a doubly explanatory item response model (latent regression linear logistic test model) [LRLLTM]).

The formulas of these four models are given in Table 2. The expression θp in the table indicates that the individual parameters are randomly drawn from the universe and exhibit a normal distribution with a mean of zero. Z value indicates individual characteristics, j subindex indicates person predictors. X is the item predictors expressed with the k sub-index. The cases where the person predictors have fixed effects are expressed with ϑj, and the cases where the item predictors have fixed effects are expressed with β_k.

Table 2. Summary table of explanatory item response models

	$\eta_{pi} =$			
Model	Person Part	Item Part	Random Effect	Model Type
Rasch	θ_p	$-\beta_i$	$\theta_p \sim N(0, \sigma_\theta^2)$	Doubly Descriptive
LRM	$\sum_j \vartheta_j Z_{pj} + \theta_p$	$-\beta_i$	$\varepsilon_p \sim N(0, \sigma_\theta^2)$	Person Explanatory
LLTM	θ_p	$-\sum_k \beta_k X_{ik}$	$\theta_p \sim N(0, \sigma_\theta^2)$	Item Explanatory
LRLLTM	$\sum_j \vartheta_j Z_{pj} + \theta_p$	$-\sum_k \beta_k X_{ik}$	$\varepsilon_p \sim N(0, \sigma_\theta^2)$	Doubly Explanatory

In Table 2, the negative sign of the β coefficients in the item sections is interpreted as the item convenience coefficient and is included in the formula as a component that increases the probability of the item being answered correctly. After estimations, this value can be multiplied by minus and converted to item difficulty value (Boeck et al., 2011). "As can be seen from Table 2, person characteristics in the latent regression model, item characteristics in the linear logistic test model, and both feature groups in the latent regression linear logistic test model are added to the model and necessary estimations are made.

The framework of explanatory item response models was drawn by De Boeck & Wilson (2004), and these models are flexible and useful approaches in that they enable the inclusion of different person and item variables in the model. Although explanatory item response models appear as a relatively new approach, it is seen that these models are used in recent studies. When the literature is examined, it is more common to encounter studies that focus on variable features. For example, Briggs (2008), in his study on 10th grade students, concluded that the ethnic origin of individuals is a significant predictor of the ability parameters obtained from the science test. Atar (2011) used the variables of gender, attitude towards mathematics lesson, giving importance and self-confidence as person variables, cognitive domain and content domain as item variables and created explanatory item response models in his study conducted with TIMSS 2007 Turkey 8th grade mathematics data. It was concluded that the variable of self-confidence was an important predictor of the student's mathematics achievement, and that the variables of gender, attitude towards mathematics and giving importance to the course had no effect on mathematics achievement. In the results of the linear logistic test model analysis created with the cognitive domain and content domain, it was concluded that these variables had an effect on the item difficulty. Kahraman (2014) conducted his study with the data of multiple choice and applied test, which is the last stage of a three-stage exam attended by physician candidates. In the study in which five different models were compared, the partial credit model, which did not include the predictor variable, was used as the base model. Then, as predictor variables for this model; New models were created by adding the order of application of the item, the time spent on the item, the gender of the candidate and the multiple-choice test score separately. Among these variables, it was concluded that only the multiple choice test score was a good predictor, while the other variables were not useful as a predictor. Chen, Yang, Bulut, Cui, & Xin (2019) examined personality factors affecting drug use using explanatory item response models in their study. In addition to gender and alcohol use as person variables, anxiety sensitivity, impulsivity, sensation seeking and hopelessness variables were added. As a result; gender, alcohol use and their interaction, the interaction between gender and hopelessness, and sensation seeking were found to be significant predictors of substance use level. Apart from these studies, researches using explanatory item response models have been encountered more frequently recently (Atar & Aktan, 2013; Bulut, Palma, Rodriguez, & Stanke, 2015; Büyükkıdık & Bulut, 2022; Chiu, 2016; Kim & Wilson, 2020; Min, Zickar & Yankov, 2018;

Petscher et al, 2020; Randall, Cheong, & Engelhard, 2011; Tat, 2020; Yavuz, 2019).

Traditional item response models do not include explanatory variables regarding the differences between individuals' abilities and item difficulties. The effect of the explanatory item response models used in this study and the addition of various person and item variables to the model on the parameters were examined. Variables that can make a difference on item and ability parameters such as gender, school type, booklet type, item content were included in the model. Although there are explanatory item response model studies conducted on the data set of international exams such as PISA, TIMSS, and PIAAC in the literature, no study was found with the data of the common exams conducted throughout the province. For this purpose, it is thought that it is important to use these models, which include person and item variables, in the common exam practices that have become widespread recently and in the reporting processes afterwards.

Aim of the research

In this study, it is aimed to examine the item parameters obtained from descriptive and explanatory item response models. For this purpose, answers to the following three questions will be sought in the research.

1. What are the item parameters obtained from Rasch, linear logistic test, latent regression and latent regression linear logistic test models?
2. How is the model data fit obtained from Rasch, linear logistic test, latent regression and latent regression linear logistic test models?
3. Does the mathematics lesson performance of individuals differ in the subgroups of the variables handled with the latent regression linear logistic test model?

Method

In this study, the effect of various person and item characteristics on mathematics achievement was examined. In this respect, it can be stated that the study is in the correlational research design, which is one of the quantitative research methods. Correlational research aims to reveal the relationships between two or more variables as they are (Fraenkel, Wallen, & Hyun, 2012). In this section, the research data and the analysis of the data are given.

Research Data

The data used in the research were obtained from 365 individuals selected by cluster sampling method among the students who participated in the mathematics lesson common exam applied to the 9th grades throughout the province of Niğde. Individuals responded to 20 multiple-choice test items in the joint exam. Common exam items; It has been prepared in accordance with the secondary education curriculum with 5 domain experts, 1 measurement and evaluation expert and 1 language expert. After the pilot application was carried out, the test and item statistics were examined and the final form was created. After the final application, it was seen that the item difficulties ranged between .36 and .78 and the average difficulty of the test was .49. Item discrimination was calculated with the point biserial correlation coefficient and these values were found to vary between .41 and .89. The average of the discrimination values was calculated as .63. The KR-20 value calculated for the reliability of the test was found to be .874. According to all these results, it can be stated that the test has medium difficulty, high discrimination and reliability (Crocker & Algina, 1986).

The person variables to be used in the research were determined as gender (male-female) and school type (science-vocational-anatolia-religious). Item variables are cognitive domain (knowing - applying -reasoning), content domain (number-algebra-geometry) and booklet type (A-B) of test items. The items in the booklets are the same, only the positions of the items in the test differ. While determining the cognitive domain and content domain, the domain classification used in the TIMSS research was taken as reference. Since there is no item related to probability in the content domain, the existing items are grouped into three categories. Opinions were received from 5 domain

experts to classify the items according to cognitive and content domain. Fleiss Kappa statistics were calculated to determine the agreement among experts. Inter-expert agreement was found to be κ=.734 for the cognitive domain and κ=.819 for the content domain. These values indicate a high level of agreement among experts (Landis & Koch, 1977). Descriptive statistics on person and item variables used in the study are given in Table 3.

Table 3. Frequency table for person and item variables

Variables	Category	Frequency	Percentage
Gender	Female	163	44,6
	Male	202	55,4
School Type	Science	61	16,7
	Vocational	90	24,7
	Anatolia	158	43,3
	Religious	56	15,3
Booklet Type	A	180	49,3
	B	185	50,7
Cognitive Domain	Knowing	4	20,0
	Appyling	14	70,0
	Reasoning	2	10,0
Content Domain	Number	9	45,0
	Algebra	9	45,0
	Geometry	2	10,0

When the variables in Table 3 are examined, it is seen that the number of male students is more than female students. According to the school type, the most students are in Anatolian high schools. It can be said that the booklet type is distributed in a balanced way as A and B. 20 test items used in the common exam were prepared by domain experts according to the relevant acquisitions, and most of them are at the "Applying" level. There are only 2 items in the cognitive domain of “reasoning”. Considering the content domain distributions, it is seen that the items are mostly from the domains of numbers and algebra, and there are 2 items in the test from the domain of geometry.

Analysis of Data

Akaike Information Criteria (AIC), Bayesian Information Criteria (BIC) and -2loglikelihood (-2LL) values of Rasch, 2PL and 3PL models were examined to decide which IRT model to use in the research. Cases where these values are small indicate better model-data fit (De Ayala, 2013). The lowest AIC and BIC values were obtained for the Rasch model, and the lowest -2LL values were obtained for the 3PL model. Due to the research purpose and the simplicity of the model, it was decided to conduct the research with the Rasch model.

It was tested whether the Rasch model satisfies the assumptions of unidimensionality and local independence. A unidimensional test includes items that fall into only one dimension. Any factor that affects the reaction to the items is only considered as a random error or unpredictable dimension specific to that item, and other items are not affected by this situation. Violation of the unidimensionality assumption leads to incorrect estimation of parameters and standard errors (DeMars, 2010). To examine the unidimensionality of our research data, exploratory factor analysis based on the tetrachoric correlation matrix was applied. In cases where 1-0 is scored, that is, in calculating the relationship between two artificially paired variables, tetrachoric correlation is used. As a result of the factor analysis performed to decide on the number of dimensions, the eigenvalue of the first factor was calculated as 2.985 and the second factor as 0.281. It was seen that there was only one factor with an eigenvalue greater than 1 and it was decided that the scale was unidimensional (Kaiser, 1960). Then, the local independence assumption was tested. According to this assumption, individuals' responses to any two test items are statistically independent when the abilities that affect test performance are held constant. In order to provide the unidimensionality assumption, the items in the test must be related, while in the local independence assumption, the items must be independent for a certain ability level. The local independence assumption was tested with Yen's Q3 test developed by Yen (1981). The residual correlation values matrix was examined with Yen's Q3 test and it was seen that the correlation values between the item pairs varied between 0.002 and 0.197. The fact that these values are below the 0.20 cut-off point indicates that the assumption of local independence is met (Chen & Thissen, 1997).

In order to make parameter estimation with four different models to be discussed in the research, the "eirm" package developed by Bulut (2021) was used in the R program. The Rasch

model was used as the first method for estimating item difficulties. In this model, no item and person variables were added to the equation. In the second model, the linear logistic test model, item variables (cognitive domain + content domain + booklet type) were included in the process. In the latent regression test model, which is the third model, person variables (gender + school type) variables were added to the model. The last model is the latent regression linear logistic test model. In this model, both item and person variables were added. Item parameters and standard error values of 20 items were estimated for all models. In order to test whether the difference between the item parameters estimated from the four models is significant, the assumptions were checked and the ANOVA test was applied for repeated measurements. For the same item group, the outputs from each model were considered as a measure. In this way, it was examined whether at least one model differed from the others for four different model outputs. In the next step, the model data fit of the models was calculated. For this purpose, the AIC, BIC and -2LL values among the "eirm" package outputs were compared. Finally, a control variable was determined for each subgroup using the latent regression linear logistic test model, and item difficulties, standard errors, and significance values were calculated. The item parameters obtained for each variable group were compared.

Results

Findings Related to the First Sub-Problem

In the first sub-problem of the study, "What are the item parameters obtained from Rasch, linear logistic test, latent regression and latent regression linear logistic test models?" The answer to the question has been sought. For this purpose, difficulty parameters and standard errors of 20 test items were estimated with four models. The results are given in Table 4.

When Table 4 is examined, it is seen that the lowest difficulty parameters were estimated from the Rasch model, and the highest difficulty parameters were estimated from the LRLLTM. The average of the item difficulty values obtained from the Rasch model is -0.324.

Table 4. Item parameters and standard errors estimated from models

	Rasch		LLTM		LRM		LRLLTM	
Item	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
M1	-1.84	0.17	-1.78	0.19	-1.63	0.18	-1.58	0.19
M2	-0.46	0.15	-0.40	0.17	-0.29	0.17	-0.23	0.17
M3	-0.35	0.15	-0.29	0.17	-0.18	0.17	-0.12	0.17
M4	-0.05	0.15	0.01	0.18	0.12	0.17	0.18	0.18
M5	0.72	0.16	0.79	0.18	0.91	0.17	0.97	0.18
M6	0.06	0.15	0.13	0.18	0.24	0.17	0.29	0.18
M7	-1.14	0.15	-1.07	0.18	-0.95	0.17	-0.89	0.18
M8	-0.19	0.15	-0.13	0.18	-0.02	0.17	0.04	0.17
M9	-0.13	0.15	-0.07	0.18	0.04	0.17	0.10	0.18
M10	0.06	0.15	0.13	0.18	0.24	0.17	0.29	0.18
M11	-0.29	0.15	-0.23	0.17	-0.12	0.17	-0.06	0.17
M12	-0.24	0.15	-0.18	0.17	-0.07	0.17	-0.01	0.17
M13	-0.38	0.15	-0.32	0.17	-0.21	0.17	-0.16	0.17
M14	-0.80	0.15	-0.74	0.18	-0.62	0.17	-0.57	0.18
M15	-1.33	0.16	-1.26	0.18	-1.13	0.17	-1.08	0.18
M16	0.80	0.16	0.86	0.18	0.99	0.18	1.04	0.18
M17	0.10	0.15	0.16	0.18	0.27	0.17	0.33	0.18
M18	-0.38	0.15	-0.32	0.17	-0.21	0.17	-0.16	0.17
M19	-1.00	0.15	-0.94	0.18	-0.82	0.17	-0.76	0.18
M20	0.36	0.15	0.42	0.18	0.54	0.17	0.59	0.18

This situation is interpreted as the general difficulty of the test for a student with an average ability level. According to
the LRLLTM, the average of the item difficulty values was calculated as -0.089, which can be interpreted as the test being of medium difficulty. When the standard error values were examined, lower standard error values were obtained for all items with the Rasch model, while values close to each other were obtained in the other models. According to all models, the easiest item of the test is item 1, while item 16 is the most difficult. It can be stated that all items appear more easily when person and item variables are added to the Rasch model, which is the descriptive model.

The significance of this difference between the difficulty parameters estimated from these four models was tested with the ANOVA test for repeated measurements. In this analysis, it was seen that the Mauchly sphericity assumption was not met ($w=.164$, $p<.05$). As a result of the Greenhouse-Geisser test, it was concluded that the item difficulties estimated according to at least one model differed significantly ($p<.05$). The models were compared in pairs in order to determine from

which model the item difficulties differ, and the results are given in Table 5.

Table 5. Pairwise comparison of models

Model	Sum of Squares	Degrees of Freedom	Mean Square	F	p
Rasch vs. LLTM	.078	1	.078	3958.333	.000
LLTM vs. LRM	.271	1	.271	2509.710	.000
LRM vs. LRLLTM	.062	1	.062	2364.636	.000

When Table 5 was examined, it was concluded that the difference between the item difficulties estimated from the models was significant ($p<0.05$).

Findings Related to Second Sub-Problem

In the second sub-problem of the research, "How is the model-data fit in Rasch, LLTM, LRM and LRLLTM? The answer to the question has been sought. In order to examine the model data fit, -2LL, AIC and BIC fit values were compared. The results are given in Table 6.

Table 6. AIC, BIC and -2 LL values of models

	AIC	BIC	-2LL
RASCH	8092.900	8237.700	8050.800
LLTM	8094.400	8246.100	8050.400
LRM	7726.200	7898.600	7676.200
LRLLTM	7727.200	7906.500	7675.200

AIC and BIC values do not directly provide data on model data fit and usability of a model. These values become more meaningful by comparing the models. It is interpreted that models with lower AIC and BIC values are more compatible with the data (Blosis et al., 2007). When Table 6 is examined, it is seen that the model with the lowest AIC value is LRM and the highest one is LLTM. Similarly, the model with the lowest BIC values is LRM and the highest is LLTM. When the -2LL values were examined, the lowest value was obtained in the LRLLTM, and the highest value was obtained in the Rasch model. The smaller of the three values is interpreted as a better model-data fit. Based on these data, it can be said that LRM and LRLLTM have better model-data fit than Rasch and LLTM.

Findings Related to Third Sub-Problem

In the third sub-problem of the study, "Does the mathematics lesson performance of individuals differ in the subgroups of the variables handled with the latent regression linear logistic test model? The answer to the question has been sought. In the study, gender (male-female) and school type school type (science-vocational-anatolia-religious) variables were considered as person predictors. "Male" students for gender and "anatolian high school" for school type were determined as the control variable. Cognitive domain (knowing - applying -reasoning), content domain (number-algebra-geometry) and booklet type (A-B) variables were considered as item predictors. "Reasoning" for the cognitive domain, "algebra" for the content domain, and "booklet A" for the booklet type were determined as the control variables. Obtained item parameters, standard errors, z and p values are given in Table 7.

Table 7. Item parameters obtained with LRLLTM

Category	Estimate	Std.Error	z value	p value
Gender (Female)	-0.161	0.097	1.659	0.218
School (Science)	-3.264	0.201	16.172	0.000
School (Religious)	0.116	0.162	-0.718	0.472
School (Vocational)	1.132	0.148	-7.624	0.000
Cognitive (Knowing)	-1.278	0.125	10.165	0.000
Cognitive (Applying)	-1.033	0.113	9.121	0.000
Content (Geometry)	-0.352	0.111	3.161	0.001
Content (Number)	0.323	0.061	-5.223	0.000
Booklet(B)	-0.106	0.105	1.004	0,315

When Table 7 is examined, it can be said that the difference in item difficulties between male and female student groups according to the gender variable is not statistically significant ($p>0.05$). When the school type variable is examined, there is no statistically significant difference between Anatolian high school students and imam hatip high school students in terms of item difficulties. The average item difficulty for science high school students is -3,264. This is interpreted as an average

of 3.264 logit easier to find test items for science high school students when compared to Anatolian high school students. On average, it was found to be 1.132 logit more difficult for vocational and technical Anatolian high school students.

When a comparison is made according to the cognitive domain, it can be interpreted that the items at the knowledge level are 1.278 logit easier than the reasoning level, and the items at the application level are 1.033 logit easier than the reasoning level. When the content domain are examined, it can be stated that the items in the geometry domain are on average 0.352 logit easier than the items in the algebra domain, and the items in the number domain are on average 0.323 more difficult than the items in the algebra domain. According to the booklet type variable, there was no statistically significant difference between the groups who answered the A and B booklets in terms of average item difficulties.

Discussion, Conclusion and Recommendations

Standard item response models are insufficient to explain the differences in predicted person and item parameters and their reasons. Explanatory Item Response Models offer flexible and more comprehensive statistical approaches that allow item and person characteristics to be added to the model. In this study, an application of Explanatory Item Response Models was carried out on the data of the common mathematics course exam conducted throughout the province of Niğde. In the first stage, the item difficulties estimated from the descriptive and explanatory item response models for the items in the achievement test were examined. Minimum difficulty parameter values were estimated for all items with the Rasch model. When person and item variables were added to the model, it was observed that the difficulty parameters increased. In the LRLLTM model, in which both person and item variables were added together, the highest difficulty parameter values were obtained. The difference between the mean difficulty values obtained from the models was found to be significant. These results are similar to the studies in the literature. In the study conducted by Atar (2011) on TIMSS 2007 mathematical data, it was concluded that item difficulties differed when person and item variables were added to the Rasch model. Atar and Aktan (2013) compared the 2PL IRT model with the Latent Regression 2PL model, which was created by adding person variables to this model, in their study on TIMSS 2007 science data. As a result of adding person predictors to the model, it was found that the parameters estimated by the two models differed. Tat (2020), in his study on simulation data in different subgroups, compared the item difficulties obtained from the Rasch model and three different exploratory item response models. Although item difficulty values differed for some subgroups in the study, no significant difference was observed between the difficulties obtained from the four models.

In the second sub-problem of the research, the model-data fit values obtained from four different models were compared. In this section where AIC, BIC and -2LL values are examined, it is seen that the Rasch model has worse model-data fit than other models. It was concluded that the LRM had better model-data fit. It can be stated that the person variables (gender and school type) added to the model contribute to the increase in model data fit. These results are similar to the studies in the literature. Atar and Aktan (2013) compared the model data fit of the 2 PL model and the latent regression 2 PL model, and better model data fit values were obtained in the latent regression 2 PL model. In Tat (2020) study, it was stated that the latent regression and linear logistic test model had better model-data fit than the Rasch model.

In the third sub-problem of the research, the item parameters obtained from the LRLLTM model; gender, school type, cognitive domain, content domain and booklet type variables were discussed at the level of subgroups. There are many studies in the literature examining the relationship between gender and mathematics achievement (Yücel & Koç, 2011; Cheema & Galluzzo, 2013; Ellison & Swanson, 2018). However, it is seen that studies within the framework of explanatory item response models are limited. Atar and Aktan (2013) conducted their study with person explanatory item response models and stated that the gender variable did not

have a significant effect on explaining students' science achievement. In the study conducted by Atar (2011) using explanatory item response models, it was stated that the gender variable did not have a significant effect in explaining the differences in mathematics achievement. In this study, it was seen that the difference between the item difficulty levels did not show a statistically significant difference according to gender. When the comparisons according to school types are examined; While the items were easier for science high school students, it was seen that the items were more difficult for vocational high school students. Berberoğlu and Kalender (2005) stated in their study that in many national and international studies, student achievements differ significantly according to school types and that while science high school students achieve high success, vocational high school students achieve lower success.

When the findings obtained in terms of item predictors were examined, it was concluded that the items at the reasoning level were more difficult than the items at the knowing and applying level. These findings are similar to the results obtained by Atar (2011) in his study. According to the content domain classification, it was concluded that the items in the geometry domain were easier than the items in the algebra domain. Contrary to these results, Atar (2011) concluded that the items in the geometry domain are on average 0.39 logit more difficult than the items in the algebra domain. It is thought that the main reason for the difference in the results of the research is the low number of items in the geometry domain in this study and the item characteristics. Another variable discussed in the research is the booklet type. For the booklet type variable, it was seen that the difference between the average item difficulties obtained from the A and B booklets was not significant. This is actually an expected result. Although the items in both booklets are exactly the same, the position of the items in the test differs. These ranking differences did not differentiate the obtained item parameters according to the booklet type.

In the literature, there are also studies in which explanatory item response models were conducted with data from different courses other than mathematics. Büyükkıdık & Bulut (2022) included some of the variables in this study in their research. In the study, gender and school type were used as individual variables and content area was used as item variable. The effects of different variables and their interactions on students' responses to science questions in the exam were analyzed. As a result, it was found that female students were more likely to answer the items correctly than male students and private school students were more likely to answer the items correctly than public school students. In terms of content, biology items were found to be easier than physics items. According to these results, it is seen that even if similar variables are used in the studies, the results obtained may vary according to the courses and the data set used.

Explanatory item response models, unlike traditional models, provide the opportunity to include person and item variables in a single model. It was seen that better model data fit values were obtained from these models and the models differed item parameters. It is thought that the widespread use of these relatively new models in educational research will be beneficial. In this study, the Rasch model and explanatory item response models were examined. Comparisons of the 2PL and 3PL models with the explainer response models can be made in other studies. Studies using explanatory item response theory models can be conducted in data sets that use multi-category scored and multidimensional items. In these models, it is possible to observe the person and item interaction effect. More comprehensive studies can be conducted to examine the common effects of different person and item variables in national and international education studies. These models can also be used in DIF determination and computerized adaptive testing (CAT) studies. This study was studied with a limited research group. Considering the importance of sample size for parameters obtained with IRT, it is thought that working with larger sample groups will allow for more accurate estimation of model parameters.

References

Atar, B. (2011). Tanımlayıcı ve açıklayıcı madde tepki modellerinin TIMSS 2007 Türkiye matematik

verisine uyarlanması. *Eğitim ve Bilim*, *36*(159).

Atar, B., & Aktan, D. Ç. (2013). Birey açıklayıcı madde tepki kuramı analizi: örtük regresyon iki parametreli lojistik modeli. *Eğitim ve Bilim*, *38*(168).

Baker, F. B. (2001). The basics of item response theory. http://ericae. net/irt/baker.

Berberoğlu G. ve Kalender İ. (2005). Öğrenci Başarısının Yıllara, Okul Türlerine, Bölgelere Göre İncelenmesi: ÖSS ve PISA Analizi, ODTÜ Eğitim Bilimleri ve Uygulama Dergisi, Sayfa 27-28.

Blozis, S. A., Conger K. J., & Harring, J. R. (2007). Nonlinear latent curve models for multivariate longitudinal data. International Journal of Behavioral Development: Special Issue on Longitudinal Modeling of Developmental Processes, 31, 340-346

Boeck, P. de, Cho, S. J., & Wilson, M. (2011). Explanatory secondary dimension modeling of latent differential item functioning. *Applied Psychological Measurement*, *35*(8), 583–603.

Boeck, P. de., &Wilson, M. (2004). *Explanatory item response models*. New York, NY: Springer New York.

Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, *21*(2), 89–118.

Bulut, O. (2021). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses*, R package version 0.4. doi: 10.5281/zenodo.4556285 Available from https://CRAN.R-project.org/package=eirm.

Bulut, O., Palma, J., Rodriguez, M. C., & Stanke, L. (2015). Evaluating measurement invariance in the measurement of developmental assets in Latino English language groups across developmental stages. *Sage Open*, *5*(2), 2158244015586238.

Büyükkıdık, S., & Bulut, O. (2022). Analyzing the Effects of Test, Student, and School Predictors on Science Achievement: An Explanatory IRT Modeling Approach. *Journal of Measurement and Evaluation in Education and Psychology*, *13*(1), 40-53.

Cheema, J. R., & Galluzzo, G. (2013). Analyzing the gender gap in math achievement: Evidence from a large-scale US sample. *Research in Education*, *90*(1), 98-112.

Chen, F., Yang, H., Bulut, O., Cui, Y., & Xin, T. (2019). Examining the relation of personality factors to substance use disorder by explanatory item response modeling of DSM-5 symptoms. *PloS One*, *14*(6), e0217630. https://doi.org/10.1371/journal.pone.0217630

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*(3), 265-289.

Chiu, T. (2016). *Using Explanatory Item Response Models to Evaluate Complex Scientific Tasks Designed for the Next Generation Science Standards* (Doctoral dissertation, UC Berkeley).

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

De Ayala, R. J. (2022). *The theory and practice of item response theory, Second Edition.* Guilford Publications.

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

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

Ellison, G., & Swanson, A. (2018). *Dynamics of the gender gap in high math achievement* (No. w24910). National Bureau of Economic Research.

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

Fleiss,J.L.(1971) "Measuring nominal scale agreement among many raters." *Psychological Bulletin*, Cilt 76, Sayi 5 say. 378-382

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). How to design and evaluate research in education (8th edt.). New York: McGram-Hill Companies.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. *Measurement methods for the social sciences series.* Newbury Park, Calif.: Sage Publications.

Kahraman, N. (2014). An explanatory item response theory approach for a computer-based case

simulation test. *Eurasian Journal of Educational Research*, *14*(54), 117–134. https://doi.org/10.14689/ejer.2014.54.7

Kim, J., & Wilson, M. (2020). Polytomous item explanatory item response theory models. *Educational and Psychological Measurement*, *80*(4), 726-755.

Landis, J. R. ve Koch, G. G. (1977) "The measurement of observer agreement for categorical data", *Biometrics*. Cilt. 33, say. 159-174 Kaiser, H. F. (1960). The application of electronic computers to factor analysis. Educational and Psychological Measurement, 20(1), 141–151.

Min, H., Zickar, M., & Yankov, G. (2018). Understanding item parameters in personality scales: An explanatory item response modeling approach. *Personality and Individual Differences*, *128*, 1–6. https://doi.org/10.1016/j.paid.2018.02.012

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous Item Response Theory models. Applied Psychological Measurement, 24(1), 24-50

Petscher, Y., Compton, D. L., Steacy, L., & Kinnon, H. (2020). Past perspectives and new opportunities for the explanatory item response model. *Annals of Dyslexia*, *70*(2), 160-179.

Randall, J., Cheong, Y. F., & Engelhard, G. (2010). Using explanatory item response theory modeling to investigate context effects of differential item functioning for students with disabilities. *Educational and Psychological Measurement*, *71*(1), 129–147.

Sijtsma, K. (2020). *Measurement models for psychological attributes: Classical test theory, factor analysis, item response theory, and latent class models*. CRC Press.

Tat, O. (2020). *Açıklayıcı Madde Tepki Modellerinin Bilgisayar Ortamında Bireye Uyarlanmış Testlerde Kullanımı*. [Doktora Tezi]. Hacettepe Üniversitesi, Eğitim Bilimleri Enstitüsü, Ankara.

Yavuz, H. C. (2019). The effects of log data on students' performance. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(4), 378-390.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. Applied Psychological Measurement, 5, 245–262.

Yücel, Z., & Koç, M. (2011). İlköğretim öğrencilerinin matematik dersine karşı tutumlarının başarı düzeylerini yordama gücü ile cinsiyet arasındaki ilişki. *İlköğretim Online*, *10*(1), 133-143.