



Investigation of Variables Predicting Response Time for PISA Mathematics Items

PISA Matematik Sorularında Yanıtlama Süresini Yordayan Değişkenlerin İncelenmesi

Semirhan Gökçe^{a*}, Arzu Aydoğan Yenmez^a

^aNiğde Ömer Halisdemir University, Niğde, Turkey

Abstract

One of the most essential contributions of technological developments in educational measurement is providing an environment for administering computer-based test items. Due to its technological infrastructure, computer-based tests have some advantages over paper and pencil based tests such as fast score reporting, supporting innovative item formats and recording response time. The purpose of this study is to investigate student characteristics predicting the response time for different item types and mathematical processes in PISA 2012 mathematical literacy items. A computer-based testing environment have been developed and 26 released mathematics items were used. The participants of the study were 124 ninth grade students of a high school in the Mediterranean region. Within the scope of the study, multiple linear regression analysis was conducted to predict the response time with gender, daily computer usage, mathematics exam score, Turkish Language and Literature exam score and average response length (only for open-ended items). In the study, various models were tested regarding item type (multiple-choice and open-ended) and mathematical processes (formulate, employ and interpret). The results of the study indicated that mathematics exam score was not a significant predictor of response time in all models. Moreover, it was found that gender was a predictor of response time and male students used less time for responding to the items than female students by holding the remaining predictor variables constant.

Keywords: computer-based tests, gender, mathematics achievement, PISA, response time.

Öz

Gelişen teknolojinin eğitimde ölçme ve değerlendirme sürecine sağladığı en önemli katkılardan biri hiç şüphesiz testlerin bilgisayar ortamında uygulanabilmesidir. Kâğıt-kalem kullanılarak yanıtlanan testler ile karşılaştırıldığında katılımcıların test puanlarını sınav bitiminde öğrenebilmesi ve yenilikçi farklı soru formatlarını desteklemesi gibi üstünlükleri olan bilgisayar ortamındaki testler, sahip olduğu teknolojik altyapı sayesinde katılımcıların her sorunun yanıtlanmasında harcadığı süreyi de kaydedebilmektedir. Bu çalışmanın amacı, PISA 2012 matematik okuryazarlığı maddelerinde farklı soru türleri ve matematiksel süreçler için yanıtlama süresini yordayan öğrenci özelliklerini incelemektir. Bu amaçla, Uluslararası Öğrenci Değerlendirme Programı (PISA) 2012 uygulamasının açıklanan 26 matematik okuryazarlığı sorusu kullanılarak bilgisayar ortamında test uygulaması geliştirilmiştir. Çalışmanın katılımcılarını, Akdeniz Bölgesindeki bir lisede okuyan 124 dokuzuncu sınıf öğrencisi olmaktadır. Bu çalışma kapsamında cinsiyet, günlük bilgisayar kullanım süresi, matematik yazılı, Türk Dili ve Edebiyatı yazılı notu ve (açık uçlu sorulardaki) ortalama yanıt uzunluğu gibi değişkenler ile cevap süresini tahmin etmek için çoklu doğrusal regresyon analizi yapılmıştır. Çalışmada, madde tipi (çoktan seçmeli ve açık uçlu) ve matematiksel süreçler (formüle etme, işe koşma ve yorumlama) ile ilgili çeşitli modeller test edilmiştir. Araştırma bulguları, hiçbir modelde sınıf içi matematik sınav notunun yanıtlama süresini yordamadığını göstermiştir. Bununla birlikte, cinsiyetin yanıtlama süresini yordayan bir değişken olduğu görülmüş ve modelde yer alan diğer değişkenler sabit tutulduğunda erkeklerin kızlara göre soruların yanıtlanmasında daha az süre kullandıkları belirlenmiştir.

Anahtar Kelimeler: bilgisayar ortamında testler, cinsiyet, matematik başarısı, PISA, yanıtlama süresi

*ADDRESS FOR CORRESPONDENCE: Asst. Prof. Dr. Semirhan Gökçe, Department of Computer Education and Instructional Technology, Faculty of Education, Niğde Ömer Halisdemir University, Niğde, Turkey, E-mail address: semirhan@gmail.com, Tel: +90 (388) 225 44 02. ORCID ID: 0000-0002-4752-5598.

Assoc. Prof. Dr. Arzu Aydoğan Yenmez, Department of Mathematics and Science Education, Faculty of Education, Niğde Ömer Halisdemir University, Niğde, Turkey, E-mail address: aydogan.arzu@gmail.com, Tel: +90 (388) 225 43 53. ORCID ID: 0000-0001-8595-3262.

Received Date: November 5th, 2019. Acceptance Date: April 24th, 2020.

1. Introduction

Tests have an essential role in educational measurement. Although, paper and pencil-based tests are very popular and widely used in all over the world, the technological developments have been stimulating a slight shift to computerized administration of tests for decades. This shift is attributed to the superior characteristics of computer-based tests such as fast score reporting (Eggen, 2007; Hambleton, Swaminathan & Rogers, 1991), supporting innovative item formats (Jodoin, 2003; Sireci & Zenisky, 2006) and measuring response time (Schnipke, 1995). In order to plan a computerized administration, a well-designed software is required. This computer software should have a user friendly environment which heeds on the representation of test items, switching between items, termination of the test, score reporting and saving the response time.

In educational tests, the performances of the participants are more or less affected by time limits (Lu & Sireci, 2007). In case of a limited time, a group of participants could not fully consider some of the items. In this situation, these tests are called speed tests. The purpose of speed tests is to measure how quickly the participants can respond to the items (Schnipke, 1995). On the contrary, the tests that are usually more difficult and in which the participants are given enough time to attempt all the items are called power tests (Schnipke, 1995; Lu & Sireci, 2007). In power tests, the difference between two scores is attributed to the difference in performance. Being on two ends, these two tests denotes that the testing time is important in educational measurement and also it depends on the purpose of the test.

Computer-based tests are able to record the response time of the participants for each item. On the one hand, response time provides information about the participants' item-based efforts in the test. On the other hand, the investigation of response time helps both to determine the optimum time for the test and to reveal the participants who have rapid-guessing behavior. Recent studies on response time focuses on determination of differential item functioning (DeMars & Wise, 2010; Hamilton, 1999), expression of response time as a measure of participants' motivation and effort (DeMars, Bashkov & Socha, 2013; Wise & DeMars, 2006; Wise & Kong, 2005), examining the relationship between response time and test scores (Hornke, 2000), identifying aberrant response-time patterns in the tests (van der Linden & van Krimpen-Stoop, 2003; van der Linden & Guo, 2008), identifying cheat attempts (van der Linden, 2009) and the determination of item selection in computerized adaptive tests (van der Linden, 2008).

Several large scale assessments have recently been administered as computer-based. Until 2015, paper and pencil test was the only option for the Programme for International Student Assessment (PISA) conducted by the Organization for Economic Co-operation and Development (OECD) but since then the tests were also delivered via computer (OECD, 2017). PISA is a three-year study and evaluates the knowledge and skills acquired by 15-year-old students. In this study, released mathematical literacy items of PISA 2012 applications were used. The mathematical literacy in PISA practices is defined as the capacity of the individual to formulate, use and interpret mathematics in various contexts of daily life, and this capacity consists of using mathematical concepts, operations and tools to reason, explain and predict a phenomenon mathematically (OECD, 2010). In PISA, mathematics proficiency is defined as the capacity of individuals to formulate, employ and interpret mathematics in a variety of contexts. The term describes the capacities of individuals to reason mathematically and use mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena (OECD, 2014). Mathematical processes have been classified as: (1) how effectively students are able to recognize and identify opportunities to use mathematics in problem situations and then provide the necessary mathematical structure needed to formulate that contextualized problem into a mathematical form, (2) how well students are able to perform computations and manipulations and apply the concepts and facts that they know to arrive at a mathematical solution to a problem formulated mathematically and (3) how effectively students are able to reflect upon mathematical solutions or conclusions, interpret them in the context of a real-world problem, and determine whether the results or conclusions are reasonable (OECD, 2013). These three categories are abbreviated as -formulate, -employ and -interpret, respectively. Each PISA mathematics item is labeled in either of these mathematical processes.

While taking the computer-based test, there are a list of factors affecting performance. Harrison and Rainer (1992) indicated that personal and background characteristics such as gender, age, preceding computer experience, concerns and attitudes towards computers were closely related to computer skills. Especially, the participants having advanced computer skills adapt the computerized testing environment more easily. Also, related studies investigated gender differences in computer usage (Schumacher & Morahan-Martin, 2001), computer competence (Corston & Colman, 1996), interest towards computers (Krendl, Broihier & Fleethood, 1989; Stoilescu & Egodawatte, 2010)

and purpose of using computers (Wilson, 2004). Furthermore, there are studies discussing the variables influencing gender differences such as item response format, item design characteristics, and time limit (Arendasy & Sommer, 2010; Voyer & Doyle, 2010).

This study focuses on the factors affecting the response time of the participants for different item types and mathematical processes. Analyzing multiple-choice and open-ended item types together with formulating, employing and interpreting mathematical processes in response time regression models, were expected to provide information about the participant characteristics predicting response time in computer-based PISA mathematics assessments. Moreover, investigating the response time on the items measuring different mathematical processes would provide evidence for the validity of PISA 2012 mathematical literacy test.

The purpose of this study is to investigate student characteristics predicting the response time for different item types and mathematical processes in PISA 2012 mathematical literacy items. The questions sought in the scope of the research are given below.

1. Which variables are involved in the model predicting the response time of the students with respect to multiple-choice and open ended item types in PISA mathematical literacy assessment?
2. Which variables are involved in the model predicting the response time of the students with respect to different mathematical processes in PISA mathematical literacy assessment?

2. Method

In this study, we investigated student characteristics predicting the response time for different item types and mathematical processes in PISA 2012 mathematical literacy items. This study was based on the relational model. In accordance with this model, the existence and degree of the relationships between dependent and independent variables are tried to be revealed (Crano & Brewer, 2002). In the first group of analysis, we focused on item type and investigated the variables predicting response time in multiple-choice and open-ended items. In the second group of analysis, we focused on mathematical processes and investigated the variables predicting response time in formulating, employing and interpreting items. In the analysis, the response time was the dependent variable and gender, daily computer usage, mathematics exam score, Turkish language and literature (TLL) exam score and average response length (only for open-ended items) were the predictor variables.

2.1. Participants

PISA focuses on assessing the extent to which 15-year-old students have acquired key knowledge and skills. Thus, with considering this information the participants of this study consisted of 124 ninth grade students attending to a public school in the Mediterranean Region. The participants voluntarily attended to the study. Their school is located in the city center and the socio-economic status of these participants is about average.

Each participant had sufficient time to complete the test. Response time analysis indicated that three students completed the test by marking and typing randomly and two students closed the application directly without seeing any questions in the test. Therefore, the analysis was conducted based on 119 students (59 females and 60 males).

2.2. Data collection tools

A questionnaire was administered to the participants in order to collect information about the predictor variables. After this questionnaire, a computer-based test was administered. The testing environment was developed by using C# programming language and the test covered the released items of PISA 2012 mathematical literacy test. Item information table regarding item description, item types and mathematical processes are given in Table 1.

Table 1
Item information table of released PISA 2012 mathematics items

Item description	Item type	Mathematical process
Apartment purchase	open-ended	formulate
Drip rate (item 1)	open-ended	employ
Drip rate (item 2)	open-ended	employ
Charts (item 1)	multiple-choice	interpret
Charts (item 2)	multiple-choice	interpret
Charts (item 3)	multiple-choice	employ
Sailing ships (item 1)	multiple-choice	employ

Sailing ships (item 2)	multiple-choice	employ
Sailing ships (item 3)	open-ended	formulate
Sauce	open-ended	formulate
Ferris wheel (item 1)	open-ended	employ
Ferris wheel (item 2)	multiple-choice	formulate
Climbing Mount Fuji (item 1)	multiple-choice	formulate
Climbing Mount Fuji (item 2)	open-ended	formulate
Climbing Mount Fuji (item 3)	open-ended	employ
Helen the cyclist (item 1)	multiple-choice	employ
Helen the cyclist (item 2)	multiple-choice	employ
Helen the cyclist (item 3)	open-ended	employ
Which car? (item 1)	multiple-choice	interpret
Which car? (item 2)	multiple-choice	employ
Which car? (item 3)	open-ended	employ
Garage (item 1)	multiple-choice	interpret
Garage (item 2)	open-ended	employ
Revolving door (item 1)	open-ended	employ
Revolving door (item 2)	open-ended	formulate
Revolving door (item 3)	multiple-choice	formulate

As shown in Table 1, there are 13 multiple-choice and 13 open-ended items in the test. As the items are grouped based on the mathematical processes, formulating has 8 items (3 multiple-choice and 5 open-ended), employing has 14 items (6 multiple-choice and 8 open-ended) and 4 items measure interpreting skills (4 multiple-choice but no open-ended). Sample screenshots from the test environment for both multiple-choice and open-ended items in are given in Figure 1 and Figure 2, respectively.

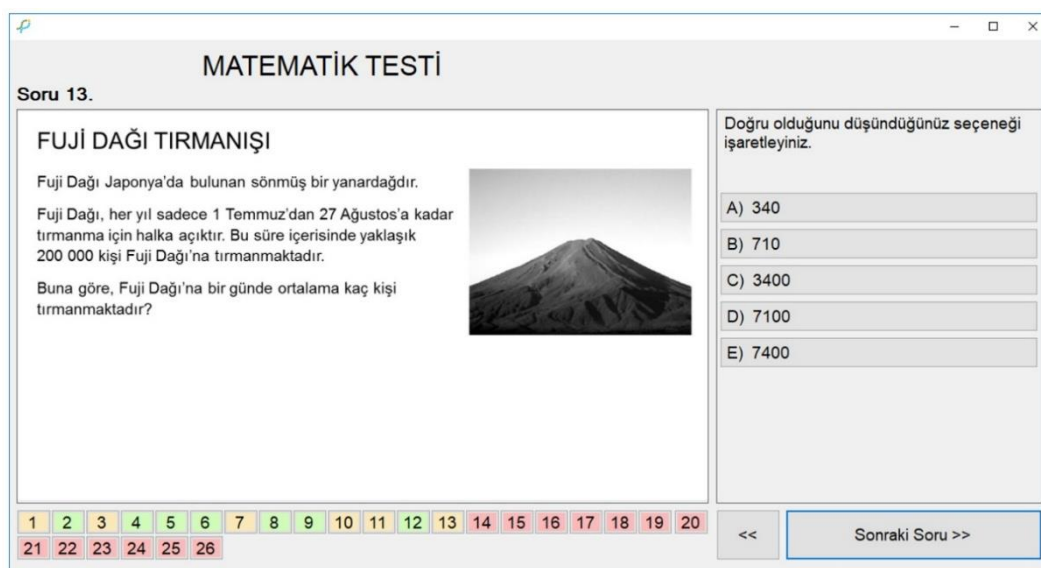


Figure 1. A screenshot of a multiple-choice item that measures formulating process in the computer-based test

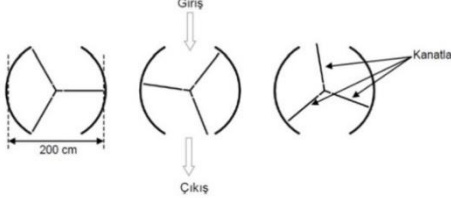
MATEMATİK TESTİ

Soru 24.

DÖNER KAPI

Bir döner kapının, daire şeklinde bir alan içerisinde dönen üç kanadı vardır. Bu alanın iç çapı 2 metre (200 santimetre)'dir. Üç kapı kanadı, bu alanı üç eşit bölüme ayırmaktadır.

Yandaki plan, yukarıdan bakıldığında bu üç kapı kanadının üç farklı konumunu göstermektedir.



İki kapı kanadı arasındaki açı kaç derecedir?

Yanıtınızı aşağıda boş bırakılan alana giriniz.

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
21 22 23 24 25 26

<< Sonraki Soru >>

Figure 2. A screenshot of an open-ended item that measures employing process in the computer-based test

As shown in Figure 1 and 2, a navigation menu located at the bottom of the application was used for facilitating the transition between items. In the navigation menu, different colors were used to indicate the answered, missing, not reached items. Here, not reached items were indicated as red, missing items were indicated in yellow and the answered (the response could either be correct or incorrect) indicated in green. Thus, if a participant does not give any response to an item, this property will provide an opportunity to review the item just by clicking item number. Otherwise, it is not possible to keep the correct response time records for each item. During test administration a timer recorded the response time and when a participant returned to an item, the timer updated the response time, accordingly. At the end of the test, item level and total response times of the participants were reached.

After test administration, the scoring of the responses was carried out by using the PISA scoring rubric. This document includes sample responses for ease of scoring. The responses to multiple-choice items were automatically scored by the computer but the scoring of responses to open-ended items were completed by two faculty members specialized in mathematics education. The inter-rater reliability between these scorers was found as .996. The responses having inconsistent scoring of the scorers were analyzed and these responses were inserted to the key as sample responses. After scoring all the responses, the total scores of the participants were calculated.

2.3. Data analysis

As mentioned previously, we investigated the participant characteristics predicting the response time for different item types (multiple-choice and open-ended items) and mathematical processes (formulating, employing and interpreting) in PISA 2012 mathematical literacy items. Response time variable stands for the average time (in terms of seconds) spent on the items. For gender, 0 represented female students and 1 represented male students. Daily computer usage indicated the participants' average time (hours) spent on desktops, laptops and/or tablets. Mathematics exam score and TLL exam score variables were out of 100 and stands for the participants' written exam scores held in the classroom. Average response length in open-ended items referred to the average number of characters used in these type of items.

The regression analysis was used to model the relationship between two or more predictor variables (student characteristics) and a dependent variable Y (response time) by fitting a linear equation. In a multiple linear regression, there are more than one predictors so the equation could be stated as,

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \dots + \beta_k \cdot X_k + e_i$$

where β_0 is the regression constant, $\beta_1, \beta_2, \dots, \beta_k$ are the parameters to be estimated and e_i is the error of prediction (Stevens, 2002). This equation indicates the change in the dependent variable Y with one-unit change in either of the predictor variable X_i for $i=1, 2, \dots, k$ while holding the remaining predictor variables constant (Tabachnick & Fidell, 2007).

The multiple linear regression analysis has some assumptions. First of all, errors of prediction are assumed to be independent with constant variance and normally distributed with a mean value of 0 (Stevens, 2002). For this assumption, Durbin-Watson statistics were calculated for multiple-choice and open-ended response times as 2.160 and 2.268, respectively. When these values were close to 2, this indicates that error terms were independent of each other. In order to test normality, standardized residuals of the response time were calculated. The histogram of these residuals and Q-Q plots indicated fair approximation to normal distribution. Another assumption is homoscedasticity which refers to the variances of the predictions determined by regression remain constant (Knaub, 2007). For homoscedasticity, the scatter plot of unstandardized predicted values versus studentized residuals was created for each model. Since the points were almost equally distributed in the plots, we could say that homoscedasticity assumption was met. Another assumption of multiple linear regression is linearity. This term focuses on the linear relationship which exists between dependent variable and each predictor variable. In order to observe the relationship, normal P-P plots of regression standardized residuals were created. There was linearity as assessed by partial regression plots in both multiple-choice and open-ended regression models and plots of studentized residuals against the predicted values. On the other hand, if there were moderate to high inter-correlations among predictor variables then the multi-collinearity problem would occur. Hence, the correlations among predictor variables were calculated and the values were presented in Table 2.

Table 2
Correlation coefficients among predictor variables

	Daily computer usage	Math exam score	TLL exam score	Avg. response length*
Daily computer usage	1			
Math exam score	-.029	1		
TLL exam score	-.097	.212**	1	
Avg. response length*	-.126	-.012	.210**	1

* Average response length exists for only open-ended items

** Correlation is significant at .05 level.

In Table 2, five of the correlations were not significant. The remaining comparisons were significant but the correlations were low among these variables. For multi-collinearity assumption, the variance inflation factor (VIF) which is the quantity $1/(1-R_j^2)$ should not exceed 10 (Hair et al., 2010; Myers, 1990). Here, R_j is the squared multiple correlation for predicting the j th predictor from all other predictor variables (Stevens, 2002). In any case, small VIF values indicates low correlation among variables under ideal conditions. In the regression analysis, VIF values were calculated and all of them were about 1.00. Additionally, the tolerance value of collinearity statistics lower than .10 indicates a potential problem of multi-collinearity (Hair et al., 2010). It was seen that, the tolerance values were higher than the critical value (located between .89 and .95). Thus, all these findings showed that the multi-collinearity assumption was met. Additionally, the outlier analysis was performed and Cook's distance, a method used to determine the effect of data on regression models, were calculated. When the Cook's distance value is greater than .85, the relevant data is defined as an outlier in the regression analysis for the models with three or more independent variables, (McDonald, 2002). Cook's distance statistics for outlier analysis of both item types and mathematical processes are given in Table 3.

Table 3
Cook's distance statistics

Category	Sub-category	Minimum	Maximum	Mean	SD
Item type	Multiple-choice	.000	.153	.010	.022
	Open-ended	.000	.136	.009	.019
Mathematical process	Formulating	.000	.066	.009	.013
	Employing	.000	.069	.009	.014
	Interpreting	.000	.063	.009	.012

Table 3 gives hints about the absence of outliers in the data predicting response times of tests with both item types and mathematical processes.

In multiple linear regression analysis, stepwise method was used to build the models. This method determines the best option by adding and removing the predictor variables. The variables are chosen based on the estimated coefficients. At each stage a test is made to determine the least useful predictor and in each step the importance of each predictor is reassessed (Stevens, 2002).

3. Findings

The findings of the multiple linear regression analysis were discussed within the scope item type and mathematical processes. Since there were both multiple-choice and open-ended items in formulation and employment processes but only multiple-choice items in interpretation process, it was not convenient to build and test the regression models for each mathematical process. In other words, response length predictor variable should be included in formulation and employment processes since they have open-ended items but should not be included in interpretation process since there was no open-ended item in this process. As a solution, we separated response time based on both mathematical process and item type. There would be two models for item type (one for multiple-choice and the other for open-ended) and five models for mathematical processes (one for formulating in multiple-choice, one for formulating in open-ended, one for employing in multiple-choice, one for employing in open-ended and finally one for interpreting in multiple-choice items). Since there were no open-ended items for interpreting process, this model was out of scope. As stated earlier, response time was the dependent variable whereas gender, math exam score, TLL exam score and daily computer usage were the predictor variables in each model. Additionally, the average response length variable was included into the models using open-ended items.

3.1. Models predicting the response time regarding item types

The results of the regression models predicting the response time of multiple-choice and open-ended items are presented in Table 4 below.

Table 4
Model summary predicting response time regarding item type

Item type	Variables	Unstandardized Coefficients		Standardized Coefficients Beta	p
		B	Std. Error		
Multiple-choice	- (Constant)	137.299	9.845		.000
	- daily computer usage	-6.664	1.389	-.391	.000
	- gender	-10.452	3.195	-.272	.001
	- TLL exam score	-.346	.121	-.231	.005
Open-ended	- (Constant)	168.472	15.019		.000
	- avg. response length	.691	.089	.541	.000
	- daily computer usage	-10.080	2.133	-.333	.000
	- TLL exam score	-.504	.188	-.189	.009
	- gender	-11.924	4.880	-.175	.016

As shown in Table 4, except math exam score all the remaining variables are significant predictors of response time for the models regarding item type. For the regression model predicting response time in multiple-choice items, the adjusted R-square was calculated as .273 indicating that 27.3% of the variance on response time was explained by the predictor variables included in the model. F value was statistically significant with $F(1,115)=8.211$ and $p=.005$. Since 0 represents female and 1 represents male, male students used 10.452 seconds less time to respond each multiple-choice item than female students by holding the remaining predictor variables constant.

For the regression model predicting response time in open-ended items, the adjusted R-square was found as .464. This value showed that 46.4% of the variance on response time was explained by the predictor variables included in the model. F value was statistically significant with $F(1,114)=5.971$ and $p=.016$. The average response length, daily computer usage, TLL exam score and gender variables entered the model but again the math exam score variable was excluded. As far as the gender was concerned, a male student used 11.924 seconds less to respond an open-ended item than a female student by holding the remaining predictor variables constant.

3.2. Models predicting the response time regarding mathematical processes

In the computer based test, there are both multiple-choice and open-ended items measuring formulation and employing processes. However, there are only multiple-choice items for measuring interpreting process. So, it is not possible to predict response time for open-ended items in interpreting process. Model summary predicting the response time for formulating, employing and interpreting mathematical processes are given in Table 5.

Table 5
 Model summary predicting response time regarding mathematical processes

Mathematical process	Item type	Variables	Unstandardized Coefficients		Standardized Coefficients Beta	p
			B	Std. Error		
Formulating	Multiple-choice	- (Constant)	148.836	15.189		.000
		- daily computer usage	-7.673	2.142	-.311	.001
		- gender	-13.343	4.930	-.240	.008
		- TLL exam score	-.381	.186	-.175	.043
	Open-ended	- (Constant)	176.251	19.594		.000
		- avg. response length	.713	.116	.482	.000
		- TLL exam score	-.694	.246	-.225	.006
		- daily computer usage	-7.344	2.782	-.210	.009
Employing	Multiple-choice	- (Constant)	190.010	17.486		.000
		- daily computer usage	-8.765	2.466	-.308	.001
		- TLL exam score	-.678	.215	-.271	.002
		- gender	-12.446	5.675	-.194	.030
	Open-ended	- (Constant)	165.368	17.307		.000
		- avg. response length	.696	.102	.502	.000
		- daily computer usage	-9.832	2.458	-.299	.000
		- gender	-13.670	5.623	-.184	.017
Interpreting	Multiple-choice	- TLL exam score	-.455	.217	-.157	.038
		- (Constant)	62.990	2.658		.000
		- daily computer usage	-3.093	1.153	-.243	.008
		- gender	-5.317	2.599	-.185	.043

Table 5 provides coefficients predicting response time in mathematical processes. In these regression models, math exam score was not a significant predictor of response time for mathematical processes. On the other hand, the remaining predictor variables existed in at least one of the models. First of all, daily computer usage has negative coefficients in all models. This means that by holding the remaining predictor variables constant, students who spend more time on computers in daily life are expected to use less time in answering the items in this computer-based test. Similarly, gender has also negative coefficients in all the models. This implies that male students use less time than females provided that the remaining predictor variables are constant. Another predictor variable is the TLL exam score and this variable entered all the models except interpreting process with negative coefficients. These coefficients indicated that by holding the other predictor variables constant, students having higher TLL exam scores spend less time on the formulating and employing processes. As stated earlier, average response length is a predictor variable for open-ended items and entered to the models predicting response time in formulating and employing processes. The adjusted R-square and F statistics of the models are given in Table 6.

Table 6
 Explained variance statistics in the models of mathematical process

Mathematical process	Item type	Adjusted R-square	F	df1	df2	p
Formulating	Multiple-choice	.174	4.169	1	115	.043
	Open-ended	.319	5.663	1	114	.019
Employing	Multiple-choice	.177	4.809	1	115	.030
	Open-ended	.417	4.390	1	114	.038
Interpreting	Multiple-choice	.102	4.184	1	116	.043

The adjusted R-square values given in Table 6 shows the percentage of variation explained by the predictor variables affecting explaining the dependent variable. For example, for the open-ended items measuring employing process 41.7% of the variance was explained by average response length, daily computer usage, gender and TLL exam score variables.

4. Conclusion and Discussion

The purpose of this study is to investigate student characteristics predicting the response time for different item types and processes in PISA 2012 mathematical literacy items. It was found that except mathematics exam score, all the remaining variables are significant predictors of response time for the models regarding item type. The reason behind the inexistence of math exam score variable in the models could be attributed to the differentiation of cognitive processes measured by PISA assessments and classroom tests. The main purpose of PISA is not the extent to which the students learn the topics covered in their national curriculum (EARGED, 2005). In fact, PISA have been conducted to measure the knowledge and skills required for effective participation in social life rather than the level of access to school curricula (Berberoğlu & Kalender, 2005). The studies analyzing the classroom assessments and large scale assessments indicated that the classroom mathematics exam questions and PISA mathematics literacy items were measuring different skills (Güler, Özdemir & Dikici, 2012; Karaman & Bindak, 2017).

In the models regarding item type, TLL exam score was found as a predictor of response time and had negative coefficients. In other words, the students having higher TLL exam scores used less time to provide their responses. This result could be explained by the reading load (number of words at the item root) of PISA mathematics items requiring reading comprehension skills and the interpretation of tables and figures. Reading comprehension is known as the effective and meaningful integration of the existing knowledge with the information provided in the text. Paris and Stahl (2005) stated reading comprehension is a process involving the interpretation of the information in the text, the use of background information, and the structuring of a compatible presentation on the subject's memory. In addition, reading was defined as an activity that accompanies the resources of the processor with limited capacity including the combination of perceptual and cognitive processes based on mutual interaction with the goal of meaningful message (Chang, 2003). In the related literature, there were studies pointing out the relationship between reading comprehension and academic achievement. According to a study in which Bloom participated, four achievement tests in reading comprehension, language-literature, science, and mathematics were administered to secondary school students from 15 countries, in their mother tongue. The results of the study indicated that the correlation coefficient between reading comprehension scores and the remaining three scores (language-literature, science and mathematics scores) is quite high (Bloom, as cited in Nas, 2003, p. 136). In another study, it was shown that the academic achievement of the students with advanced reading proficiencies were significantly higher than the students with intermediate and basic reading proficiencies. Besides, the academic achievement of the students with intermediate reading proficiencies were significantly higher than the students with basic reading proficiencies (Ateş, 2008). In our study, the average TLL exam scores of female and male students were 76.3 and 70.6, respectively. Therefore, female students could employ their language skills in comprehending items and express their responses better in open-ended items.

Daily computer usage and gender were the other student characteristics predicting response time in the models regarding item type. In both models, daily computer usage had negative coefficients which indicated that the participants spending more time on desktops, laptops and tablets need less time for responding items by holding the remaining predictor variables constant. This result could be attributed to the familiarity of using technological devices such as the fluent use of the keyboard and mouse that reduces the response time. The average values of daily computer usage were 1.71 hour for female students and 2.31 hour for male students which supports our interpretation.

The results of the study showed that the response times differed in terms of gender. In order to justify the difference, the response times of female and male students were compared for both item types. Accordingly, it was decided that the comparison could be made more accurately in multiple-choice items due to the lack of response length variable. It was determined that the average response times of female and male students in multiple-choice items were 99 and 87 seconds, respectively. Since the response length variable was not included in this model, this difference could be attributed to female students' spending more time on thinking processes. Afterwards, response time values of female and male students were calculated for open-ended items and found as 134 seconds for females and 118 seconds for males. The greater difference between the average response time in open-ended items could be explained by the typing speed. The related studies showed that that computer use differs in terms of gender. In other words, male and female students differed as long as usage time (Schumacher & Morahan-Martin, 2001), using speed (Corston & Colman, 1996), interest towards computers (Krendl, Broihier & Fleetwood, 1989; Stoilescu &

Egodawatte, 2010) were concerned. Since, in our findings females spent more time than males, this difference could also be addressed to rapid guessing behavior (DeMars, Bashkov & Socha, 2013).

The amount of variance on response time explained by the predictor variables were higher in open-ended items than multiple-choice items. This can be explained by the fact that the response length, which is one of the leading factors affecting the response time, enters the model. The average response length, i.e. number of characters (letter, number, or symbol) in the response, has positive coefficient. So, an increase in the response length was expected to increase the predicted response time. When the gender differences in average response length was concerned, female students typed more than male students. For the open ended items, female students typed 406 characters and male students typed 385 characters. It could be said that this situation had a reducing effect on response time in favor of male students.

PISA defines mathematical thinking processes in three main dimensions: formulate, employ and interpret. Formulation is defined as the ability to transform the situations that students encounter in daily life into mathematical expressions. The ability to formulate relates to basic knowledge and skills related to analyzing, installing and solving a problem. For example, using a mathematical language to transform a problem case into a mathematical notation is measured under this dimension. Employment is defined as the ability to use mathematical concepts, operations and reasoning skills in solving mathematically put forward situations. It relates to the skills of employing, performing arithmetic operations, solving equations, making symbolic operations, reading tables and graphics, and analyzing data. For example, making generalizations based on the mathematical operations and processes used in finding solutions are measured within this dimension. Interpretation is defined as the ability to interpret a problem that has been formulated and solved by transferring it to a real life situation. For example, the explanation is measured within this dimension because the mathematical result obtained in a problem and the decision made are logical or unreasonable (OECD, 2013). In this study, regression models were also designed and tested based on the mathematical processes data. In all the formulating and employing models, all the predictor variables entered to the model except the mathematical exam score. This finding was quite similar to those in the models regarding item type. A remarkable point in these analysis was the predictor variables entering the model for interpreting process in multiple-choice items. Only daily computer usage and gender entered to this model but the TLL exam score was excluded. It was seen that the item difficulties of such items is quite low, indicating that these items were easier (OECD, 2014; Stacey, 2015). This means that most of the participants gave correct answer to these items. The current research data supports this finding. Additionally, the easy items would bring forth the simplicity of understanding what is going on within the items. Because of these reasons, TLL exam score might not be a significant predictor of response time.

When the average response times per item in mathematical processes are taken into consideration, the participants spent about an average of 54 seconds in interpreting items which is definitely lower compared to the other processes. The average time spent on open-ended items measuring formulating and employing processes are about 125 and 127 seconds, respectively. On the other hand, the average response time of multiple-choice items are about 99 seconds in formulating and 116 seconds in employing. The differentiation between the average response times in each process could also be attributed to the item difficulties (OECD 2014; Stacey, 2015). When evaluated from this perspective, the outcomes of this study about response time contribute to the validity of PISA 2012 mathematics assessments. More specifically, the items developed for formulation and employment processes have been measuring higher order skills compared to interpretation. Also, the average response time for these processes provided coherent results.

This study focused on the prediction of the response time, which cannot be addressed in paper-and-pencil based tests. In terms of educational practice, the findings of this research seem to be important in terms of observing both the efforts of female and male students and determining variables predicting response times in computer-based tests. The studies investigating the response time gives important feedbacks about determining the optimum testing time per item. For further studies, some item characteristics such as the number of words, the existence of illustrations and tables in the items and even the item parameters could be used for predicting the response time for the computer-based tests.

References

- Arendasy, M., & Sommer, M. (2010). Evaluating the contribution of different item design features to the effect size of the gender difference in three-dimensional mental rotation using automatic item generation. *Intelligence*, 38, 574-581.

- Ateş, M. (2008). İlköğretim ikinci kademe öğrencilerinin okuduğunu anlama düzeyleri ile Türkçe dersine karşı tutumları ve akademik başarıları arasındaki ilişki. Doctoral dissertation, Selçuk Üniversitesi Sosyal Bilimler Enstitüsü.
- Berberoğlu, G., & Kalender, İ. (2005). Öğrenci başarısının yıllara, okul türlerine, bölgelere göre incelenmesi: ÖSS ve PISA analizi. *Journal of Educational Sciences & Practices*, 4(7), 21-35.
- Chang, H.S. (2003). Difficulties in Studying and Teaching Literature Survey Courses- in English Departments in Taiwan, Master's Thesis. <http://repositories.lib.utexas.edu/bitstream/handle/2152/489/changh036.pdf?sequence=2>
- Corston, R., & Colman, A. M. (1996). Gender and social facilitation effects on computer competence and attitudes toward computers. *Journal of Educational Computing Research*, 14(2), 171-183.
- Crano, W.D., ve Brewer, M.B. (2002). Principles and methods of social research. New Jersey, Lawrence Erlbaum Associates Publishers.
- DeMars, C. E., & Wise, S. L. (2010). Can differential rapid-guessing behavior lead to differential item functioning? *International Journal of Testing*, 10(3), 207-229.
- DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under Low-Stakes Conditions. *Research & Practice in Assessment*, 8, 69-82.
- EARGED (2005). PISA 2003 projesi ulusal nihai rapor. [Çevrimiçi: http://pisa.meb.gov.tr/?page_id=22, Erişim tarihi: 11.03.2020]
- Eggen, T. J. H. M. (2007). Choices in CAT models in the context of educational testing. In Proceedings of the 2007 GMAC conference on computerized adaptive testing.
- Güler, G., Özdemir, E., & Dikici, R. (2012). İlköğretim matematik öğretmenlerinin sınav soruları ile SBS matematik sorularının Bloom taksonomisi'ne göre karşılaştırmalı analizi. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi*, 14(1), 41-60.
- Hair, J., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis (7th ed.)*. Upper saddle River, New Jersey: Pearson Education International.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.
- Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied measurement in Education*, 12(3), 211-235.
- Harrison, A. W., & Rainer, R. K. (1992). The influence of individual differences on skill in end-user computing. *Journal of Management Information Systems*, 9(1), 93-111.
- Hornke, L. F. (2000). Item response times in computerized adaptive testing. *Psicologica*, 21(1), 175-189.
- Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, 40(1), 1-15.
- Karaman, M., & Bindak, R. (2017). İlköğretim matematik öğretmenlerinin sınav soruları ile TEOG matematik sorularının yenilenmiş Bloom Taksonomisine göre analizi. *Current Research in Education*, 3(2), 51-65.
- Knaub, J. (2007). Heteroscedasticity and homoscedasticity. *Encyclopedia of measurement and statistics*, 431-432.
- Krendl, K. A., Broihier, M. C., & Fleetwood, C. (1989). Children and computers: Do sex-related differences persist? *Journal of Communication*, 39(3), 85-93.
- Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, 26(4), 29-37.
- McDonald, B. (2002). A teaching note on Cook's distance-a guideline, *Research Letters in the Information and Mathematical Sciences*, 3, 122-128.
- Myers, R.H. (1990). *Classical and modern regression with applications*. PWS-Kent Publishing, Boston.
- Nas, R. (2003). *Türkçe Öğretimi*. Bursa: Ezgi Kitabevi.
- OECD (2010). PISA 2012 Mathematics framework. OECD Publishing.
- OECD (2013). PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy. OECD Publishing.
- OECD (2014). PISA 2012 Results: What Students Know and Can Do (Volume I: Student Performance in Mathematics, Reading and Science). OECD Publishing.
- OECD (2017). PISA 2015 Results: Excellence and Equity in Education. OECD.
- Paris, S. G., & Stahl, S. A. (2005). *Children's reading comprehension and assessment*. Routledge.
- Schnipke, D. L. (1995). Assessing Speededness in Computer-Based Tests Using Item Response Times. A paper presented at the annual meeting of the National Council on Measurement in Education April, 1995, San Francisco, CA.
- Schumacher, P., & Morahan-Martin, J. (2001). Gender, Internet and computer attitudes and experiences. *Computers in Human Behavior*, 17(1), 95-110.

- Sireci, S. G., & Zenisky, A. L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. *Handbook of Test Development*, 329-347.
- Stacey, K. (2015). The international assessment of mathematical literacy: PISA 2012 framework and items. In selected regular lectures from the 12th International Congress on Mathematical Education (pp. 771-790). Springer, Cham.
- Stevens, J. (2002). *Applied multivariate statistics for the social sciences*. Lawrence Erlbaum Associates. Mahwah, NJ.
- Stoilescu, D., & Egodawatte, G. (2010). Gender differences in the use of computers, programming, and peer interactions in computer science classrooms. *Computer Science Education*, 20(4), 283-300.
- Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics (7th ed.)*. Boston, MA: Pearson.
- van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, 33(1), 5-20.
- van der Linden, W. J. (2009). A bivariate lognormal response-time model for the detection of collusion between test takers. *Journal of Educational and Behavioral Statistics*, 34(3), 378-394.
- van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, 73(3), 365-384.
- van der Linden, W. J., & van Krimpen-Stoop, E. M. (2003). Using response times to detect aberrant responses in computerized adaptive testing. *Psychometrika*, 68(2), 251-265.
- Voyer, D., & Doyle, R. A. (2010). Item type and gender differences on the mental rotations test. *Learning and Individual Differences*, 20, 469-472.
- Wilson, B. C. (2004). A study of learning environments associated with computer courses: Can we teach them better? *Journal of Computing Sciences in Colleges*, 20(2), 267-273.
- Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, 43(1), 19-38.
- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18(2), 163-183.