ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Araştırma Makalesi (Research Article); 2022; 13(1); 23-39

Examining Students' Item Response Times in eTIMSS According to their Proficiency Levels, Self-Confidence and Item Characteristics *

Seher YALÇIN **

Abstract

The aim of this study was to examine whether the time spent answering science and mathematics items by Turkish students participating in the Trends in International Mathematics and Science Study (TIMSS) at the 8th grade level showed a significant difference according to their proficiency levels, self-confidence, and the item characteristics. This study was correlational research to explore the relationship between the variables discussed. A total of 577 students who participated in the TIMSS 2019 study at the 8th grade level in Turkey and answered the common 24 (11 mathematics and 13 science) items in Booklets 1 and 2 constituted the study participants. In the data analysis, the Kruskal Wallis-H test, Mann-Whitney U test, and Latent Class Analysis were used. As a result, it was determined that the type of item and cognitive level had a significant relation to the item response times of students. The students were found to spend more time on open-ended items than multiple-choice items. On the other hand, the time spent on items in the applying level was significantly higher than the knowledge level. However, there was no significant difference between the time spent answering items in the applying level and reasoning level. It was observed that if the students' confidence level in science was high, the rate of correct answers was high, and they answered the items in a short amount of time. Students who were somewhat self-confident in mathematics were more successful in difficult mathematics items and spent less time answering the items.

Keywords: TIMSS, response time, proficiency level, item difficulty, cognitive level, self-confidence

Introduction

The International Mathematics and Science Study (TIMSS) is conducted every four years to measure the mathematics and science achievement of 4th and 8th grade students, with the first study occurring in 1995. In the most recent TIMSS 2019 study, a transition to computer-based assessment (eTIMSS) was made. Fifty-eight countries at the 4th grade level and 39 countries at the 8th grade level participated in the TIMSS 2019 study (Mullis et al., 2020). In Turkey, the 2019 study was carried out with the participation of 4077 students in 181 schools at the 8th grade level (Ministry of National Education-MoNE, 2020). Turkish students performed at the TIMSS midpoint level by obtaining 496 points in mathematics at the 8th grade level. The rate of students with advanced mathematics proficiency is 12%; however, 20% of the students in the Turkish sample cannot reach lower mathematics proficiency. Thus, Turkey ranks 15th out of 39 countries with an average score of 515 in science at the 8th grade level, and 13% of students have advanced science proficiency. However, 12% of the 8th grade students cannot reach the lower science proficiency level (MoNE, 2020).

It is indicated in the TIMSS results that there are students in two extreme groups, very successful and unsuccessful, and as a result, the reasons for their success and failure should be discussed. The results of international assessment studies can be used to develop policies in education as well as to reveal the reasons for students' failure and the factors that affect their success. In this current study, the aim was to reveal the relationship between students' response times in science and mathematics items and student characteristics (proficiency levels, self-confidence) and item characteristics (item difficulty, item type, cognitive level) based on the 2019 eTIMSS data for students at the 8th grade level.

---

* A part of this study was presented at Assessment and Evaluation Practices in Our Education System: Findings, Solution Suggestions, and Projections for the Future Symposium. İstanbul Aydın University, İstanbul, Turkey.
** Assoc. Prof., Ankara University, Faculty of Educational Sciences, Measurement and Evaluation Department, Ankara- Turkey, yalcins@ankara.edu.tr, ORCID ID: 0000-0003-0177-6727

To cite this article:
Yalçın, S. (2022). Examining students' item response times in eTIMSS according to their proficiency levels, self-confidence, and item characteristics. *Journal of Measurement and Evaluation in Education and Psychology, 13*(1), 23-39. https://doi.org/10.21031/epod.999545

Received: 23.09.2021
Accepted: 17.12.2021

With the transition to computer-based applications, information about how long students spent answering the items has begun to be recorded. Response time can be defined as the time elapsed after the stimulus (item) is presented to an individual until that individual responds to the given stimulus (Lee & Chen, 2011). Therefore, in the psychometric, behavioral, and cognitive psychology literature, a variety of models have been proposed to model response times. The most widely used approach in the recent literature is to model response times and responses together (Su & Davison, 2019). In the field of psychometrics, it is stated that the use of item response time has become widespread in understanding the cognitive activities of individuals and determining the factors that affect a correct or incorrect response to an item (Schnipke & Scrams, 2002). Item response time has become more popular in recent years with the widespread use of computer-based applications (Lee & Chen, 2011; Lee & Haberman, 2016).

Item response time can be used for different purposes: item selection in computer adaptive testing (Lee & Haberman, 2016), relation to testing response motivation (İlgün-Dibek, 2020; Wise & DeMars, 2010; Wise & Kingsbury, 2016), revealing abnormal response behaviors (van der Linden & Guo, 2008), relation to test-taking behavior (quick-guessing behavior) (Kahraman et al., 2013; Lee & Jia, 2014), as an additional source of information in improving the precision of individuals' ability and item parameter estimation (Molenaar et al., 2015; Petscher et al., 2015) or as an indicator of student achievement (Goldhammer et al., 2014) in different contexts. Knowing the time an individual spends solving an item is necessary for the accuracy of item and ability parameter estimations as well as studies for test development in education to ensure test validity and effective examination. Not providing enough time to answer an item in the test creation process also negatively affects the validity of the obtained scores (Altuner, 2019).

The time students spend answering an item can be affected by many variables. In the literature, it is observed that item response time differs according to the student's ability level as well as the difficulty level of the items (Altuner, 2019; Goldhammer et al., 2014; İlgün-Dibek, 2020; Yavuz, 2019). In some studies (Davison et al., 2012; Goldhammer & Klein Entink, 2011; Klein Entink et al., 2009; Su & Davison, 2019), the relationship between the individual's performance/ability level/the behavior of answering the item correctly and the item response time is investigated. Also, it is investigated whether the students at the upper or lower ability level spend more or less time answering questions. In another study, it was found that students with low confidence levels spend more time answering items (Lasry et al., 2013).

In a study examining the relationship between item statistics and item response time on an English test according to different ability levels (low, medium, high), it is seen that response time varies according to the difficulty level of items. Additionally, high-level students answer easy items in a shorter time and spend more time on difficult items. It is concluded that the item discrimination index has no effect on the response time of students according to their ability levels (Altuner, 2019). While in another study conducted on the Program for International Student Assessment (PISA) items, it is found that the probability of answering an item correctly increases as the time for students to respond to the item increases, but there is no linear relationship between item difficulty and student proficiency (Yavuz, 2019).

In Halkitis et al. (1996), it is determined that the item response time is affected by the number of words, the difficulty of the item, and the level of discrimination. Goldhammer et al. (2014) examine the effect of time on reading and problem solving using the Program for the International Assessment of Adult Competencies (PIAAC) items. They find that the effect of time depends on item difficulty as well as the ability of test-takers. When studies in the literature were examined in general, it was seen that they were mostly based on large-scale study data. For example, PISA and PIAAC data were used, but there was no study addressing TIMSS items, the cognitive level of an item, and effect of item type on response time, and the psychological characteristics of students, such as self-confidence in the relevant field, which was discussed in only one study.

Purpose of the Study

This current study aimed to examine whether the time spent answering the science and mathematics items by Turkish students participating in TIMSS at the 8th grade level showed a significant difference according to their proficiency levels, self-confidence, and item characteristics (item type and cognitive level). Additionally, students' proficiency levels in the relevant field, their self-confidence, their answers to the items selected from the relevant field, and their classification according to the time they spent answering these items were examined. In this context, the sub-questions sought to be answered in this study are presented in the following:

1. Do students' response times to science and mathematics items differ significantly when students are grouped separately according to their science and mathematics proficiency levels and in terms of their self-confidence levels in science and mathematics?

2. Is there a significant difference in the averages of students' item response times to science and mathematics items according to their cognitive level and the type of item?

3. How are the students classified according to their science proficiency levels, their self-confidence in science, their answers to the four items (the three most difficult and one of the easiest), and the time they spent answering these items?

4. How are the students classified according to their mathematics proficiency levels, their self-confidence in mathematics, their answers to the four items (the three most difficult and one of the easiest), and the time they spent answering these items?

Method

Model of the Research

This current study was a correlational study, which was aimed at exploring the relationship between the time spent by Turkish students answering the items from the science and mathematics items of the eTIMSS, their proficiency levels in that area, students' self-confidence, and the item characteristics (item type and cognitive level) (Fraenkel et al., 2012).

Sample

In TIMSS, a two-stage stratified clustering model is used. In the first stage, the schools that will participate in the exam are selected from the sample. In the second stage, the classes belonging to the schools that are chosen for the sample are selected according to the grade that will participate in the exam. Thus, 577 students who participated in the TIMSS 2019 study from Turkey at the 8th-grade level and answered the common 24 (11 mathematics and 13 science) items in Booklets 1 and 2 constituted the participants of this current study. Of these students, 293 (50.8%) were girls, 280 (48.5%) were boys, and the gender information of four students (0.7%) was unavailable.

Data Source and Data Collection Methods

In TIMSS, 14 different booklets are used, organized according to the subject areas and cognitive areas appropriate for the grade level (Ruddock et al., 2008). Each booklet is applied to 250-300 students, and the booklets are linked by common items. To include more individuals in the sample of this study, individuals who answered the common items in the first and second booklet were included in the study. Since there is only one-time information regarding the items with common roots among the common items, these items were excluded from the analysis. The data were obtained from the TIMSS web page of the International Association for the Evaluation of Educational Achievement (IEA), which is open to all researchers (IEA, n.d.). Therefore, an ethics committee decision was not required for this study.

The learning domain, cognitive level, item type, and item difficulty values of 24 items belonging to the science and mathematics achievement tests used in this study, which were calculated according to the individuals who answered these items from all participating countries on the TIMSS 2019 study, are presented in Table 1 and Table 2 (Fishbein et al., 2021). Additionally, the average, standard deviation, skewness, and kurtosis coefficients of the response times of the items are provided in Tables 1 and 2.

Table 1
Information on Science Items and Item Response Times

	Information of items				Item Response Times			
Item code	Learning area	Cognitive level	Item type	Item difficulty	Mean	Standard deviation	Skewness coefficient	Kurtosis coefficient
SE72072	B	Knowing	MC	.556	50.98	42.97	3.81	25.43
SE72029	B	Knowing	MC	.539	30.51	23.17	2.52	8.77
SE72902	B	Reasoning	OE	.514	117.81	67.02	1.14	1.77
SE72077	B	Applying	MC	.571	64.26	33.44	1.97	8.31
SE72103	C	Knowing	OE	.504	49.9	34.59	2.51	11.49
SE72110	C	Applying	OE	.329	72.37	38.94	2.28	9.82
SE72148	C	Knowing	MC	.413	45.25	34.09	4.1	34.33
SE72200	P	Applying	MC	.449	38.01	26.72	4.27	29.31
SE72275	P	Knowing	MC	.729	45.91	29.83	2.72	12.9
SE72244	P	Applying	OE	.431	93.71	54.39	1.45	3.8
SE72301	ES	Knowing	MC	.467	37.85	25.57	2.79	14.49
SE72721	ES	Reasoning	MC	.643	43.43	32.12	5.92	70.15
SE72335	ES	Applying	MC	.542	62.75	57.26	6.21	59.43

Note: B: Biology, C: Chemistry, P: Physics, ES: Earth Sciences, MC: Multiple choice, OE: Open-ended.

The 13 science items in Table 1 are common items in both Booklet 1 and Booklet 2 of the TIMSS. When the items were examined in terms of learning domain, it could be seen that four items were in Biology, three items in Chemistry, three items in Physics, and three items in Earth Sciences. Six of the items were at the knowing cognitive level, five at the level of practice, and two at the level of reasoning. While nine of the items were multiple-choice, four were open-ended. The open-ended items were converted into 1-0 data. When the difficulty level of the items was examined, it was seen that the most difficult item was the item coded "SE72110", but it could also be stated that all the items were of medium difficulty (Miller et al., 2009).

When the information on the item response times in Table 1 was examined, the item with the highest item response time average was the item coded "SE72902" (M = 117.81, SD: 67.02). When the skewness coefficients of the item response times were examined, it could be seen that they varied between 1.14 and 6.21, and the kurtosis coefficients varied between 1.77 and 70.15. As a result, it was determined that the skewness and kurtosis coefficients exceeded the acceptable values. Following, when the mean, median, mode values, and histogram graph of the distribution were also examined, consequently, the data did not meet the normal distribution assumption.

Table 2

Information on Mathematic Items and Item Response Times

Information of items					Item Response Times			
Item code	Learning area	Cognitive level	Item type	Item difficulty	Mean	Standard deviation	Skewness coefficient	Kurtosis coefficient
ME72025	N	Applying	MC	.496	55.63	34.87	2.55	14.09
ME72017	N	Reasoning	OE	.20	101.56	76.75	1.9	5.16
ME72190	N	Knowing	OE	.57	72.55	47.66	2.3	9.08
ME72068	A	Knowing	MC	.649	59.28	39.48	1.75	4.99
ME72076	A	Knowing	MC	.435	40.54	31.87	3.96	30.24
ME72056	A	Applying	OE	.393	78.07	54.35	4.7	49.38
ME72098	A	Applying	OE	.309	99.6	68.98	2.09	7.48
ME72103	A	Applying	MC	.469	117.83	74.14	2.66	15.52
ME72121	G	Applying	MC	.595	53.33	40.52	3.25	18.1
ME72227	DP	Knowing	OE	.411	64.3	39.98	1.63	3.64
ME72209	DP	Reasoning	OE	.175	161.08	107.66	1.89	5.33

Note: N: Numbers, A: Algebra, G: Geometry, DP: Data and Probability, MC: Multiple choice, OE: Open-ended.

The 11 math items in Table 2 are common math items in both Booklet 1 and Booklet 2. When the items were examined in terms of learning domain, it could be seen that three items were in Numbers, five items in Algebra, one in Geometry, and two items in Data and Probability. Four of the items were at the knowledge cognitive level, five at the level of application, and two at the level of reasoning. While five of the items were multiple-choice, six were open-ended. The open-ended items were converted into 1-0 data. When the difficulty level of the items was examined, it could be stated that the two most difficult items were the items coded "ME72209 and ME72017", while the rest of the items were of medium difficulty (Miller et al., 2009).

When the information on the item response times in Table 2 was examined, the item with the highest item response time average was the item coded ME72121 (M = 117.83, SD: 74.14). When the skewness coefficients of the item response times were examined, it was seen that they varied between 1.63 and 4.7, and the kurtosis coefficients varied between 3.64 and 49.38. As a result, it was determined that the skewness and kurtosis coefficients exceeded the acceptable values. Following, when the mean, median, mode values, and histogram graph of the distribution were also examined, the data did not meet the normal distribution assumption.

The reliability coefficients were also calculated for the studied group, separately for the items in the field of science and mathematics. The Cronbach Alpha coefficient calculated for 13 items in the field of science was .73, and the Cronbach Alpha coefficient calculated for 11 items in the field of mathematics was .85. Reliability coefficients in both domains indicated sufficient reliability.

Data Analysis

Within the scope of the first sub-objective of this study, the Kruskal Wallis-H test was used to determine whether the response times of students to the science and mathematics items differed significantly when the students were grouped separately according to their science and mathematics proficiency levels and according to their self-confidence levels in science and mathematics (I am very confident, a little confident, not confident). Since the data did not meet the normal distribution assumption, non-parametric tests were used.

Five plausible values (PV1-PV5) for science and mathematics were used to determine the level of proficiency of the students. For each plausible value, the students were coded as “0” for each possible value below 400 points, “1” for 400-475, “2” for 476-550, “3” for 551-625, and “4” for 625 points. These cut-off scores are the TIMSS proficiency level scores. In cases where the proficiency codes for each plausible value were not the same, the students were divided into classes according to their proficiency levels, taking the most repeated proficiency level. Since science and mathematics item response times, which were dependent variables, did not satisfy the assumption of normal distribution, non-parametric tests were used. Within the scope of the second sub-objective of this study, whether the mean response times of students in science and mathematics items differed significantly according to the type of item (multiple-choice, open-ended), the Mann-Whitney U test, and whether there was a significant difference according to cognitive level (knowing, applying, reasoning) was tested with the Kruskal Wallis-H test. Also, the Bonferroni correction was used to control for Type I error during pairwise comparison of groups with significant differences.

Within the scope of the third and fourth sub-objectives of this study, students were able to determine how they were classified according to their science/mathematics proficiency levels, their self-confidence in science/mathematics, and their answers to four items (the three most difficult and one of the easiest), and the time they spent answering those items. Latent Class Analysis was used to determine the classes of the students according to the variables discussed. In other words, LCA was performed to see the grouping that emerged when all the variables considered in the study were evaluated together. The three most difficult and the easiest items in the related booklet were chosen because the time students spent answering the items made more difference between difficult and easy questions. In the analysis of the data, final student weighting was used while analyzing student responses. SPSS and Latent Gold programs were used in the analyses, and the significance level was .05.

Results

The results regarding whether the response times of the students to the items showed a significant difference when the students were grouped according to their science proficiency levels are provided in Table 3.

Table 3

Examination of Students' Response Times in Science Items According to their Proficiency Levels

Item codes	Category	N	Mean rank	Chi-square	df	Significant difference
SE72072	Sub-level	74	290.47	3.418	4	
	Low level	118	268.46			
	Intermediate	170	285.66			
	Top level	133	292.71			
	Advanced level	80	311.70			
SE72029	Sub-level	74	310.40	10.139	4	
	Low level	118	302.75			
	Intermediate	170	303.82			
	Top level	133	260.56			
	Advanced level	80	257.51			
SE72902	Sub-level	74	255.88	7.198	4	
	Low level	118	267.81			
	Intermediate	170	295.61			
	Top level	133	309.02			
	Advanced level	80	296.36			
SE72077	Sub-level	74	343.28	23.687*	4	0>2, 0>3, 1>3
	Low level	118	320.47			
	Intermediate	170	276.37			
	Top level	133	242.13			
	Advanced level	80	289.94			

Table 3 (continued)

Item codes	Category	N	Mean rank	Chi-square	df	Significant difference
SE72103	Sub-level	74	306.82	16.176*	4	0>4, 2>4, 3>4
	Low level	118	282.91			
	Intermediate	170	311.85			
	Top level	133	287.55			
	Advanced level	80	224.81			
SE72110	Sub-level	74	295.42	26.610*	4	2>4, 2>3
	Low level	118	280.83			
	Intermediate	170	332.80			
	Top level	133	269.94			
	Advanced level	80	223.06			
SE72148	Sub-level	74	282.93	17.723*	4	1>4, 2>4, 3>4
	Low level	118	299.76			
	Intermediate	170	310.96			
	Top level	133	290.24			
	Advanced level	80	219.16			
SE72200	Sub-level	74	261.51	10.626*	4	2>4, 3>4
	Low level	118	285.00			
	Intermediate	170	305.43			
	Top level	133	306.28			
	Advanced level	80	245.57			
SE72275	Sub-level	74	318.14	11.900*	4	1>4, 2>4
	Low level	118	303.16			
	Intermediate	170	298.61			
	Top level	133	270.26			
	Advanced level	80	241.50			
SE72244	Sub-level	74	305.74	18.396*	4	1>4, 2>4
	Low level	118	312.71			
	Intermediate	170	305.51			
	Top level	133	269.42			
	Advanced level	80	225.45			
SE72301	Sub-level	74	294.92	11.987*	4	1>4
	Low level	118	319.59			
	Intermediate	170	286.56			
	Top level	133	284.18			
	Advanced level	80	237.31			
SE72721	Sub-level	74	304.93	7.632	4	
	Low level	118	309.76			
	Intermediate	170	293.07			
	Top level	133	263.73			
	Advanced level	80	262.94			
SE72335	Sub-level	74	247.16	5.755	4	
	Low level	118	292.04			
	Intermediate	170	299.94			
	Top level	133	291.68			
	Advanced level	80	277.37			

Note: $^*p<.05$, Sub-level: 0, Low level:1, Intermediate: 2, Top level: 3, Advanced level

It was determined that the response times of students to the items did show a significant difference in eight items when the students were grouped according to their science proficiency levels ($p<.05$). In the dual proficiency groups with a significant difference, it was observed that the mean rank of the response times of students at the lower proficiency level was higher than the response times of students at the middle or upper proficiency level. In other words, as the proficiency level of the students' increased, their response times decreased.

For example, it was determined that students with moderate proficiency with an open-ended medium difficulty level item at the Applying level in the field of Chemistry coded "SE72110" spent significantly more time answering the item than those at the advanced and high levels. However, there was no significant difference between the time spent by students at the upper and advanced proficiency levels. In Table 4, it is shown whether the response times of students to the items showed a significant difference when the students were grouped according to their mathematical proficiency levels.

Table 4
Examining the Response Times of the Students in Mathematics Items According to their Proficiency Levels

Item codes	Category	N	Mean rank	Chi-square	df	Significant difference
ME72025	Sub-level	111	271.65	8.652	4	
	Low level	144	285.45			
	Intermediate	143	305.42			
	Top level	103	310.16			
	Advanced level	73	248.56			
ME72017	Sub-level	111	236.10	129.613*	4	3>1, 3>0, 3>2, 4>0, 4>2, 4>1, 4>3
	Low level	144	224.40			
	Intermediate	143	250.96			
	Top level	103	370.78			
	Advanced level	73	444.22			
ME72190	Sub-level	111	269.67	15.385*	4	3>4, 1>0, 1>4
	Low level	144	319.33			
	Intermediate	143	275.84			
	Top level	103	311.78			
	Advanced level	73	240.42			
ME72068	Sub-level	111	250.55	60.531*	4	3>4, 2>4, 1>4, 1>0, 2>0, 2>3, 1>3
	Low level	144	353.98			
	Intermediate	142	316.67			
	Top level	103	259.17			
	Advanced level	73	191.86			
ME72076	Sub-level	110	252.01	43.656*	4	3>4, 2>4, 1>4, 2>0, 2>3, 1>3, 1>0
	Low level	144	341.22			
	Intermediate	141	312.91			
	Top level	103	265.56			
	Advanced level	73	205.16			
ME72056	Sub-level	110	240.18	13.852*	4	2>0, 3>0
	Low level	144	291.00			
	Intermediate	142	311.52			
	Top level	103	305.19			
	Advanced level	73	272.38			
ME72098	Sub-level	110	200.05	68.397*	4	4>2, 3>0, 3>2, 1>0, 4>0, 3>1, 4>1, 2>0
	Low level	144	265.58			
	Intermediate	143	284.96			
	Top level	103	370.52			
	Advanced level	73	346.42			
ME72103	Sub-level	110	233.29	16.619*	4	1>0, 2>0, 3>0, 4>0
	Low level	144	294.69			
	Intermediate	142	288.43			
	Top level	103	320.96			
	Advanced level	73	298.15			

Table 4 (continued)

Item codes	Category	N	Mean rank	Chi-square	df	Significant difference
ME72121	Sub-level	110	297.63	25.505*	4	2>4, 2>3, 1>3, 1>4
	Low level	144	319.83			
	Intermediate	143	307.39			
	Top level	103	248.07			
	Advanced level	73	221.22			
ME72227	Sub-level	109	218.72	35.684*	4	2>0, 3>0, 3>1, 4>0, 3>4
	Low level	141	269.65			
	Intermediate	143	317.48			
	Top level	103	338.41			
	Advanced level	73	274.64			
ME72209	Sub-level	110	274.83	4.812	4	
	Low level	138	277.20			
	Intermediate	143	274.42			
	Top level	103	290.50			
	Advanced level	73	320.27			

Note: *$p<.05$, Sub-level: 0, Low level:1, Intermediate: 2, Top level: 3, Advanced level

It was determined that the students' response times to the items did show a significant difference in nine items when the students were grouped according to their mathematical proficiency levels ($p< .05$). In the dual proficiency groups with a significant difference, it was observed that the mean rank of the response times of students at the lower proficiency level was higher than the response times of students at the middle or upper proficiency level. In other words, as the proficiency level of students increased, their response times decreased. Only in item "ME72017", was the situation reversed. Thus, it was determined that students with high proficiency spent significantly more time in answering an item than those with low and medium proficiency, and those with advanced proficiency spent significantly more time than those with low, medium, and high proficiency. In other words, as the proficiency level of the students' increased, their response times also increased. This was thought to be due to the fact that the item was open-ended at the reasoning level in the "Fractions and Decimal Numbers" subject area, and the item difficulty level was "0.18", that is, it was a very difficult item. Additionally, it was determined that students with high level proficiency spent significantly more time than advanced students in answering the item coded "ME72190", in the mathematics area, numbers, the knowledge level, and in the open-ended question, for the medium difficulty question. In the "ME72098" coded algebra subject area, an open-ended difficulty item in the application level, it was determined that students with high level proficiency spent significantly more time answering the item than those at the intermediate level. Students with high level proficiency in the "ME72227" coded mathematics area, data and probability subject area, open-ended in knowledge level, and difficulty item spent significantly more time answering the item than the lower level students. Also, it was determined that those at the advanced proficiency level spent significantly less time than those at the higher proficiency level. The results were examined as to whether the response times of students to the items differed significantly according to their self-confidence levels in science, which are presented in Table 5.

Table 5

Examination of Students' Response Times on Science Items According to Their Confidence Levels in Science

Item codes	Category	N	Mean rank	Chi-square	df	Significant difference
SE72072	I am very confident	187	302.15	3.655	2	
	I am a little confident	221	285.33			
	I am not confident	163	268.38			
SE72029	I am very confident	187	291.36	1.057	2	
	I am a little confident	221	289.71			
	I am not confident	163	274.83			
SE72902	I am very confident	187	311.60	9.194*	2	1>3
	I am a little confident	221	284.95			
	I am not confident	163	258.04			
SE72077	I am very confident	187	286.44	.262	2	
	I am a little confident	221	289.47			
	I am not confident	163	280.79			
SE72103	I am very confident	187	283.54	.212	2	
	I am a little confident	220	289.47			
	I am not confident	163	282.39			
SE72110	I am very confident	187	281.36	.199	2	
	I am a little confident	220	288.62			
	I am not confident	163	286.04			
SE72148	I am very confident	187	264.53	6.549*	2	2>1
	I am a little confident	220	306.08			
	I am not confident	163	281.79			
SE72200	I am very confident	187	294.57	.863	2	
	I am a little confident	220	282.06			
	I am not confident	163	279.73			
SE72275	I am very confident	187	287.15	.628	2	
	I am a little confident	220	290.30			
	I am not confident	163	277.12			
SE72244	I am very confident	187	275.25	1.685	2	
	I am a little confident	220	296.15			
	I am not confident	163	282.89			
SE72301	I am very confident	187	258.65	7.471*	2	3>1
	I am a little confident	219	293.81			
	I am not confident	163	303.39			
SE72721	I am very confident	187	274.40	1.184	2	
	I am a little confident	219	289.02			
	I am not confident	163	291.76			
SE72335	I am very confident	187	279.37	2.344	2	
	I am a little confident	218	297.48			
	I am not confident	163	273.02			

*$p<.05$

When Table 5 is examined, it can be seen that the response times of the students to the items differed in three items according to their self-confidence levels in science ($p<.05$). In the “SE72902” coded science item, the students who were very confident spent significantly more time answering the question than those who were not confident. This could have been due to the item being an open-ended item at the level of reasoning. On the other hand, in the item coded "SE72148", the time spent by the students who were somewhat confident in answering the question was significantly longer than those who did not. This item was a medium difficulty multiple-choice item. In the science item coded "SE72301", the time spent by the students who were not self-confident to answer the question was significantly longer than the students who were very confident. This item was a multiple-choice item at the knowledge level that could be considered easy. The results of the analysis of the students' response times to mathematics items according to their self-confidence levels in mathematics are presented in Table 6.

Table 6
Examination of Students' Response Times on Mathematics Items According to Their Confidence Levels in Mathematics

Item codes	Category	N	Mean rank	Chi-square	df	Significant difference
ME72025	I am very confident	77	281.77	.459	2	
	I am a little confident	195	290.40			
	I am not confident	295	280.35			
ME72017	I am very confident	77	394.23	54.119*	2	1>2, 1>3, 2>3
	I am a little confident	195	300.46			
	I am not confident	295	244.34			
ME72190	I am very confident	77	287.44	.047	2	
	I am a little confident	195	284.24			
	I am not confident	295	282.94			
ME72068	I am very confident	76	227.72	12.738*	2	3>1
	I am a little confident	195	277.72			
	I am not confident	295	301.69			
ME72076	I am very confident	76	245.37	6.017	2	
	I am a little confident	195	277.37			
	I am not confident	293	295.55			
ME72056	I am very confident	76	271.23	1.013	2	
	I am a little confident	195	291.59			
	I am not confident	294	280.35			
ME72098	I am very confident	77	345.48	18.065*	2	1>3
	I am a little confident	195	294.57			
	I am not confident	294	259.92			
ME72103	I am very confident	76	271.06	1.424	2	
	I am a little confident	195	293.71			
	I am not confident	294	278.98			
ME72121	I am very confident	77	282.53	.552	2	
	I am a little confident	195	276.93			
	I am not confident	294	288.11			
ME72227	I am very confident	77	307.74	5.454	2	
	I am a little confident	194	293.29			
	I am not confident	291	266.69			
ME72209	I am very confident	77	283.88	2.153	2	
	I am a little confident	194	293.67			
	I am not confident	290	271.76			

*p<.05

As seen in Table 6, students' response times to mathematics items differed significantly in three items according to their self-confidence levels in mathematics (p<.05). In the math item coded "ME72017", the time spent by students who were very confident in answering the question was significantly longer than the students who were a little confident or not self-confident. Additionally, the time spent by the students who were somewhat confident in answering the question was significantly higher than the students who were not confident. This could have been due to the fact that the item was very difficult and open-ended at the level of reasoning. In the item coded “ME72068”, the time spent by the students who were not self-confident to answer the question is significantly higher than the students who were very confident. The multiple choice was an easy, knowledge-level item. In the item coded "ME72098", the time spent by the students who were very confident in answering the question was significantly longer than those who did not. This may be due to the fact that the item was quite difficult and open-ended at the application level. The results regarding whether the mean scores of students' item response times differed significantly according to the type of item are presented in Table 7.

Table 7

Examining Students' Item Response Time Averages According to Item Type

Category	N	Mean rank	Sum of ranks	Mann-Whitney U	Z
Multiple choice	14	8.57	120.00	15.000	-3.220*
Open-ended	10	18.00	180.00		

*p<.05

As seen in Table 7, the type of item had a significant difference in the students' item response time averages (Z=-3.220, p<.05). Students spent more time on open-ended questions than multiple-choice questions. The results regarding whether the mean scores of students' item response times showed a significant difference according to the cognitive level of the item are presented in Table 8.

Table 8

Examining Students' Item Response Time Averages According to Cognitive Level

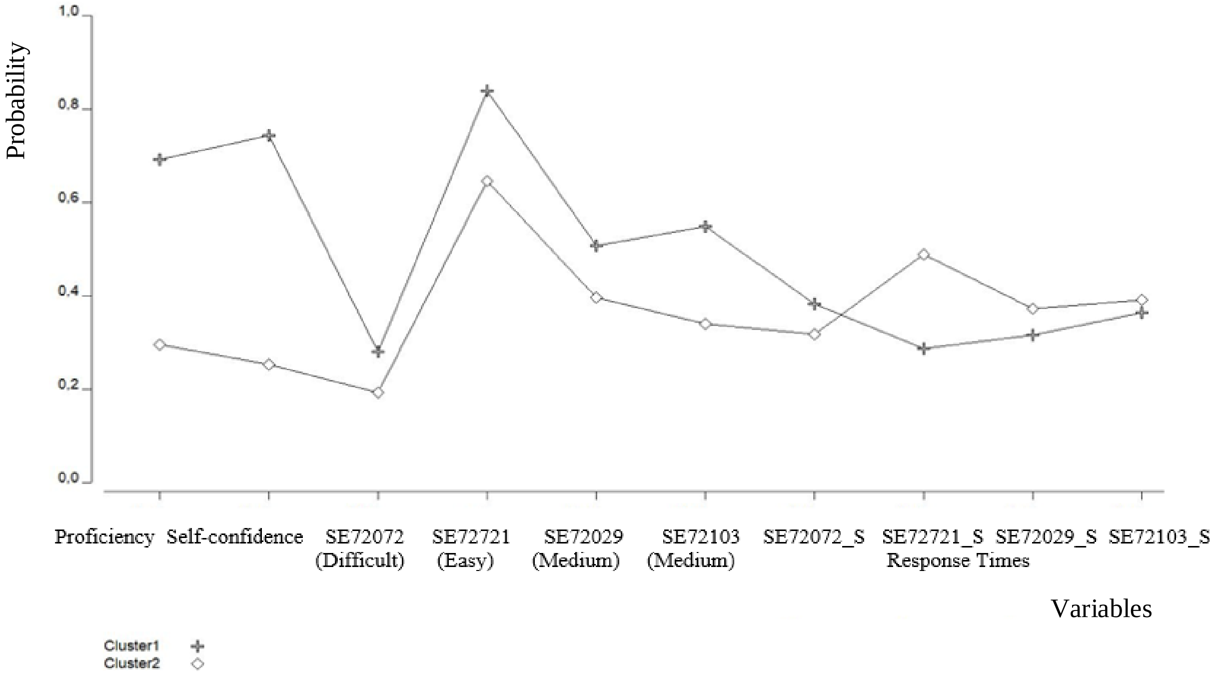
Category	N	Mean rank	Chi-square	df	Significant difference
Knowing	10	8.10	7.260*	2	2>1
Applying	10	14.70			
Reasoning	4	18.00			

*p<.05

As seen in Table 8, the cognitive level of the item had a significant difference in students' item response times (χ^2(2) = 7.260, p< .05). When Bonferroni correction was applied in the comparison between paired groups, there was no significant difference between any two subgroups. Whereas when the correction was not applied, the time spent on the questions at the application level was significantly higher than at the knowledge level. However, the correction results applied to reduce the Type 1 error were considered. Two latent clusters were formed according to the students' science proficiency levels, their self-confidence in science, their answers to the four science items, and the time they spent answering the items. The results are provided in Figure 1.

Figure 1

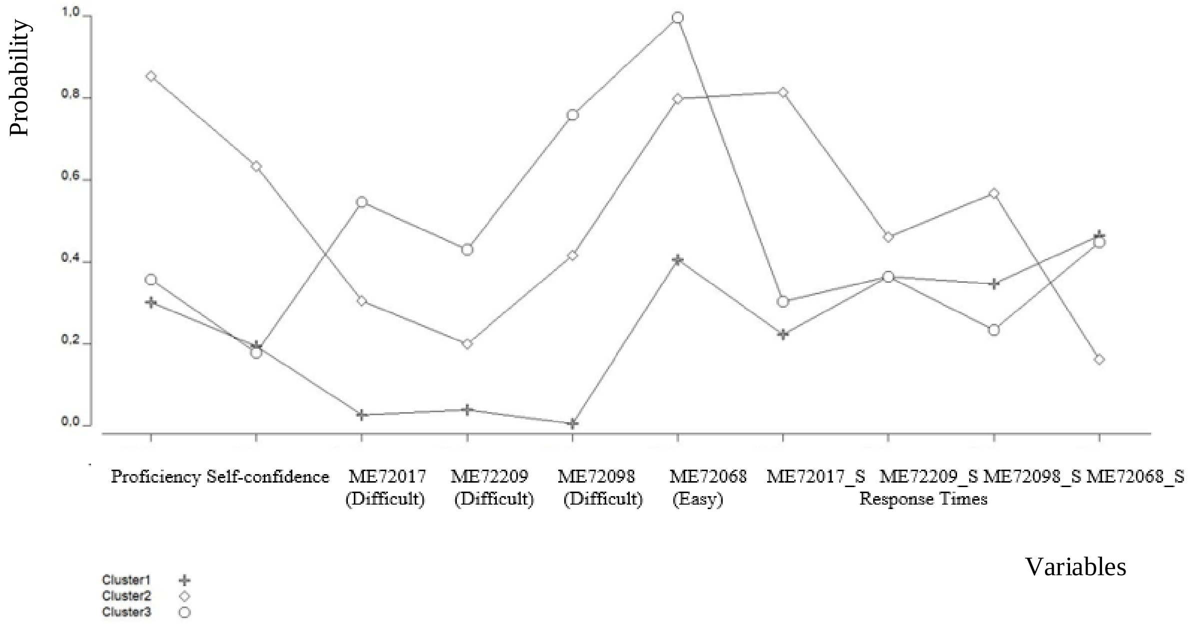
Latent Clusters Formed According to the Variables Considered in the Field of Science



As seen in Figure 1, the students in Cluster-1 are the students who had high proficiency in science, high self-confidence in science, and were more likely to answer the four items correctly than those in the first cluster. However, it was seen that they differed in terms of the time spent answering those items. The student group in the second cluster, on the other hand, consisted of students with low science proficiency, low self-confidence in science, and were less likely to answer the four items correctly than those in the first cluster. The time spent by students in this group to answer the items was lower except for one item. Thus, it was observed that students with high self-confidence in science spent less time on easy and medium difficulty questions, while they spent more time on difficult questions. In general, a pattern such as students' high proficiency, high confidence level, high item response rates, and low response time emerged in science items. Three latent clusters were formed according to the students' mathematical proficiency levels, their self-confidence in mathematics, their answers to the four items, and the time they spent answering the items. The results are provided in Figure 2.

Figure 2

Latent Clusters Formed According to the Variables in the Field of Mathematics



As seen in Figure 2, students in Cluster 1 and Cluster 3 consisted of students with low mathematics proficiency and low self-confidence in mathematics. However, students differed in terms of providing correct answers to the items. While students in Cluster 1 answered all questions incorrectly except for the easy item, it was seen that students in Cluster 3 answered those questions correctly at a higher rate than students in Cluster 2 with high self-confidence and proficiency. When the time spent by the students to answer the items was analyzed, it was seen that students in Cluster 1 and Cluster 3 spent similar times in answering the items, but students in Cluster 3 answered the questions correctly, similar to the time that students in Cluster 1 spent on the items they answered incorrectly. When the responses to mathematics items were evaluated in general, low proficiency was consistent with low self-confidence and low item correct answer patterns, whereas high proficiency was not valid for high confidence levels. Students who were somewhat self-confident were better at difficult mathematics items and spent less time answering those items.

Discussion and Conclusion

In this study, it was investigated whether the time spent answering science and mathematics items by Turkish students who participated in eTIMSS at the 8th grade showed a significant difference according to their proficiency levels, students' self-confidence, and the item characteristics (item type and cognitive level). When the students' response times to the items were grouped according to their science and mathematics proficiency levels, it was seen that there was a significant difference in most items. Thus, it was determined that as the proficiency level of students increased, their response times generally decreased. However, this pattern was not valid for all items. When the literature was reviewed, it was seen that it was in parallel with the general pattern detected within this study. In a study conducted by Su and Davison (2019), it is observed that individuals with high abilities give correct answers faster in the assessment of reading comprehension, and correct response times varied according to individuals with different abilities. In other studies, it is consistent with the finding that highly talented individuals respond more quickly (Goldhammer & Klein Entink, 2011; Petscher et al., 2015), which was in line with the findings from this current study. Additionally, in Goldhammer et al. (2014), a study of PIAAC data, it is seen that the time spent in reading skill tasks decreases as the skill level increases, while the time spent in problem solving tasks increases as the skill level increases. These heterogeneous differences suggest that time on the task is not a uniform interpretation but a function of task difficulty and individual skill. In a study of PISA data, it is determined that the time spent answering an item is not in a linear relationship with item difficulty and student proficiency (Yavuz, 2019). When the findings of the studies in the literature and the findings obtained from this current study were evaluated together, it was seen that although the cognitive competence of students had a positive effect on the correct response to an item in a short time, this finding was not valid under all conditions.

It was observed that the response times of students to the items differed for some items according to their self-confidence levels in science and mathematics. It was also observed that students with high self-confidence in science spent less time on easy and medium difficulty questions, while they spent more time on difficult questions. Additionally, it was found that students who were somewhat self-confident in mathematics were more successful in difficult mathematics items and spent less time answering these items than highly confident students. Furthermore, it was shown that the subject area and cognitive level of the item had an effect on self-confidence. In one study in the literature (Lasry et al., 2013), it is determined that students with low confidence levels spend more time answering items. This finding is consistent with the findings obtained for easy and medium-difficulty questions in the field of science.

It was determined that the type of item and cognitive level had a significant difference in the item response times of the students. Students spent more time in open-ended questions than in multiple-choice questions. In another study (Birgili, 2014), it is stated that students made more effort to answer open-ended questions than multiple-choice questions. In a study conducted with TIMSS 2015 data (İlhan et al., 2020), it is found that students had more difficulty in open-ended items than in multiple-choice items. When the research findings in terms of cognitive level are examined, the difference in cognitive level is not significant between the subgroups. In other words, there is no significant difference between the time spent answering the questions in the pairwise comparison of cognitive levels. It was seen that the studies in the literature are focused on the difficulty/ease of items rather than the cognitive level. Also, in a study conducted with TIMSS 2015 data (İlhan et al., 2020), they state that item type and cognitive level have a statistically significant effect on the difficulty index. In this context, considering the item type and cognitive level separately in the analyzes may have affected the findings. The joint effect of these two variables may have a different effect than the separate effects of the two variables. Therefore, it is recommended to consider this situation in other studies.

The proficiency levels of students in the relevant field, their self-confidence, their answers to items selected from the relevant field, and their classification according to the time they spent answering those items were examined. In science items, a pattern such as high proficiency of students, high confidence level, high item response rates, and low response time emerged. This situation may have been related to the items in the booklet generally being of medium difficulty or it could be explained by the students having known what they could do in the field of science and that their confidence levels were high. In mathematics items, this situation was consistent with low proficiency, low self-confidence, and low item correct answer patterns, but not for high levels. Thus, it was determined that students who were somewhat self-confident in mathematics were more successful in difficult mathematics items and spent less time answering those items. As was seen from the visuals of the latent clusters formed, it was determined that the time spent on the task was not linearly related to the difficulty and individual competencies of the task. This finding was also consistent with the findings of other researchers in the literature (Goldhammer et al., 2014; Yavuz, 2019). Considering the studies regarding item difficulty, Altuner (2019) finds that students with high ability levels answer easy items in a shorter time and spend more time on difficult items, which was consistent with this current study. Goldhammer et al. (2014), in their study of PIAAC data, find that students spend more time on difficult tasks and less time on tasks that require routine operations. Unlike the findings obtained in this study, another study conducted on PIAAC data concluded that individuals spare very little time on very difficult items (İlgün-Dibek, 2020). This is explained by the low risk of the study and the low motivation of individuals to answer difficult questions in this case. In another study conducted on PISA items, it is found that the probability of answering the item correctly increases as the time for students to respond to the item increases, but there is no linear relationship between item difficulty and student proficiency (Yavuz, 2019).

The fact that the data in this current study was obtained from a low-risk study for students may have also affected their response times. However, unlike the low-risk PIAAC study in Turkey, the motivation of students who participate in international large-scale studies such as TIMSS and PISA is considered to be higher than the PIAAC participants. This may due to these assessments being applied to students during their compulsory education and that the studies were carried out by the MoNE to increase the students' motivation to participate in these types of assessments. In another study, it is stated that students spend an average level of effort while answering questions in international large-scale measurement studies such as TIMSS and PISA (Wise & DeMars, 2010).

Considering the effect of self-confidence levels on the time spent for students to solve an item, it is recommended to carry out studies regarding students' self-confidence. By asking the students to think aloud while solving questions, it can be determined how confident they are in answering questions. With quick feedback, students' self-confidence can be increased. The fact that TIMSS is not one of the high-risk studies for students could have affected students' response times and motivation to respond. Also, students in this study were not conscious of the effect of time. The effect of time can be tested with experimental studies in which it is said that time will also be taken into account. Additionally, similar work can be replicated in high-risk studies for students. It should be noted that the findings obtained in this current study were limited to students who answered common items in two booklets from the TIMSS. A similar study can be made for the items in all the booklets of the TIMSS study, and as a result, patterns can be detected. Thus, it was determined that many different variables have a relation to students' correct responses from item to item as well as response times in both areas. In this context, it is recommended that studies regarding the time students spend answering items should be designed as studies that model student characteristics and item characteristics together. In addition, in the light of the latest technological developments, studies can be conducted in which the time spent responding to the item with eye-tracking devices is modeled together. Finally, non-parametric tests were used as the dependent variable did not meet the assumption of normal distribution. If this assumption is met, results can be compared according to parametric methods.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Secondary data were used in this study. Therefore, ethical approval is not required.

References

Altuner, F. (2019). *Examining the relationship between item statistics and item response time* [Master's Thesis, Mersin University]. Retrieved from http://tez2.yok.gov.tr/

Birgili, B. (2014). *Open ended questions as an alternative to multiple choice: Dilemma in Turkish examination system* [Master's Thesis, Middle East Technical University]. Retrieved from http://tez2.yok.gov.tr/

Davison, M. L., Semmes, R., Huang, L., & Close, C. N. (2012). On the reliability and validity of a numerical reasoning speed dimension derived from response times collected in computerized testing. *Educational and Psychological Measurement*, *72*(2), 245-263. https://doi.org/10.1177/0013164411408412

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 user guide for the international database*. TIMSS & PIRLS International Study Center, Boston College.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th ed.). McGraw-Hill.

Goldhammer, F., & Klein Entink, R. H. (2011). Speed of reasoning and its relation to reasoning ability. *Intelligence*, *39*(2-3), 108-119. https://doi.org/10.1016/j.intell.2011.02.001

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., & Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, *106*(3), 608-626. https://doi.org/10.1037/a0034716

Halkitis, P. N., Jones, J. P., & Pradhan, J. (1996, April 8-12). *Estimating testing time: The effects of ıtem characteristics on response latency* [Paper presentation]. Annual meeting of the American Educational Research Association, New York.

International Association for the Evaluation of Educational Achievement (IEA). (n.d.). *Trends in international mathematics and science study: Data & tools*. https://www.iea.nl/index.php/data-tools/repository/timss

İlgün-Dibek, M. (2020). Silent predictors of test disengagement in PIAAC 2012. *Journal of Measurement and Evaluation in Education and Psychology*, *11*(4), 430-450. https://doi.org/10.21031/epod.796626

İlhan, M., Boztunç Öztürk, N., & Şahin, M. G. (2020). The effect of the item's type and cognitive level on its difficulty index: The sample of TIMSS 2015. *Participatory Educational Research*, *7*(2), 47-59. https://doi.org/10.17275/per.20.19.7.2

Kahraman, N., Cuddy, M. C., & Clauser, B. E. (2013). Modeling pacing behavior and test speededness using latent growth curve models. *Applied Psychological Measurement*, *37*(5), 343–360. https://doi.org/10.1177/0146621613477236

Klein Entink, R. H., Fox, J.-P., & Van Der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika*, *74*(1), 21–48. https://doi.org/10.1007/s11336-008-9075-y

Lasry, N., Watkins, J., Mazur, E., & Ibrahim, A. (2013). Response times to conceptual questions. *American Journal of Physics*, *81*(9), 703-706. https://doi.org/10.1119/1.4812583

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359–379.

Lee, Y.-H., & Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data, *International Journal of Testing*, *16*(3), 240-267. https://doi.org/10.1080/15305058.2015.1085385

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-scale Assessments in Education*, *2*(1), 1–24. https://doi.org/10.1186/s40536-014-0008-1

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th ed.). Prentice Hall.

Ministry of National Education. (2020). *TIMSS 2019 Türkiye ön raporu*. Eğitim Analiz Değerlendirme Raporları Serisi, No: 15.

Molenaar, D., Tuerlinckx, F., & van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, *50*(1), 56–74. https://doi.org/10.1080/00273171.2014.962684

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. Boston College, TIMSS & PIRLS International Study Center.

Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2008). Developing the TIMSS 2007 mathematics and science assessments and scoring guides. In J. F. Olson, M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 13-44). International Study Center, Boston College.

Petscher, Y., Mitchell, A. M., & Foorman, B. R. (2015). Improving the reliability of student scores from speeded assessments: An illustration of conditional item response theory using a computer-administered measure of vocabulary. *Reading and Writing*, *28,* 31–56. https://doi.org/10.1007/s11145-014-9518-z

Schnipke, D. L., & Scrams, D. J. (2002). Exploring issues of examinee behavior: Insights gained from response-time analyses. In N. C. Mills., M. T. Potenza, J. J. Fremer & C. W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 237-266). Psychology Press.

Su, S., & Davison, M. L. (2019). Improving the predictive validity of reading comprehension using response times of correct item responses. *Applied Measurement in Education*, *32*(2), 166-182. https://doi.org/10.1080/08957347.2019.1577247

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384. https://doi.org/10.1007/s11336-007-9045-8

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*(1), 27-41. https://doi.org/10.1080/10627191003673216

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student test-taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement*, *53*(1), 86-105. https://doi.org/10.1111/jedm.12102

Yavuz, H. Ç. (2019). The effects of log data on students' performance. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(4), 378-390. https://doi.org/10.21031/epod.564232