ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(3); 183-192

The Comparison of PISA 2015-2018 Mathematics Trend Items Based on Item Response Times

Muhsin POLAT* Hülya KELECİOĞLU**

Abstract

This study aims to explore the intricate relationship between students' response times, item characteristics, and the effort invested during the Programme for International Student Assessment (PISA) 2015 and 2018 cycles. Through the analysis of data obtained from 69 mathematics trend items, administered in a computer-based format across both PISA 2015 and 2018 cycles with a focus on the Türkiye sample, this research investigates the dynamics of students' response times and their implications on effort and item characteristics. The findings reveal a significant increase in students' mean response times in the 2018 cycle compared to 2015, indicating potentially heightened effort and solution behavior. Notably, item formats exerted a substantial influence on response times, with open-ended items consistently eliciting longer response times compared to multiple-choice items. Additionally, a correlation between response times and item difficulty emerged, suggesting that more challenging items tend to consume more time, possibly due to the complexity of involved cognitive processes. Item-based effort, assessed through Response Time Fidelity (RTF) indices, highlighted that the majority of students exhibited solution behavior across both cycles to the items. Moreover, a decrease in the proportion of students displaying rapid-guessing behavior was observed in the 2018 cycle, potentially reflecting increased engagement with the assessment. While providing insights into the interplay of response times, item characteristics, and effort, this study emphasizes the need for further exploration into the multifaceted nature of effort in educational assessments. Overall, this research contributes valuable perspectives on nuances surrounding test performance and effort evaluation within PISA mathematics assessments.

Keywords: item response time, PISA, response-time effort, rapid-guessing

Introduction

In the Programme for International Student Assessment (PISA) and other tracking assessment procedures, where students' acquired knowledge and skills are evaluated, the expected behavior of students is to consciously apply their acquired knowledge and skills to respond to items. The scores obtained from these tests are used to assess individuals' knowledge and abilities in terms of what they know and what they can do. During these assessments, it is assumed that individuals engage in effortful attempts to answer the test items. In fact, this assumption is made across all measurement processes (Wise & Kong, 2005). An individual engages in an interaction with test items, exhibiting effort to respond to each item to the best of their ability. Test-taking effort is commonly characterized by a student's active involvement and dedication to resources aimed at achieving the most favorable outcome on the examination (Debeer et al., 2017). In high-stakes tests, individuals are expected to exhibit this behavior because the test scores hold significance for the individual. In placement exams for institutions, university entrance exams, course passing exams, and similar assessments, students consciously exhibit effort to respond to test items. However, in low-stakes national and international monitoring tests, students can respond to test items without exhibiting much effort and complete the test within a very short period of time. In this case, it can lead to variance that is unrelated to the structure that the test aims to measure (construct-irrelevant variance), and test scores may underestimate the examinee's true ability. Failing to account for the impact of effort exerted during testing could compromise the validity

* National Educational Expert, Republic of Türkiye Ministry of National Education, Ankara-Türkiye, muhsinpolat58@gmail.com, ORCID ID: 0009-0003-2897-3189
** Prof. Dr. Hacettepe University, Faculty of Education, Ankara- Türkiye, hulyakelecioglu@gmail.com, ORCID ID: 0000-0002-0741-9934

To cite this article:
Polat, M. & Kelecioğlu, H. (2024). The comparison of PISA 2015-2018 mathematics trend items based on item response times. *Journal of Measurement and Evaluation in Education and Psychology, 15*(3), 183-192. https://doi.org/10.21031/epod.1398317

Received: 1.12.2023
Accepted: 24.09.2024

of test outcomes (Michaelides & Militsa, 2022). In cases of a lack of effort from individuals, two situations will emerge regarding test scores. The first situation is that when individuals do not exhibit sufficient effort, it will lead to a negative bias in the test scores. The second situation is that when individuals exhibit different levels of effort, it will result in variability in effort bias among individuals (Wise et al., 2006). Due to the construct-irrelevant variance caused by effort bias, the estimated levels of individual ability derived from these test scores will likely be lower than the actual ability levels. Individuals with high motivation and who exhibit effort in solving the test items will generally have higher test performance compared to those who show low effort (Eklöf et al., 2014; Rios & Guo, 2020). Therefore, in such a situation where variance unrelated to the construct occurs, the validity of evaluations made based on test scores and the decisions made will be low. Validity is necessary for interpreting test scores and using these scores for any purpose (AERA et al., 2014). The impact of factors not aligned with the test's objectives on test scores diminishes their validity, resulting in flawed deductions drawn from these scores.

The low effort that individuals put into answering test items directly impacts the validity of the test. Therefore, measuring individuals' efforts and revealing their impact is highly important. In the literature, efforts are measured using various methods. The first one is administering a self-report scale after the test for individuals to indicate how much effort they exerted. In this method, Likert-type scales are commonly used to reveal individuals' efforts. However, the objectivity of the information gathered about effort through self-reporting may be low (Wise & DeMars, 2006). Another method developed to measure individuals' efforts is person-fit statistics. Person-fit statistics aim to identify individuals' abnormal responses. For this purpose, each individual's response pattern is compared with measurement models (Meijer, 1996). Abnormal responses can include copied, careless, or random answers. Therefore, since abnormal responses don't solely indicate a lack of effort, using this statistic might not yield accurate results. Another method developed to identify responses without exhibiting effort, also used in this study, involves evaluating item response times. In this method, effort is defined by individuals' response times to items (Schnipke & Scrams, 1997; Wise & Kong, 2005). According to this method, individuals develop two types of behavior when encountering an item. First, individuals exhibit effort to answer the item correctly and attempt to solve it; this behavior is termed "solution behavior". Second, individuals do not contemplate the item and answer it without attempting to solve it; this behavior is termed "rapid-guessing behavior". Response times expended by individuals to solve the item are used to distinguish between these two behaviors (Wise & Kong, 2005).

Computerized tests enable the gathering of response times and diverse process-related data. Such tests make it possible to measure response time at the item level. The response time denotes the duration test-takers require to answer a specific item within the test context (Lee & Jia, 2014). Response time has been seen as a valid indicator of test-taking effort (Wise & Kong, 2005). As response time offers insights into examinee test-taking behavior on a per-item basis, it empowers researchers to monitor potential fluctuations in effort throughout the testing session (Wise & Kingsbury, 2016). For instance, item position within a test has been extensively examined as a significant factor influencing examinee effort; specifically, effort tends to decrease towards the end of a testing session (Debeer et al., 2014; Pools & Monseur, 2021). Item characteristics with less reading material and more answer options were associated with less rapid-guessing behavior (Setzer et al., 2013). DeMars (2000) indicated evidence for higher item non-response and lower effort in low-stakes constructed responses compared to multiple-choice items (DeMars, 2000). Therefore, examining response times across item types and characteristics between PISA cycles will enable the assessment of test-takers' efforts.

Wise and Kong (2005) developed the Response Time Effort (RTE) index to determine students' behavioral types toward items and their effort directed toward the test. This index utilizes the response time when students encounter an item to ascertain whether they exhibit solution behavior or engage in rapid-guessing. By assessing students' responses to all items on the test collectively, the Response Time Effort (RTE) index is constructed to represent the effort exhibited by students. The RTE indices range between 0 and 1, representing the proportion of solution behavior displayed during the test. As the RTE value approaches 1, it indicates that the student exhibited strong effort to solve the test, while a value closer to 0 suggests minimal effort was exerted. Wise (2006) similarly utilized item response times to

develop an index that indicates how much effort was exhibited on each item in the test. This index, named Response Time Fidelity (RTF), demonstrates the extent to which items are solved with solution behavior by students. The RTF index for items ranges between 0 and 1. As the index approaches 1, it indicates that a large number of students exhibited effort on the item, while a value closer to 0 suggests minimal effort from students. As a result, the RTE index reflects an individual student's effort across all items, while the RTF represents the collective effort of all students on a single item. The RTE index signifies effort pertaining to an individual, whereas the RTF indicates effort directed at a particular item. The normative threshold-setting methods were employed to generate RTE and RTF indices and determine individuals' behavioral types toward items. In these methods, the mean item response time for each item is calculated, and threshold points are established by taking a given percentage of these calculated item response times (Wise & Ma, 2012). In this study, thresholds were determined by taking the 10th and 20th percentiles of response times: NT10 and NT20 methods. It is believed that the outcome of the study will contribute to a better understanding of the results from the PISA assessment.

Item response time can be utilized for various purposes, such as item selection in computer-adaptive testing (Lee & Haberman, 2016), its relationship with student motivation (Wise & Kingsbury, 2016), detecting abnormal response behaviors (van der Linden & Guo, 2008), its association with test-taking behaviors (rapid-guessing or solution-based behavior) (Wise, 2006), and serving as an additional source of information to improve the accuracy of ability and item parameter estimations (Petscher et al., 2015; Wise & DeMars, 2006). Understanding the time and effort individuals spend on solving items is crucial for minimizing errors in item and ability parameter estimations, as well as reducing measurement error in test scores. Additionally, analyzing item response times contributes to a deeper understanding of the interaction between respondents and test items (Ju, 2021). This study offers a detailed comparison of item response times from the PISA 2015 and 2018 assessments, focusing on item format, item parameters, and response time fidelity differences. By examining these variations, the research provides insights into how different item types influence response behaviors, contributing to more accurate estimations of student ability and item performance. Furthermore, understanding these time differences helps improve the design of future assessments, ensuring that test validity and fairness are maintained. This analysis also adds to the existing literature by exploring the impact of test formats on response time dynamics in large-scale international assessments.

Purpose of the Study

Within this study, mean response times for trend mathematics items shared between the PISA 2015 and PISA 2018 cycles were compared based on item formats, assessment frameworks in which items were placed, and item parameters. Additionally, RTF indices for items were compared using the NT10 and NT20 methods for both assessment cycles.

Method

Data from 69 trend items in the mathematics subtest of the PISA 2015 and 2018 cycles, specific to Türkiye, were used in this study. Both cycles were conducted in a computer-based format, referred to as Computer Based Assessment (OECD, 2023). In this study, three variables are used: response time, response (coded and scored) and item characteristics (item difficulty, item discrimination and item format). These variables are available in PISA 2015 and 2018 public use datasets and technical reports. Response time, a continuous variable derived from the process data associated with each item, indicates the total duration spent by individual students on the respective items. The scored (by computer) and coded (by human) response variables consist of seven categories as follows: 0 = No Credit, 1 = Full Credit, 5=Valid Skip, 6 = Not Reached, 7 = Not Applicable, 8 = Invalid, and 9 = No Response (OECD, 2023). Categories 5, 6, 7, and 8 are recoded as missing values, and category 9 is recoded as 0, employing the identical coding rules for missing scores as delineated in PISA's methodology (OECD, 2023). Item difficulty is the proportion of correct responses for all items and item discrimination is the item-total correlation. These variables were used to examine the relationship between item characteristics and item response times, as well as to analyze the RTF indices for the items in both 2015 and 2018. Wise (2006) developed the SBij equation to measure Response Time Fidelity (RTF). Here, Ti represents the threshold point that delineates between rapid-guessing behavior and solution behavior for each item i,

while RTj denotes the response time for item i by individual j. The threshold points are determined based on normative threshold values developed by Wise and Kong (2006) using the NT10 and NT20 criteria. NT10 and NT20 cutoffs are obtained by calculating the mean response time for each item separately and taking ten and twenty percent of these calculated mean response times, respectively. Accordingly, the behavior of individual j toward item i is calculated as follows.

$$SB_{ij} = \begin{Bmatrix} 1\ if\ RT_{ij}\ \geq\ T_i \\ 0\ otherwise. \end{Bmatrix} \quad (1)$$

SBij represents the behavior of individual j toward item i and is binary. Accordingly, if an individual's response time for an item is greater than the cutoff point, the behavior is considered solution-oriented. If it is smaller, it's regarded as rapid-guessing behavior.

Response Time Fidelity (RTF), which illustrates individuals' behaviors toward an item, is calculated using SB values for each item by all individuals as follows.

$$RTF_i = \frac{\sum_{j=1}^{n} SB_{ij}}{N} \quad (2)$$

The value N in the formula represents the total number of individuals responding to the item. The Response Time Fidelity (RTF) index, developed based on individual behaviors, has been determined for each item, and the variation between years has been examined. As the items are common, the difference between RTF indices across cycles has been analyzed, and descriptive statistics, graphics, and estimates were created by using “data.table” package in the R software (Barrett et al., 2024). The mean response times of individuals and item characteristics in PISA 2015 and PISA 2018 were compared using a t-test. Spearman's correlation coefficient was used to estimate the correlation between variables. The Wilcoxon Test was used to compare response times of item characteristics over cycles. The information for each trend item has been compared with mean response times, and differences between cycles have been investigated.

Sample

In this study, data from trend mathematics items of Türkiye students used in the PISA 2015 and 2018 cycles were employed. The total number of Türkiye students who participated in the mathematics test was 5895 in 2015 and 6890 in 2018, including 50% female and 50% male students in 2015, and 49.6% female and 50.4% male students in 2018. Due to the presence of only 24 forms containing the 69 trend items related to mathematical literacy, and since not all students received the math items, data from 2408 students were used in PISA 2015 and data from 3718 students were used in PISA 2018. These math trend items were distributed in various forms due to matrix sampling. The sample size for each item ranged from 423 to 741 in 2015 and from 1100 to 1156 in 2018. Türkiye participated in the assessment in a computer-based format (MEB, 2019).

Results

This research examined the relationship between the characteristics of the 69 trend items in the mathematics subtest and item response times across the PISA 2015 and PISA 2018 cycles. Furthermore, the Response Time Fidelity (RTF) index for the trend items was investigated concerning the PISA 2015 and 2018 cycles.

The summary descriptive statistics of the mean response times given by individuals to the 69 trend items in the mathematics literacy subtest of PISA 2015 and PISA 2018 are presented in the table below.

Table 1.
The Summary Results of Item Response Times (in seconds) by year

	PISA 2015	PISA 2018
min	38.14	39.74
mean	99.38	118.26
median	95.03	112.39
max	179.33	208.28

The data in the table reveal an increase in students' response times to the items in 2018 compared to 2015. The mean response time expended by individuals across the 69 mathematics items was 99 seconds in 2015, while it reached 118 seconds in 2018. The mean response times of individuals in PISA 2015 and PISA 2018 were compared using a t-test. As a result of comparing the mean response times between the two groups, the estimated t-value was 3.49. This finding indicates a statistically significant difference in the mean response times between the two groups ($p < .05$).

The mean response time of items was compared across years based on item formats. The obtained results are presented in the table below.

Table 2.
Response Times According to Item Formats Across Years

		2015	2018		
Item Format	N	MRT (sec)	MRT (sec)	Z	p
Simple Multiple-Choice Computer Scored	16	82.20	98.48	-3.516	0.00
Complex Multiple-Choice Computer Scored	13	89.52	103.75	-3.296	0.00
Open Response Computer Scored	22	98.27	113.84	-4,107	0.00
Open Response Human Coded	18	123.13	151.73	-3,724	0.00

As seen in Table 2, the highest mean response times belong to open-ended items scored by scorers for both cycles. The lowest mean response time is for simple multiple-choice items automatically scored by the computer for both cycles. Across all item formats, response times were higher in 2018. Additionally, despite both being open-ended items, the response time for items scored by the computer is lower than the response time for items scored by scorers. To compare the response time difference according to item format, for all item formats, the Wilcoxon signed-rank test was conducted. Response time differences for all item formats over two PISA cycles are statistically significant.

The item response times related to mathematical processes, which are sub-dimensions of mathematical literacy assessment framework, were examined across years. The resulting outcomes are presented in the table below.

Table 3.
Response Times for Mathematical Process Across Years

		2015	2018		
Processes	N	MRT (sec)	MRT (sec)	Z	p
Formulating Situations Mathematically	21	95.62	116.88	-4.107	0.00
Interpreting, Applying and Evaluating Mathematical Outcomes	19	98.15	115.15	-3.823	0.00
Employing Mathematical Concepts, Facts and Procedures	29	102.91	121.30	-4.703	0.00

As seen in Table 3, the mean response time for items related to the employing mathematical concepts, facts, and procedures within mathematical processes is the highest in both cycles. The mean response time increased across all processes in the year 2018. To compare the response time difference across processes, the Wilcoxon signed-rank test was performed for all item processes. According to

mathematical processes of math items, the differences in response times are statistically significant between the two PISA cycles.

The relationship between item response times and item parameters was compared across years. The obtained results are shown in the table below. Item parameters including item difficulty and item discrimination index were calculated according to item response theory.

Table 4.
The Correlations Between Item Parameters and Item Response Times

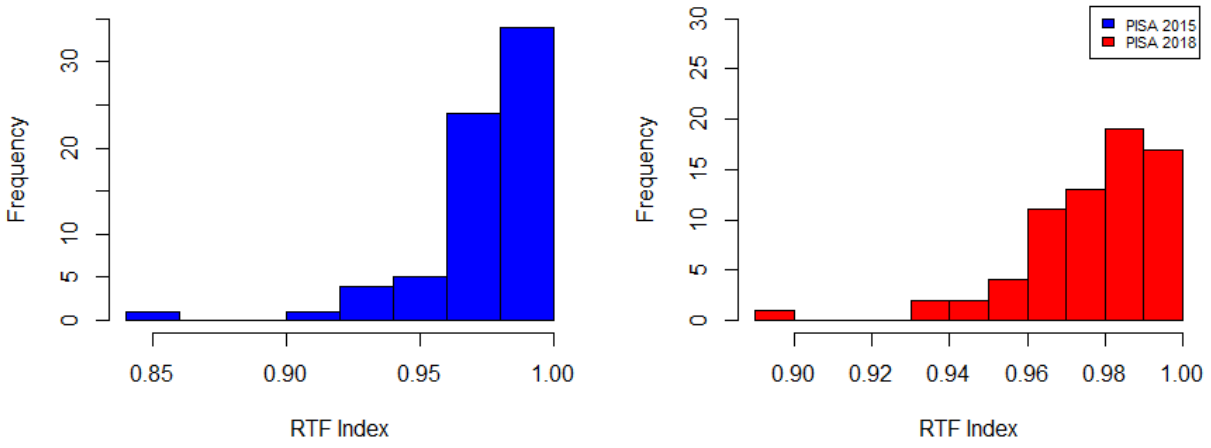
	Year	Item Difficulty	Item Discrimination
Item Response Time	2015	0,26**	0,20
	2018	0,27**	0,21

**p<0,01

Upon reviewing Table 4, a statistically significant positive correlation between the mean item response times and item difficulty is evident for both cycles. The correlation between the variables was moderate. It can be inferred that as items become more challenging, item response times tend to increase. However, no significant correlation was observed between item response times and item discrimination.

RTF indices for items were calculated based on the NT10 and NT20 methods for both cycles. According to the NT10 method, the mean RTF index for 2015 was 0.97. This suggests that 3% of students exhibited rapid-guessing behavior while answering the items. In 2018, the mean RTF index for items reached 0.98, with only 2% of students displaying rapid-guessing behavior. There was a decrease in the percentage of students exhibiting rapid-guessing behavior between the years. The histogram of item RTF indices based on cycles for the NT10 method is depicted below.

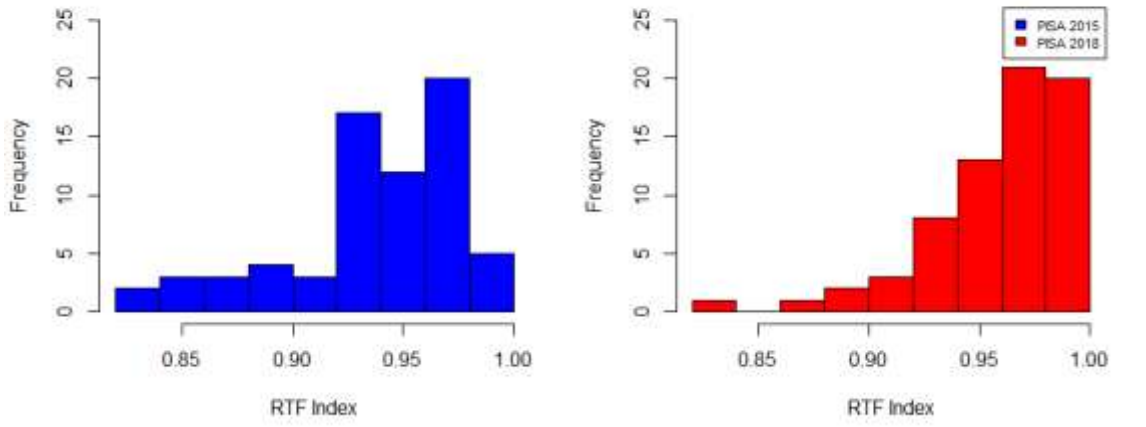
Figure 1.
Histograms of Item RTF Indices Based on Years According to the NT10 Method



According to Figure 1, the RTF indices are skewed to the left in both cycles. For 2015, the RTF indices range from 0.86 to 0.99, while for 2018, they vary between 0.89 and 0.99. Consequently, for each item, rapid-guessing behavior was exhibited by at least one individual.

Below is the histogram illustrating the item RTF indices based on cycles for the NT20 method.

Figure 1.
Histograms of Item RTF Indices Based on Years According to the NT10 Method



Based on the NT20 method, the mean RTF indices in 2015 was 0.94. Consequently, 6% of students showed rapid-guessing behavior while responding to the items. In 2018, the mean RTF index for items was 0.96, and the proportion of students displaying rapid-guessing behavior was 4%. There was a decline in the percentage of students exhibiting rapid-guessing behavior between the years according to this method.

According to Figure 2, the RTF indices obtained using the NT20 method are skewed to the left in both cycles. For 2015, the RTF indices range from 0.82 to 0.99, while for 2018, they vary between 0.84 and 0.99. The majority of individuals have exhibited solution behavior toward the items.

Discussion

In this study, the relationship between the characteristics of the 69 trend items in the mathematics subtest and item response times across the PISA 2015 and PISA 2018 cycles was examined. The mean response time for all items increased in the 2018 cycle, and this increase is statistically significant. Considering the relationship between response time and students' efforts toward the items, it can be suggested that in the 2018 administration, students exhibited more effort and solution behavior while answering the items. Türkiye's mathematics performance also improved in the PISA 2018 administration (MEB, 2019). Eklöf et al. (2014) found a statistically significant relationship between students' efforts obtained through self-reporting and their test performance scores. Therefore, the increase in Türkiye mathematics performance in PISA 2018 should be investigated to determine whether the increase is related to the increase in response time.

In the PISA, the mathematics item formats were analyzed according to the mean item response times, and the response times were compared between the two cycles. The item format that individuals spent the most time solving was open-ended items, while multiple-choice items required less time from individuals for both PISA cycles. Additionally, despite both being open-ended items, the response time for items scored by human raters was longer than for those scored automatically by computers for both PISA cycles. For all item formats, the mean item response time has increased, and response time differences in all item formats over two PISA cycles are statistically significant. A review of the literature revealed that the findings of this study align with the broader trends observed in previous research. Kuang and Sahin (2023) showed that disengagement has focused on one type of

disengagement, namely rapid-guessing, to multiple-choice items. In addition, Wise and Gao (2017) found that selected response item formats led to rapid-guessing and occasional rapid-omits. Yalçin (2022) showed that students spent more time on open-ended questions than multiple-choice questions. Birgili (2014) found that students exerted more effort when responding to open-ended questions compared to multiple-choice ones. Similarly, a study using TIMSS 2015 data (İlhan et al., 2020) revealed that students faced greater difficulty with open-ended items than with multiple-choice items.

Item Response Fidelity (RTF) indices, allowing determination of the proportion of behavior exhibited by individuals toward items, were established for each item. Additionally, RTF indices were compared between the two cycles. The RTF index indicating solution behavior exceeds 80% for each cycle. The majority of individuals displayed solution behavior towards the items. Similarly, the proportion of students displaying rapid-guessing behavior in the PISA 2018 administration decreased compared to the rate in the PISA 2015 administration.

The weighted sub-dimension within the mathematics assessment framework is centered around mathematical processes. These encompass "formulating situations mathematically", "employing mathematical concepts, facts, and procedures"' and "interpreting, applying, and evaluating mathematical outcomes" (OECD, 2023). Item response times were investigated within these processes. It was observed that items related to employing mathematical concepts, facts, and procedures had the longest mean item response times. Within the utilization process, the focus is on how individuals apply mathematical concepts, facts, and procedures in decision-making (MEB, 2019). Consequently, this phase demands individuals to engage their reasoning skills. Given that this skill involves problem analysis, correlating problem stages, making inferences, and proposing solutions, items related to the application of these reasoning skills may naturally require more time due to the complexity of such cognitive processes.

The relationship between mean item response times and item parameters was compared across the years. There was no significant relationship found between item response time and item discrimination in both cycles. However, there was a significant positive relationship between item response time and item difficulty for both cycles. Similar to these findings, Altuner (2019), demonstrated in their study a negative moderate relationship between item response time and item difficulty index and a significant relationship with the item discrimination index. Consistent with this study, item response time tends to increase for difficult items, while it significantly decreases for easier items.

The fact that PISA is not considered a high-stakes assessment for students may have influenced their response times and effort levels. Additionally, students in this study may not have been fully aware of the time constraints. It is important to note that the current study's findings are limited to students who answered mathematics trend items across two PISA cycles. Future research could extend this analysis to include items from all PISA tests to identify broader patterns. This study highlighted that various factors influence students' response times. Therefore, it is recommended that future studies examine the effect of response time on item and ability parameters. Additionally, leveraging recent technological advancements, such as eye-tracking devices, could provide a more detailed understanding of response times. Finally, this study used non-parametric tests due to the violation of normality assumptions and sample size. Should these assumptions be met in future studies, results could be compared using parametric methods.

In the study, a pattern of rapid-guessing behavior was observed more frequently in multiple-choice items, which indicates a high likelihood of guesswork. Therefore, it is recommended that the predominant use of short-response items, particularly for assessing the skills measured by the items where rapid-guessing is observed, be adopted. This approach could lead to more accurate estimations of students' ability levels.

Declarations

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as this study uses data shared with the public.

Author Contribution: Muhsin POLAT: conceptualization, investigation, methodology, data analysis, visualization, writing - review & editing. Hülya KELECİOĞLU: conceptualization, methodology, supervision, writing - review & editing.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington DC.

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/

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., & Hocking, T. (2024). *data.table: Extension of 'data.frame'*. R package version 1.14.8. https://CRAN.R-project.org/package=data.table

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, 39(6), 502-523. doi:10.3102/1076998614558485

Debeer, D., Janssen, R., & Boeck, P. D. (2017). Modeling skipped and not-reached items using IRTrees. *Journal of Educational Measurement*, 54(3), 333-363. doi:10.1111/jedm.12147

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, 13(1), 55-77. doi:10.1207/s15324818ame1301_3

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross- national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education*, 27(1), 31-45. doi:https://doi.org/10.1080/08957347.2013.853070

Kuang, H., & Sahin, F. (2023). Comparison of disengagement levels and the impact of disengagement on item parameters between PISA 2015 and PISA 2018 in the United States. *Large-scale Assessments in Education*, 11(4). doi:10.1186/s40536-023-00152-0

Lee, Y. H., & Haberman, S. J. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing*, 16(3), 240-267. doi: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(8), 1-24. doi:10.1186/s40536-014-0008-1

MEB. (2019). *PISA 2018 Türkiye Ön Raporu*. Ankara: Milli Eğitim Bakanlığı.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education*, 9(1), 3-8. doi:10.1207/s15324818ame0901_2

Michaelides, M. P., & Militsa, I. (2022). Response time as an indicator of test-taking effort in PISA: country and item-type differences. *Psychological Test and Assessment Modeling*, 64(3), 304-338.

OECD. (2023). *OECD-PISA*. Retrieved from PISA 2018 Technical Report: https://www.oecd.org/pisa/data/pisa2018technicalreport/

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. doi:10.1007/s11145-014-9518-z

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: evidence from the English version of the PISA 2015 science test. *Large-scale Assessments in Education*, 9(10), 1-31. doi:10.1186/s40536-021-00104-6

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, 33(4), 263-279. doi:10.1080/08957347.2020.1789141

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response times with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement*, 34, 213-232. doi:10.1111/j.1745-3984.1997.tb00516.x

Setzer, J. C., Wise, S. L., Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education*, 1, 34-49. doi:10.1080/08957347.2013.739453

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

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, 19(2), 95-114. doi:10.1207/s15324818ame1902_2

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. doi:10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education*, 30(4), 343-354. doi:10.1080/08957347.2017.1353992

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. doi:10.1111/jedm.12102

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. doi:10.1207/s15324818ame1802_2

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. *Paper presented at the 2012 annual meeting of the national council on measurement in education.* Vancouver: Canada.

Wise, S. L., Bhola, D. S., & Yang, S.-T. (2006). Taking the time to improve the validity of low-stakes tests: the effort-monitoring CBT. *Educational Measurement Issues and Practice*, 25(2), 21-30. doi:10.1111/j.1745-3992.2006.00054.x