International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 174–193
https://doi.org/10.21449/ijate.1343248
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Examination of response time effort in TIMSS 2019: Comparison of Singapore and Türkiye

Esin Yilmaz Kogar [1*], **Sumeyra Soysal** [2]

[1]Niğde Ömer Halisdemir University, Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Niğde, Türkiye
[2]Necmettin Erbakan University, Ahmet Keleşoğlu Faculty of Education, Department of Educational Sciences, Division of Educational Measurement and Evaluation, Konya, Türkiye

ARTICLE HISTORY
Received: Aug. 15, 2023
Accepted: Dec. 12, 2023

Keywords:
Response time effort,
Rapid guessing behavior,
Solution behavior,
Cognitive domain,
Content domain.

Abstract: In this paper, it is aimed to evaluate different aspects of students' response time to items in the mathematics test and their test effort as an indicator of test motivation with the help of some variables at the item and student levels. The data consists of 4th-grade Singapore and Turkish students participating in the TIMSS 2019. Response time was examined in terms of item difficulties, content and cognitive domains of the items in the mathematics test self-efficacy for computer use, home resources for learning, confident in mathematics, like learning mathematics, and gender variables at the student level. In the study, it was determined that all variables considered at the item level affected the response time of the students in both countries. It was concluded that the amount of variance explained by the student-level variables in the response time varied for each the country. Another finding of the study showed that the cognitive level of the items positively related to the mean response time. Both Turkish and Singaporean students took longer to respond to data domain items compared to number and measurement and geometry domain items. Additionally, based on the criterion that the response time effort index was less than .8, rapid-guessing behavior, and therefore low motivation, was observed below 1% for both samples. Besides, we observed that Turkish and Singaporean students were likely to have rapid guessing behavior when an item in the reasoning domain became increasingly difficult. A similar result was identified in the data content domain, especially for Turkish graders.

1. INTRODUCTION

Today's conditions have brought about the necessity of digitalisation in many areas of human life such as trade, health and education. Especially, in the field of education, both courses and exams have started to be conducted with online or computer-based applications, and these applications have become more common and important in recent years. Computer-based applications have also been included in the cycle of large-scale international assessments such as the Trends in International Mathematics and Science Study (TIMSS), the Programme for International Student Assessment (PISA). In 2019, eTIMSS was added to TIMSS as a computer-based "eAssessment system", measuring the same mathematics and science

*CONTACT: Esin YILMAZ KOGAR esinyilmazz@gmail.com Niğde Ömer Halisdemir University, Faculty of Education University, Faculty of Education, Department of Educational Sciences, Niğde, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

constructs using the same assessment items as possible with "paperTIMSS", which is in the paper-and-pencil format as in previous TIMSS cycles (Mullis et al., 2016). Besides, eTIMSS provides more detailed information about students and allows different assessments to be made.

Computer-based testing applications, such as eTIMSS, allow for the collection of a chronology of test takers' interactions with test items throughout the assessment process (Organisation for Economic Co-operation and Development [OECD], 2015). Moreover, such applications make it possible to obtain measures of response time per item, which is difficult to measure in pen-and-paper assessments. The term response time (RT) refers to the time it takes test takers to respond (react) to a particular item (stimulus) in the test (Lee & Chen, 2011). This enables the analysis of test-takers' efforts to take the test through objective records of their actions rather than relying on self-reported assessments of their behaviour (Lee & Jia, 2014; Wise & Kong, 2005). Item response time, which is more easily obtainable through computer adaptive testing, helps understand which factors affect how quickly an examinee answers an item, and thus, helps test developers estimate the time required to answer the total of the test (Bergstrom et al., 1994). The use of process data from large-scale tests, such as eTIMSS, PISA, to determine test effort or grader behaviour has great potential for educational assessment.

The common purpose of large-scale international assessments is to evaluate education systems worldwide by testing and analyzing the abilities and understanding of students of different ages in participating countries/economies. Since the scores on such tests, so-called low-stakes tests, do not indicate any personal conclusions about the test taker's performance, individuals may be reluctant to demonstrate the full range of their knowledge, skills or attitudes. Therefore, since it is unclear whether test takers are motivated enough, test scores may not represent their true ability level and may not serve as a valid measure of their abilities. In this context, many researchers have investigated the function of test-taking motivation during low-stakes assessments of their performance (e.g., Barry & Finney, 2009; Eklöf, 2007; Wise & Kong, 2005; Wise & DeMars, 2005). For this purpose, Wise and Kong (2015) established a relationship between motivation to take the test and response time.

Unlike the studies that relied on examinee self-reports to measure test-taking effort, Wise and Kong (2005) developed a measure, called response time effort (RTE), using item response times, which represents a direct observation of the test taker's behavior and whose collection is unobtrusive and nonreactive (examinees with computer-based test will typically be unaware). Baumert and Demmrich (2001, p.441) defined this term as the following: “test-taking effort as a student’s engagement and expenditure of energy toward the goal of attaining the highest possible score on the test.” Test-taking motivation is also identified as “the willingness to engage in working on test items and to invest effort and persistence in this undertaking”; in short, it is the motivation of the individual to achieve a high-performance level in a test (Eklöf, 2010). Based on the relationship between these two phenomena, Wise and Kong (2005) and Wise (2017) interpreted the RTE as an indicator of test motivation, which contains two types of behaviour. The first of these behaviours is characterized by active engagement in seeking the correct answers to test items, known as solution behavior (SB), while the second is marked by quick responses in a mostly random manner, referred to as rapid-guessing behaviour. Accordingly, the researchers assume that less-motivated examinees are likely to exhibit rapid-guessing behavior, while high-motivated examinees are likely to display SB.

Exams such as the Scholastic Aptitude Test and the American College Test conducted in the USA are high-stake tests where students try getting admission from a college or university, and their performance in these exams directly concerns them (Sundre & Kitsantas, 2004). In contrast, international large-scale assessments classified as low-stake tests (PISA, TIMMS, NAEP, PIRLS, etc.) provide national-level reports without individual results for students, teachers, or parents. Therefore, students' motivation to take the exam may emerge as a problem.

In the literature, the variability of test-taking effort and motivation of the examinees during a low-stakes assessment was examined in various aspects and conditions. Some researchers have focused on the relationship between test-taking effort, test motivation and test performance (e.g., Cole et al., 2008; Lundgren & Eklöf, 2020; Slim et al., 2020; Wise & DeMars, 2005), some researchers have examined predictor variables of examinee's RT (e.g., Baumert & Demmrich, 2001; Gershon et al., 1993; Lundgren & Eklöf, 2020; Wolgast et al., 2020) or properties of the test affected by the RT (e.g., Fan et al., 2012; Wang & Hanson, 2005; Weirich et al., 2017). The number of studies that test effort based on item-response time is more than the ones that are based on self-report measures. The majority of studies employing RT indices have consistently found that a significant percentage of examinees, ranging from 74% to 99%, demonstrated response time values greater than .90. However, it has been noted that the behavior of 1-23% of examinees raised concerns, as they displayed rapid-guessing behaviour on more than 10% of the test items (e.g., Setzer et al., 2013; Swerzewski et al., 2011; Wise & DeMars, 2005; Wise & Kong, 2005). These findings demonstrate that when RT indices were utilized as a measure of test effort, the majority of examinees consistently displayed diligent efforts in answering the items, with minimal within-examinee variation in their levels of effort throughout the test. Wise and DeMars (2005) raised concerns about the potential for examinees to provide inaccurate or insensible responses on self-report scales, perhaps, which may have led to the limited use of self-report methods to examine test-taking efforts in a few studies (Barry et al., 2010; Myers & Finney, 2021; Wolgast et al., 2020).

In the world, which has already been increasingly digitilized for the last few decades, the place and importance of computerized and online learning environments in education has gradually increased with the significant impact of the COVID-19 pandemic. Besides, the effect of digitalization could be seen in international large-scale assessments such as the TIMSS. In the last few cycles of these assessments, a paper-pencil format or a computer-based "e" version has been presented to selected countries. However, over 50% of the 64 nations involved in the TIMSS 2019 opted to conduct the "e" version of the assessments, while the remaining countries followed the traditional approach of administering the TIMSS using pen and paper, as done in previous cycles (Martin et al., 2020).

Although the most basic variable that may have an effect on students' RT is the ability level of the student, different variables may also affect RT. For example, studies in the literature show that RT is also related to different item-level variables such as item difficulty level, content area, and conginitve domain (Bridgeman, & Cline, 2000; Goldhammer et al., 2014; Hess et al., 2013; İlgün-Dibek, 2020; Lee & Jia, 2014; Wang, 2017; Yalçın, 2022; Zenisky & Baldwin, 2006). Besides, student-level variables may also be related to RT. Among such variables, there are studies addressing gender (Hess et al., 2013; İlgün-Dibek, 2020; Setzer et al., 2013) and self-confidence (Yalçın, 2022) in terms of RT. Cooper (2006) and Zhang et al. (2016) reported that test outcomes were affected by students' comfort and self-confidence levels in using computers and tablets. Considering this situation, we wanted to examine the effect of computer self-efficiency computer on response time in our study. In addition, in TIMSS 2019, countries, not students, decided which version of the paper TIMSS or eTIMSS would be implemented in countries. Given that not all individuals have the same opportunities, students who are not familiar with the use of such digital devices may experience difficulties in computer-based assessments (Bennet et al., 2008; Chen et al., 2014; Pommerich, 2004). Since this familiarity is obviously related to home resources, the home resources variable was also considered in the study. Although studies on RT analyses are available in the literature, there are fewer studies on RT analyses for country comparisons (see, İlgün-Dibek, 2020; Rios & Guo, 2020; Michaelides et al., 2020). Therefore, in the current study, we aimed to gather evidence on possible differences in response time efforts between countries, which is intended to increase the validity of cross-country comparisons. Thus, we believe that this study will contribute to

the literature on cross-country comparisons of response time effort. We have summarised the aim and sub-problems of our research in detail below.

1.1. Purpose of the Study

The aim of the current study is to evaluate different aspects of students' RT and test-taking motivation using some item- and student-level variables, based on data from the 2019 TIMSS 4th grade samples from Türkiye and Singapore. In this context, the following subquestions, which the study is intended to answer, are presented as follows:

1) Does the mean response time of students in the mathematics test, each for Türkiye and Singapore samples, significantly differ according to content domain, cognitive domain, and item difficulty to which the items belong?
2) Is the mean response time of students in the mathematics test, each for Türkiye and Singapore samples, significantly predicted by self-efficacy for computer use, home resources for learning, like learning mathematics and gender?
3) How is the response time effort of students in the mathematics test, each for Türkiye and Singapore samples?

2. METHOD

2.1. Datasets

The data of the study consists of 4th-grade Singapore and Türkiye students participating in the TIMSS 2019, which can be downloaded from the International Association for the Evaluation of Educational Achievement (IEA) website. The reason why we included Türkiye and Singapore in this research is that we wanted to compare Singapore, which had the highest performance in mathematics with 625 points, with our country which ranked 23rd with 523 points. 5986 students (47.9% girl) participated in Singapore and 4028 students (52.1% girl) participated within Türkiye. The total number of items analyzed was 159. For 27 derived items where students were asked to give more than one answer or a multi-part answer, the response time (total time on screen as seconds) was divided by the number of items contained in the derived item. Similarly, item difficulty statistic for a derived item was rearranged to represent the mean difficulty of the items it contained.

2.2. Variables in Interest

2.2.1. *Item-level variables*

2.2.1.1. Content Domain (CnD). One of the dimensions, which each of the paper TIMSS and eTIMSS assessment frameworks is organized around and specifies the subject matter to be assessed. In the 4th-grades, a mathematics test consists of three content domains, which are apportioned as follows: number (50%), measurement and geometry (30%) and data (20%) (Martin et al., 2020). In this study, the items were coded as 1 = numbers, 2 = measurement and geometry, 3 = data through the analysis.

2.2.1.2. Cognitive Domain (CD). Paper TIMSS and e-TIMSS assessment frameworks are each structured around a dimension that outlines the specific cognitive processes to be assessed. For all grades, a mathematics test consists of three cognitive domains, which is apportioned as follows: knowing (40%), applying (40%) and reasoning (20%) (Martin et al., 2020). In this study, the items were coded as 1 = knowing, 2 = applying, 3 = reasoning through the analysis.

2.2.1.3. Item Difficulty (*p*). It was calculated by dividing the number of test takers who answered correctly by the total number of test takers. Item percent correct statistics was used from the TIMSS 2019 International Database. Although there are different classifications for item difficulty index, the three-category classification as referred by Crocker and Algina (1986,

p.324) was used in this study to avoid too many categories: hard (0 to .39), moderate (.40 to .60) and easy (.61 to 1.00). For item-level analysis, the distribution of items across the three cognitive domains and the three content domains by item difficulty is summarized in Table 1.

Table 1. *Descriptive statistics for items.*

Country	Cognitive Domain	Content Domain	Item Difficulty			Total
			Hard	Moderate	Easy	
Singapore	Knowing	Numbers	-	-	31	31
		Measurement and Geometry	-	3	15	18
		Data	-	-	8	8
		Total	-	3	54	57
	Applying	Numbers	1	2	34	37
		Measurement and Geometry	-	3	14	17
		Data	-	1	12	13
		Total	1	6	60	67
	Reasoning	Numbers	1	8	4	13
		Measurement and Geometry	3	3	8	14
		Data	1	1	6	8
		Total	5	12	18	35
Türkiye	Knowing	Numbers	3	9	19	31
		Measurement and Geometry	4	7	7	18
		Data	1	3	4	8
		Total	8	19	30	57
	Applying	Numbers	6	21	10	37
		Measurement and Geometry	3	11	3	17
		Data	3	6	4	13
		Total	12	38	17	67
	Reasoning	Numbers	9	3	1	13
		Measurement and Geometry	5	7	2	14
		Data	2	2	4	8
		Total	16	12	7	35

As shown in Table 1, each cognitive domain and content domain contained a considerable number of items. The number of items for Numbers, Measurement and Geometry, and Data content domains is 81, 49 and 29, respectively. The number of items for knowing, applying and reasoning cognitive domains is also 57, 67 and 35, respectively. For Singapore sample, while the percentage of correct answers to the items in the knowing and applying domains by the students is quite high, the items within the reasoning domain are a substantial amount of medium and easy difficulty level. Additionally, it can be said that Singaporean students do not have difficulty in the data domain, but they have some difficulties in the measurement and geometry domain. In the Türkiye sample, the items within knowing were mostly on the easy difficulty level, while the items within applying and reasoning were classified as medium and high. It was observed that Turkish students had more difficulties as the cognitive domain level of the contents increased.

2.2.2. *Student-level variables*

2.2.2.1. Gender. This variable was coded as 1 = Girl and 2 = Boy throughout the analysis.

2.2.2.2. Students Like Learning Mathematics (SLM). The scale has nine items with a 4-point response key ranging from agree a lot to disagree a lot, which covers students' attitudes toward mathematics and studying mathematics. The total scale score is divided into three categories: very much like (score at or above 10.2), somewhat like (between 10.2–8.4) and do not like (at or below 8.4). The percentages of students in Singapore to this variable are as follows: 36.9% very much like mathematics learning, 40.0% somewhat like learning mathematics, and 22.9% do not like learning mathematics. For Türkiye, the classification is as follows: 64.8% very much like learning mathematics, 25.4% somewhat like learning mathematics, and 9.2% do not like learning mathematics.

2.2.2.3. Students Confident in Mathematics (SCM). This scale measures how confident students feel about their ability in mathematics, in terms of their level of agreement with nine statements with a 4-point response key ranging from agree a lot to disagree a lot. The total scale score is divided into three categories: very confident (score at or above 10.7), somewhat confident (between 10.7–8.5), and not confident (at or below 8.5). The percentages of students in Singapore by this variable are as follows: 20.7% very confident in mathematics, 42.0% somewhat confident in mathematics, and 37.1% not confident in mathematics. For Türkiye, the classification is as follows: 33.2% very confident in mathematics, 41.3% somewhat confident in mathematics, and 23.4% not confident in mathematics.

2.2.2.4. Home Resources for Learning (HRL). This measurement scale combines data gathered from fourth-grade students and their parents. The students supplied details regarding the number of books and other study supports in their households, while the parents provided information concerning the number of children's books, the educational levels of the parents, and the occupational status of the parents. The total scale score is divided into three categories: many resources (score at or above 11.8), some resources (between 11.8–7.4) and few resources (at or below 7.4). High scores indicate that the student has more home resources. The percentages of students in Singapore by this variable are as follows: 28.3% many resources, 66.6% some resources, and 1.7% few resources. For Türkiye, the classification is as follows: 4.6% many resources, 64.1% some resources, and 24.3% few resources.

2.2.2.5. Self-Efficacy for Computer Use (SEC). We could not find any detailed information for this variable in the TIMSS 2019 technical report. As data used in this paper in Singapore, 50.0% of the students are in the high self-efficacy, 39.3% of the students are in medium self-efficacy, 2.4% of the students are in low self-efficacy category. In Türkiye, 63.8% of the students are in the high self-efficacy, 33.0% of the students are in medium self-efficacy, 2.6% of the students are in the low self-efficacy category.

2.2.2.6. Response Time Effort (RTE). As explained earlier, the test-taking effort is assessed by analyzing item response times, focusing on two distinct behaviors known as *solution behavior (SB)* and *rapid-guessing behavior*. SB refers to cases where examinees put effort into answering an item thoughtfully. On the other hand, examinees who quickly respond without sufficient time for reading and full consideration of an item exhibit rapid-guessing behavior (Wise & Kong, 2005). Thus, SB is considered effortful response strategies, while rapid guesses are seen as non-effortful strategies. The 10% normative threshold (NT10) methodology proposed by Wise and Ma (2012) was used to determine whether each student showed SB or rapid-guessing behaviour over the answering time. As part of this approach, the initial step involved computing the average response time for each item, and then 10% of this value was used as a threshold. However, according to the recommendation made by Setzer et al. (2013),

it is advised to employ a maximum threshold of 10 seconds when utilizing this methodology. Therefore, we established the maximum threshold at 10 seconds. After one threshold (T_i) was determined for each item, the following steps were followed: For item i, there is a threshold, T_i, that represents the response time boundary between rapid-guessing behavior and solution behavior. Given an examinee *j*'s response time, RT_{ij}, to item *i*, a dichotomous index of item solution behavior, SB_{ij}, is computed as in Equation 1 (Wise & Kong, 2005, pp.167-168).

$$SB_{ij} = \begin{cases} 1 \; if \; RT_{ij} \geq T_i, \\ 0 \quad otherwise, \end{cases} \tag{1}$$

The 10% normative threshold (NT10) methodology, recommended by Wise and Ma (2012), was used to determine which behavior each student displayed according to the response time. After calculating all $SB_{ij,}$ we computed RTE indices as a summary measure of effort for a test (Wise & Kong, 2005). More precisely, the RTE indicates the percentage of items in which an examinee demonstrates solution behavior. As denoted in Equation 2, the overall RTE index for examinee *j* on the test is calculated as follows:

$$RTE_j = \frac{\sum SB_{ij}}{k}, \tag{2}$$

where *k* is equal to the number of items in the test. RTE scores range between 0 and 1, reflecting the proportion of test items for which the examinee demonstrated SB. Consequently, higher RTE values suggest that the examinee likely approached the test items with sufficient effort, while lower RTE values indicate a lack of substantial effort from the examinee (Setzer et al., 2013). For interpreting the behavior type of each grader, the RTE score was divided into three categories: high effort (above .90), medium effort (between .90–.80) and low effort (below .80) in this study (Wise & Kong, 2005). Then, robustly, the grader in the low effort category was acknowledged as having rapid-guessing behavior, and the ones in the high effort category had SB.

2.3. Analysis Procedures

The Statistical Package for the Social Science (SPSS) version 24.0 (IBM Corp., Armonk, NY) program was used for the item-level analysis. First, mean RTs for content and cognitive domains were determined for each item. In the TIMSS 2019 data, there also are many items that were combined, or derived, for scoring purposes, which are called derived items. In our study, the RTs for the derived items were calculated by dividing the number of items contained in these items. Then, the difficulty of each item was calculated. The factorial analysis of variance (ANOVA) was conducted used to determine whether the mean RTs differed according to cognitive domain, content domain, and difficulty level. To determine which categories of these variables differed significantly from each other, the Scheffe test was used, which is one of the post hoc tests. In terms of the second research question, software named the International Association for the Evaluation of Educational Achievement (IEA) International Database Analyzer (IDB) (IEA, 2017) was used to conduct multiple regression analysis, sampling design, sampling weights and plausible values should be considered when analyzing large-scale assessments such as the TIMSS to avoid biased results. With the IDB Analyzer, which made this possible, student total weights (TOTWGT) were used in student level analyses. With the IDB Analyzer, SPSS syntax that considers the sampling weights was generated and multiple regression analysis was performed on SPSS using this syntax. SLM, SCM, HRL and SEC were continuous predictor variables and gender was a dummy-coded predictor variable where girls were the reference group. If the absolute value of the *t*-test is greater than 1.96, the result can be regarded as statistically significant ($p < .05$). Therefore, significance tests are conducted by *t*-value. Partial eta squared (η^2) effect sizes were calculated to determine the proportion of

unique variance of each variable in the analysis. The effect sizes were interpreted using the following benchmarks given by Cohen (1988): small (.01), medium (.06), and large (.14).

3. FINDINGS

3.1. Findings for the First Research Question

First, for item-level analysis, descriptive statistics of mean RT by item difficulty, cognitive domain, and content domain are summarized in Table 2.

Table 2. *Descriptive statistics of mean response time by item difficulty, cognitive domain, and content domain.*

		Item Difficulty							
		Hard		Moderate		Easy		Total	
Country	CD by CnD	M	SD	M	SD	M	SD	M	SD
Singapore	Knowing								
	Numbers	-	-	-	-	37.43	17.21	37.43	17.21
	M & G	-	-	40.64	1.48	38.03	13.37	38.46	12.18
	Data	-	-	-	-	64.99	13.81	64.99	13.81
	Total	-	-	40.64	1.68	41.68	18.34	41.63	17.85
	Applying								
	Numbers	66.14	-	67.98	35.40	52.64	16.99	53.83	17.78
	M & G	-	-	55.45	24.86	65.97	37.25	64.11	34.95
	Data	-	-	112.51	-	74.64	27.30	77.55	28.17
	Total	66.14	-	69.14	31.42	60.15	26.26	61.04	26.42
	Reasoning								
	Numbers	207.94	-	128.77	48.96	64.69	28.72	115.15	57.40
	M & G	102.85	68.26	73.01	20.39	47.01	14.99	64.55	38.03
	Data	182.08	-	138.52	-	79.71	56.80	99.86	61.90
	Total	139.72	70.44	115.64	47.64	61.84	37.48	91.41	55.08
	Total	127.46	69.80	91.64	48.82	52.82	26.75	60.77	37.38
Türkiye	Knowing								
	Numbers	66.47	29.57	50.63	13.77	53.14	23.94	53.70	21.73
	M & G	48.21	2.59	52.57	11.15	53.15	22.27	51.83	14.97
	Data	117.92	-	85.71	5.67	75.57	17.33	84.67	18.54
	Total	63.77	28.51	56.89	17.16	56.13	23.44	57.45	22.08
	Applying								
	Numbers	70.97	30.37	73.28	20.42	68.76	20.45	71.69	21.64
	M & G	72.74	33.70	94.39	34.97	42.79	11.51	81.47	36.50
	Data	114.15	50.04	114.88	46.41	88.04	36.68	106.45	42.60
	Total	82.21	38.11	85.96	33.09	68.71	26.90	80.91	32.93
	Reasoning								
	Numbers	140.83	45.50	106.94	51.31	46.24	-	125.73	51.03
	M & G	94.66	46.19	64.38	26.51	61.57	5.67	74.79	34.93
	Data	148.52	26.50	71.84	6.44	102.14	67.04	106.16	53.77
	Total	127.36	47.38	76.26	34.86	82.57	53.62	100.88	50.04
	Total	98.18	48.25	76.27	32.06	63.52	30.53	76.90	37.89

Note. CD = Cognitive Domain, CnD = Content Domain, M = Mean, SD = Standard Deviation, N = Item Numbers, M & G = Measurement and Geometry.

Table 2 displays the mean and standard deviation of mean RT by item difficulty, cognitive domain, and content domain. Whether the differences observed in Table 2 were statistically significant or not was examined by factorial ANOVA and the main and interaction effects on mean RT are presented in Table 3.

Table 3. *Factorial ANOVA of mean response times by cognitive domain, content domain and item difficulty.*

Country	Source	*df*	MS	F	η^2	Difference
Singapore	CD	2	6860.14	9.72**	.12	K<A<R
	CnD	2	8847.86	12.54**	.15	N<D, M&G<D
	P	2	9168.87	12.99**	.16	E<M<H
	CD x CnD	4	825.23	1.17	-	
	CD x P	3	2762.13	3.91**	.08	
	CnD x P	4	1723.34	2.44*	.07	
	CD x CnD x P	2	128.35	.18	-	
	Error	139	705.66			
	R^2 = .56; adj R^2 = .50					
Türkiye	CD	2	4647.91	5.18**	.07	K<A<R
	CnD	2	9912.91	11.05**	.14	N<D, M&G<D
	P	2	6576.03	7.33**	.10	E<M<H
	CD x CnD	4	550.11	.61	-	
	CD x P	4	3058.55	3.41*	.09	
	CnD x P	4	812.12	.91	-	
	CD x CnD x P	8	938.90	1.05	-	
	Error	132	897.09			
	R^2 = .48; adjR^2 = .38					

Note. CD = Cognitive Domain, CnD = Content Domain, P = Item Difficulty, MS = Mean squares, η^2 = Effect Size, K = Knowing, A = Applying, R = Reasoning, N = Numbers, M & G = Measurement and Geometry, D = Data, H = Hard, M = Moderate, E = Easy, R^2 = .556 and adj R^2 = .495 for Singapore, R^2 = .478 and adj R^2 = .375 for Türkiye, *p < .05. ** p < .01.

As shown in Table 3, all main effects (the cognitive domain ($F_{Singapore}$ (2, 139) = 9.72, p < .01; $F_{Türkiye}$ (2, 132) = 11.05, p < .01), content domain ($F_{Singapore}$ (2, 139) = 12.54, p < .01; $F_{Türkiye}$ (2, 132) = 11.05, p < .01) and the item difficulty ($F_{Singapore}$ (2, 139) = 12.99, p < .01; $F_{Türkiye}$ (2, 132) = 7.33, p < .01), is significantly affected on the mean RT for both samples. According to the Scheffe test for both samples, the source of the differences was the mean RT increased from knowing to reasoning, from easy to hard, and the mean RT of the data content area was higher than that of the other content areas (see Table 2). In terms of two-way interaction, there was only a significant interaction between cognitive domain and item difficulty in the Türkiye sample, and besides a significant interaction between content domain and item difficulty in the Singapore sample, as well. Three-way-interactions did not statistically affect the mean response time. With a large effect size, the highest proportion of the variance of the mean response time in the Singapore sample was attributed to item difficulty, content domain and cognitive domain, respectively, whereas, in the Turkish sample, they were content domain, item difficulty, and cognitive domain, respectively.

3.2. Findings for the Second Research Question

The findings related to the prediction of the mean RT of the items in the mathematics achievement test according to the student-level variables are given in Tables 4 and 5.

Table 4. *Multiple regression results by content domain.*

Variables	Country											
	Singapore						Türkiye					
	Number		M & G		Data		Number		M & G		Data	
	B	β	B	β	B	β	B	β	B	β	B	β
HRL	-1.45	**-.11**	-.52	**-.04**	-.22	-.01	-1.48	**-.13**	-1.04	**-.09**	-1.64	**-.08**
SCM	-.95	**-.09**	.23	.02	.21	.01	-.78	**-.08**	.72	**.07**	.52	.03
SEC	-.75	**-.07**	-.74	**-.06**	-1.20	**-.07**	-.30	-.03	-.36	-.03	.54	.03
SLM	.39	**.04**	.40	**.04**	.48	.03	.95	**.08**	-.07	-.01	.67	.03
Gender[a]	-4.01	**-.10**	-2.67	**-.06**	-5.71	**-.09**	-4.79	**-.11**	-2.13	**-.05**	-2.26	-.03

Note. [a] Girl = 1, Boy = 2. Significant standardized weights ($p < .05$) are bold. HRL: Home Resources for Learning, SCM: Students Confident in Mathematics, SEC: Self-Efficacy for Computer Use, SLM: Students Like Learning Mathematics, M & G = Measurement and Geometry

Table 4 displays the outcomes of the multiple regression analyses conducted on the content domain. The noteworthy negative β weights associated with each predictor variable reveal that students who achieved higher scores in these variables exhibited reduced mean RTs during the TIMSS 2019. Conversely, the significant positive β weights for each predictor variable indicate that students with higher scores in these variables demonstrated increased mean RTs in the TIMSS 2019. Besides, when the results for the gender variable, which is a categorical variable, were negative, it was determined that the RTs of girls were longer than that of boys. But these variables explained 5% of the variance of the mean RT for Singapore (R^2 = .05) and 4% for Türkiye (R^2 = .04). For the number content domain, similar results in both countries for five independent variables were obtained to be significant result (standardized β weight ranges from -.13 to .08), only not for Türkiye for SEC.

For the measurement and geometry content domain, all variables without SCM significantly predicted the mean RT (standardized β weight ranges from -.06 to .04) for the Singapore sample. For Türkiye, only two of the five variables (SEC and SLM) had a non-significant effect on predicting mean RT (standardized β weight ranges from -.09 to .07). But these variables explained only 1% and 2% of the variance of mean RT for Singapore (R^2 = .01) and Türkiye (R^2 = .02), respectively.

For the data content domain, only SEC and gender variables were found significant for Singapore sample (standardized β weight ranges from -.09 to -.07) and only HRL had a significant influence for the Türkiye sample (standardized β weight -.08). But these variables were a part of the variance of the mean RT only with 1% for both samples (R^2 = .01).

Table 5. *Multiple regression results by cognitive domain.*

Variables	Country											
	Singapore						Türkiye					
	Knowing		Applying		Reasoning		Knowing		Applying		Reasoning	
	B	β	B	β	B	β	B	β	B	β	B	β
HRL	-1.24	**-.13**	-1.41	**-.12**	.79	**.03**	-1.35	**-.14**	-2.33	**-.21**	.79	.04
SCM	-.74	**-.10**	-.90	**-.10**	1.15	**.06**	-.40	-.05	-.74	**-.07**	1,76	**.10**
SEC	-.69	**-.09**	-.81	**-.08**	-1.04	**-.05**	-.47	**-.05**	.13	.01	-.57	-.03
SLM	.40	**.05**	.45	**.05**	.44	.02	.40	.04	.56	**.05**	1.14	**.06**
Gender[a]	-2.83	**-.10**	-1.82	**-.05**	-9.26	**-.12**	-1.46	**-.04**	-3.22	**-.08**	-6.53	**-.09**

Note. [a] Girl = 1, Boy = 2. Significant standardized weights ($p < .05$) are bold. HRL: Home Resources for Learning, SCM: Students Confident in Mathematics, SEC: Self-Efficacy for Computer Use, SLM: Students Like Learning Mathematics

Table 5 displays the results of the multiple regression analyses for the cognitive domain. For the knowing domain, all variables were significant predictors (standardized β weight ranges from -.13 to .05) and shared 5% of the variance of the mean RT for Singapore (R^2 = .05). For Türkiye, only two of the five variables (SEC and SLM) had a non-significant effect on predicting mean RT (standardized β weight ranges from -.14 to -.04) and the explained variance was R^2 = .05.

For the applying domain, in Singapore sample, five independent variables were obtained to be significant results (standardized β weight ranges from -.12 to .05). In Türkiye sample, SEC did not have significant standardized β *coefficient (.01; p > .05)*. These variables explained 5% of the variance of the mean RT for Singapore (R^2 = .05) and 4% for Türkiye (R^2 = .04).

For the reasoning domain, all variables without SLM variable significantly predicted the mean RT (standardized β weight ranges from -.12 to .06) for the Singapore sample. For Türkiye, only two of the five variables (SEC and HRL) had a non-significant effect on predicting mean RT (standardized β weight ranges from -.09 to .10). But these variables explained only 2% and 3% of the variance of mean RT for Singapore (R^2 = .02) and Türkiye (R^2 = .03), respectively.

3.3. Findings for the Third Research Question

The findings of the RTE of the students in Singapore and Türkiye 4th grade samples in the mathematics achievement test are presented in Table 6.

Table 6. *Percentage of response time effort categories by content and cognitive domains.*

Domain		Singapore			Türkiye		
		Low	Medium	High	Low	Medium	High
Content	Numbers	.15	.65	99.20	.35	1.02	98.63
	Measurement and Geometry	.15	1.37	98.48	.30	2.04	97.66
	Data	.64	.38	98.96	2.14	.72	97.07
Cognitive	Knowing	.20	1.50	98.40	.30	2.80	97.00
	Applying	.10	2.00	97.90	.50	3.60	96.00
	Reasoning	.30	1.34	98.36	.60	2.48	96.92

Note. High effort (above .90), medium effort (between .90–.80) and low effort (below .80)

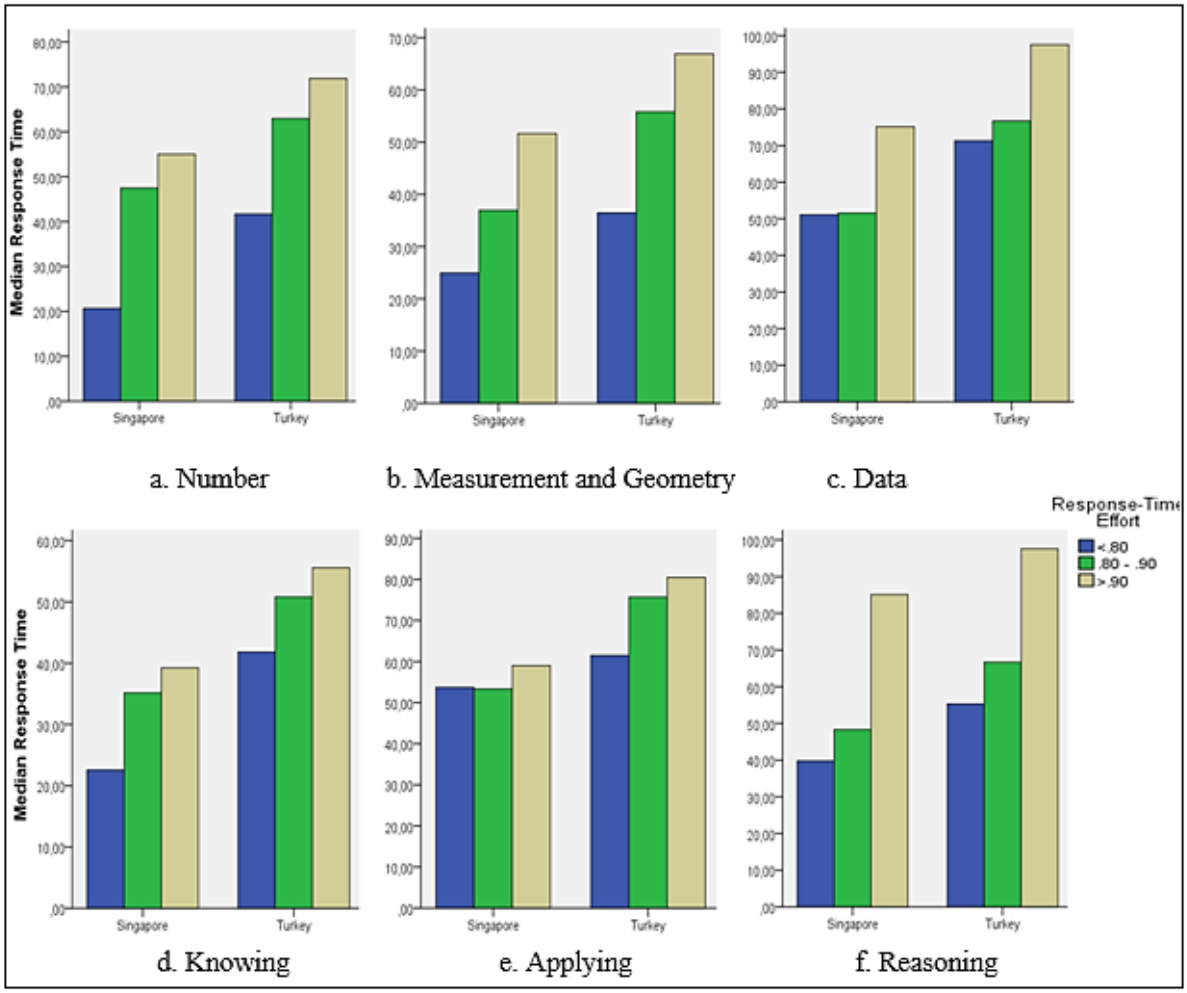
Table 6 presents the percentage of RTE categories by content and cognitive domains. The RTE index for a student was determined by calculating the average of the SB index across the items attempted by the student for each domain. More than 97% of Singaporean students and more than 96% of Turkish students had an RTE value between .9–1.0, which means those students were consistently classified as exhibiting SB across all domains throughout the test. There was a maximum .65% of Singaporean students and .60% of Turkish students (except for the data content domain, it was 2.14%) with RTE smaller than .8, which means those students are more likely to be categorized as displaying rapid guessing behavior.

Only among students with low RTE, when we examined the distribution of categories of HRL, SCM, SLM and SEC, nearly more than half of the students were in the lower category, such as low self-efficacy. The distribution by gender was balanced, that is, the tendency of male and female students to exhibit rapid-guessing behavior was similar.

The mean RT according to RTE categories for each content and cognitive domain is presented in Figure 1. As seen, the graphs for both content and cognitive domains support the assumption that students with a high RTE score consume more energy to get a good score on the test. The reason why we use the median instead of the mean in the graphs is to show the relationship between test-taking performance and RT more clearly. Because the SB index is calculated based

on the threshold, if we used the mean for graphs, it might not show the expected results due to extreme values. For example, if the threshold is 7 seconds, the student who answers an item in 1500 seconds is coded as 1, and another student who answers the same item in 8 seconds is also coded as 1. Here, the mean as the center of gravity is not valid for the distribution of RT and accordingly, the median is appropriate.

Figure 1. *Median response time according to response-time effort by content and cognitive domains.*



4. DISCUSSION and CONCLUSION

4.1. Relationship Between Content Domain, Cognitive Domain, and Item Difficulty with Mean Response Time

The results of this study showed that students' difficulty with the items was positively related to the cognitive level of the item in general. In other words, items from the knowing domain to the reasoning domain became increasingly difficult and the correct answer rate of the students decreased. Certainly, this is a predictable result as the "knowing" aspect encompasses the factual knowledge, concepts, and procedures that students are expected to be acquainted with. However, "reasoning" represents a higher-level cognitive domain that extends beyond solving routine problems, incorporating unfamiliar scenarios, intricate contexts, and multi-step challenges (Mullis & Martin, 2017, p.22). This result supports the postulate of a cumulative hierarchy of the cognitive domains. In this study, this fact was more prominently observed for Turkish students and agrees with the results of many studies on different subject areas (math or science etc.) or data sets (TIMMS or PISA etc.) as well. This result is similar to other studies

showing that the difficulty level of an item will increases according to the cognitive level (Ardıç & Soysal, 2018; İlhan et al., 2020; Koçdar et al., 2016; Nevid & McClelland 2013, Veeravagu et al., 2010). Additionally, Nehm and Schonfeld (2008), Momsen et al. (2013), İlhan et al. (2020), Ardıç and Soysal (2018) stated that item difficulty is not only affected by cognitive level but also by factors such as item type, content, and subject area of the item. In that regard, this study was similarly demonstrated that students, especially Türkiye sample, experienced more difficulties in measurement and geometry, data and numbers in content domains, respectively.

Another finding from this study showed that the content and cognitive domain of the items was positively related to the mean RT. In another word, when the cognitive level of the items increased, both Turkish and Singaporean students spent more time on the solution. Similarly, Yalçın (2022) determined that the cognitive level of the items caused a significant difference on the RTs of the students. Additionally, there were statistically significant differences between the content domains of the item in the mean RTs in both samples. Both Turkish and Singaporean students took longer to respond to data domain items compared to number and measurement and geometry domain items. Lee and Jia (2014) also examined RTs using the 8th grade mathematics items of the National Assessment of Educational Progress (NAEP). Although they found that none of the content areas (algebra, data, geometry, measurement, and number) caused particularly low RTs, the highest median RT was obtained from the number content area. This is different from the relationship between content domain and RT in our study. This difference may be due to the fact that the questions of the exams analysed in the studies were prepared at different cognitive levels. In addition, in the present study, it was determined that as the difficulty of the item increased, the student spent more time on the item. This finding is in parallel with Yang et al. (2002) who found a significant positive relationship between item difficulty and response time. The increase in students' effort while solving difficult items was also observed in the study conducted by Chae et al. (2018).

4.2. The Influence of Home Resources for Learning, Students Confident in Mathematics, Self-Efficacy for Computer Use, Students Like Learning Mathematics and Gender on Mean Response Time

HRL was negatively associated with the mean RT of 4th graders across almost all domains for both countries. This means that when a student has a higher level HRL, he or she will respond in less time to the items, and vice versa. Merely, mean RT in the reasoning domain was positively affected by the HRL score. Reasoning encompasses the application of knowledge and skills to unfamiliar contexts, encompassing the ability to draw logical inferences based on specific assumptions and rules, as well as providing justifications for the obtained results (Mullis et al., 2016, p.24). Therefore, it is an expected finding that students would perform the high-level skills required by items in reasoning in a longer time than items in lower cognitive domains. In this study, the variable with the highest impact on the mean RT was HRL. We think that we contributed to the literature by probably being the first study to examine any relationship between these two variables.

SCM was negatively correlated with the mean RT for items in both the knowing and applying cognitive domains and number content domain but was positively correlated with items in the reasoning cognitive domain and measurement and geometry content domain. This difference may be due to the difference in the difficulty levels of the items according to the content domain. Similarly, Yalçın (2022) found that students who were somewhat confident in mathematics spent less time answering difficult mathematics items than students who were very confident. However, Lasry et al. (2013) stated that students with low self-confidence spent more time to answer the items. In the study by Hoffman and Spatariu (2008), negatively correlation between self-efficacy and RT was found only for easier problems. According to the researchers,

undergraduate students with higher levels of self-efficacy may opt for automatic strategies instead, potentially allocating their time-consuming resources towards problem-solving tasks. Hoffman (2010) also observed similar relationships in his paper with pre-service teachers and ungraduated students. In this study, items in the reasoning domain and items in measurement and geometry domain are also relatively more difficult than ones in other content and cognitive domains. In this context (the two terms are not the same thing, but self-confidence and self-efficacy are so related), our findings are consistent with the paper of Hoffman and Spatariu (2008) and Hoffman (2010).

SLM was positively correlated to mean RT for both samples under most conditions. This variable had no significant effect for items in reasoning from the cognitive domain and in data from the content domain for Singapore sample. But for Türkiye sample, the higher the cognitive domain in which an item was, SLM was significantly more effective on mean RT. This is an unexpected finding because of the positive relationship between self-confidence and like learning mathematics. The students with the confidence, as some researchers reported, will be more motivated and more like learn mathematics (Hannula, 2004; Levine & Donitsa-Schmidt, 1998; Rabbani & Herman, 2017). Additionally, we positively found a correlation with approximately .63 between these variables for both samples. In this study, students who were confident in mathematics and SLM had affected the mean RT in the different way. We think that a self-report bias could affect the emergence of this dilemma. As the American Psychological Association (2022) defines, self-report bias occurs when individuals offer self-assessed measures of some phenomena and individuals may not give answers that are fully correct even if the survey is anonymous. There are many reasons for self-report bias, ranging from a misunderstanding of what a proper measurement is to social-desirability, where the respondent seeks to make a good impression in the survey or not knowing the full answer.

In the Singapore sample, the higher the students had scores on scales of SEC, the sooner they spent time in response. However, this variable did not statistically have any influence on the mean RT in the Türkiye sample. Actually, we found this finding somewhat surprising. Because the ratio of Turkish and Singaporean students who have their own computers is approximately 74% and 95%, respectively, although the scale means of the two countries are quite close. Cultural factors and individual backgrounds could play a role in this phenomenon.

Another finding, there was an influence of gender on the mean RT for both samples. Girls devoted a longer time to response than boys for all domains in the Singapore sample, but for only the number content domain and all cognitive domains in the Türkiye sample. Hunt et al. (2017) analyzed RT data using a 2 (year group, 5 and 6 graders) × 2 (gender) × 2 (problem type, two and three digits) mixed ANOVA. It can be acknowledged that the problems in their study are classified in the number content domain. Unlike our finding, they found that there was no significant main effect of gender and interactions between the independent variables.

4.3. Response Time Effort and Behavior of Students Across the Test

Wise (2017, p.55) interpreted rapid guessing as the following: "Generally, in high-stakes tests, rapid guesses represent strategic attempts to maximize one’s score, whereas in low-stakes tests they represent unmotivated test taking." However, a test taker with SB may have to display rapid guessing if the test with a time limit was about to expire. Irrespective of the reasons and the testing environment, when rapid guessing behavior is observed, it indicates that the test taker is either not engaged or minimally engaged with the test item in terms of effort. In terms of mean RT and RTE index smaller than .8, rapid-guessing behavior and, accordingly, low motivation was observed in below 1% of both samples (except for the data content domain in the Türkiye sample, it was 2.14%). Although it is a low-stakes assessment, almost all the students in the TIMSS 2019 math test showed high SB and, accordingly, high motivation. The variations in students' effort levels on low-stakes tests across different countries could be

attributed to cultural disparities in the significance placed on academic achievement. For instance, Gneezy et al. (2019) found a positive correlation between increased stakes associated with tests and performance in the United States, whereas this correlation was not observed in Shanghai. Borgonovi et al. (2021) highlighted those Asian countries like Singapore place great emphasis on international assessments, considering them as indicators of government policy effectiveness and a source of national pride. This political factor could positively affect the attitudes of Singaporean students toward international tests, and their motivation to do their best. In Türkiye, on the other hand, some studies have been conducted at the provincial and school level to ensure student motivation and to be ready for applications (Ministry of National Education, 2019).

In terms of students classified as low effort in the present study, we observed that Turkish and Singaporean students were likely to have rapid guessing behavior when an item in the reasoning domain became increasingly difficult (probably increasingly complex, also). Similar facts occurred in the data content domain, especially for Turkish graders. Although girls devoted a longer time to the response item than boys, almost no difference was observed in terms of students with low or high RTE index by gender and domains. Only for Singaporean graders, girls had a little higher test-taking effort and test motivation than boys. Unlike this finding, Zhao (2020) reported that girls were less likely to show disengaged behavior than boys in PISA 2012 assessments of computer-based mathematics. Additionally, positive but weak relationships (correlations with maximum .10) were observed between RTE and HRL, SCM, SEC and SLM, which means that graders with higher scores on these scales, he or she would have more items with solution behavior and higher test motivation. Similarly, Zhao (2020) reported a negative correlation between number-disengaged items (refers to rapidly selecting a response to multiple-choice items or omitting items) mathematics interest, and math self-efficiency.

4.4. Limitations and Suggestions for Future Research

In our study, the TIMSS 2019 User Guide (Fishbein et al., 2021) was used for the classification of the cognitive domains in which the items were included. The difficulty of placing the items in a particular cognitive domain precisely can be considered a limitation. For example, some items are likely to belong to more than one cognitive domain, or some experts may disagree on the cognitive categorization of some items. The other limitation of the study, the NT10 methodology was used to display which type of behavior students displayed. But various methods for setting the threshold have been suggested, including the use of mixture modeling (Schnipke & Scrams, 1997), visually inspecting the response time distribution (DeMars, 2007; Wise, 2006), using item characteristics (Wise & Kong, 2005), setting a common threshold across all items (Wise et al., 2004), cumulative proportion method (Guo et al., 2016), mixture log-normal (Rios & Guo, 2020). Therefore, a similar study topic using different threshold methodology can be suggested for future research. Like the study of Walkington et al. (2019), the influence of language features of mathematic problems, such as the number of sentences, pronouns, or problem topics, on student response time could be examined because systematically varying readability may demand affect student performance by different researchers. Since the TIMSS 2019 questions could not be fully accessed, this research was insufficient to examine how the characteristics of the items that need to be examined in person will affect the response time. As another research topic, the effect of students' familiarity and confidence in using a computer or tablet can be examined on test-taking efforts in computer-based test assessment under various conditions, as well.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors.

Authorship Contribution Statement

Esin Yilmaz Kogar: Problem Statement, Investigation, Methodology, Visualization, Formal Analysis, and Writing-original Draft. **Sumeyra Soysal**: Investigation, Methodology, Visualization, Formal Analysis, and Writing-original Draft.

Orcid

Esin Yimaz Kogar https://orcid.org/0000-0001-6755-9018
Sumeyra Soysal https://orcid.org/0000-0002-7304-1722

REFERENCES

American Psychological Association. (2022). Self-report bias. In APA dictionary of psychology. https://dictionary.apa.org/self-report-bias

Barry, C.L, & Finney, S.J. (2009). *Exploring change in test-taking motivation.* Northeastern Educational Research Association

Barry, C.L., Horst, S.J., Finney, S.J., Brown, A.R., & Kopp, J.P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10,* 342–363. https://doi.org/10.1080/15305058.2010.508569

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: the effects of incentives on motivation and performance. *European Journal of Psychology of Education, 14*, 441–462. http://www.jstor.org/stable/23420343

Bennett, R.E., Brasell, J., Oranje, A., Sandene, B., Kaplan, K., & Yan, F. (2008). Does it matter if I take my mathematics test on a computer? A second empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment, 6*(9), 1-39. https://files.eric.ed.gov/fulltext/EJ838621.pdf

Bergstrom, B.A., Gershon, R.C., & Lunz, M.E. (1994, April 4-8). *Computer adaptive testing: Exploring examinee response time using hierarchical linear modeling.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA. https://files.eric.ed.gov/fulltext/ED400287.pdf

Borgonovi, F., Ferrara, A., & Piacentini, M. (2021). Performance decline in a low-stakes test at age 15 and educational attainment at age 25: Cross-country longitudinal evidence. *Journal of Adolescence, 92*, 114-125. https://doi.org/10.1016/j.adolescence.2021.08.011

Bridgeman, B., & Cline, F. (2000). *Variations in mean response time for questions on the computer-adaptive GRE General Test: Implications for fair assessment.* GRE Board Professional Report No. 96-20P. Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2000.tb01830.x

Chae, Y.M., Park, S.G., & Park, I. (2019). The relationship between classical item characteristics and item response time on computer-based testing. *Korean Journal of Medical Education, 31*(1), 1-9. https://doi.org/10.3946/kjme.2019.113

Chen, G., Cheng, W., Chang, T.W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education, 1*(3), 213-225. http://dx.doi.org/10.1007%2Fs40692-014-0012-z

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cole, J.S., Bergin, D.A., & Whittaker, T.A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*(4), 609–624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Cooper, J. (2006). The digital divide: The special case of gender. *Journal of Computer Assisted Learning, 22*, 320–334. https://doi.org/10.1111/j.1365-2729.2006.00185.x

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Wadsworth.

Çokluk, Ö., Gül, E., & Doğan-Gül, C. (2016). Examining differential item functions of different item ordered test forms according to item difficulty levels. *Educational Sciences-Theory & Practice, 16(1)*, 319-330. https://doi.org/10.12738/estp.2016.1.0329

DeMars, C.E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment, 12*(1), 23–45. https://doi.org/10.1080/10627190709336946

Eklöf, H. (2007). Test-taking motivation and mathematics performance in the TIMSS 2003. *International Journal of Testing, 7*(3), 311-326. https://doi.org/10.1080/15305050701438074

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education Principles Policy Practice, 17*, 345-356. https://doi.org/10.1080/0969594X.2010.516569

Fan, Z., Wang, C., Chang, H.-H., & Douglas, J. (2012). Response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics, 37*(5), 655-670. http://dx.doi.org/10.3102/1076998611422912

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 user guide for the international database* (2nd ed.). TIMSS & PIRLS International Study Center.

Gneezy, U., List, J.A., Livingston, J.A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: the role of effort on the test itself. *American Economic Review: Insights, 1*(3), 291-308. http://dx.doi.org/10.1257/aeri.20180633

Guo, H., Rios, J.A., Haberman, S., Liu, O.L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*(3), 173-183. https://doi.org/10.1080/08957347.2016.1171766

Hannula. (2004). Development of understanding and self-confidence in mathematics, grades 5-8. *Proceding of the 28th Conference of the International Group for the Psychology of Mathematics Education, 3*, 17-24. http://files.eric.ed.gov/fulltext/ED489565.pdf

Hess, B.J., Johnston, M.M., & Lipner, R.S. (2013). The impact of item format and examinee characteristics on response times. *International Journal of Testing, 13*(4), 295–313. https://doi.org/10.1080/15305058.2012.760098

Hoffman, B. (2010). "I think I can, but I'm afraid to try": The role of self-efficacy beliefs and mathematics anxiety in mathematics problem-solving efficiency. *Learning and Individual Differences, 20*(3), 276-283. https://doi.org/10.1016/j.lindif.2010.02.001

Hoffman, B., & Spatariu, A. (2008). The influence of self-efficacy and metacognitive prompting on math problem-solving efficiency. *Contemporary Educational Psychology, 33*(4), 875-893. https://doi.org/10.1016/j.cedpsych.2007.07.002

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

Koçdar, S., Karadağ, N., & Şahin, M.D. (2016). Analysis of the difficulty and discrimination indices of multiple-choice questions according to cognitive levels in an open and distance

learning context. *The Turkish Online Journal of Educational Technology, 15*(4), 16–24. https://hdl.handle.net/11421/11442

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., & 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. https://doi.org/10.1186/s40536-014-0008-1

Levine, T., & Donitsa-Schmidt, S. (1998). Computer use, confidence, attitudes, and knowledge: A causal analysis. *Computers in Human Behavior, 14*(1), 125-146. http://dx.doi.org/10.1016/0747-5632(93)90033-O

Lundgren, E., & Eklöf, H. (2020). Within-item response processes as indicators of test-taking effort and motivation. *Educational Research and Evaluation, 26*(5-6), 275-301. https://doi.org/10.1080/13803611.2021.1963940

Martin, M.O., von Davier, M., & Mullis, I.V.S. (Eds.). (2020). *Methods and procedures: The TIMSS 2019 technical report.* The TIMSS & PIRLS International Study Center. https://www.iea.nl/publications/technical-reports/methods-and-procedures-timss-2019-technical-report

Michaelides, M.P., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing, 20*(3), 187-205. https://doi.org/10.1080/15305058.2019.1706529

Ministry of National Education (2019, March 19). *Muğla İl Millî Eğitim Müdürlüğü: The TIMSS 2019* [Muğla Provincial Directorate of National Education: TIMMS 2019] https://mugla.meb.gov.tr/www/timss-2019/icerik/2298

Momsen, J., Offerdahl, E., Kryjevskaia, M., Montplaisir, L., Anderson, E., & Grosz, N. (2013). Using assessments to investigate and compare the nature of learning in undergraduate science courses. *CBE-Life Sciences Education, 12*(2), 239-249. https://doi.org/10.1187%2Fcbe.12-08-0130

Mullis, I.V.S., Martin, M.O., Goh, S., & Cotter, K. (Eds.). (2016). *The TIMSS 2015 encyclopedia: Education policy and curriculum in mathematics and science.* The TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/timss2015/encyclopedia/

Mullis, I.V.S., & Martin, M.O. (2017). *The TIMSS 2019 assessment frameworks*. The TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/timss2019/frameworks/

Myers, A.J., & Finney, S.J. (2021). Change in self-reported motivation before to after test completion: Relation with performance. *The Journal of Experimental Education, 89*, 74–94. https://doi.org/10.1080/00220973.2019.1680942

Nehm, R.H., & Schonfeld, M. (2008). Item feature effects in evolution assessment. *Journal of Research in Science Teaching, 48*(3), 237–256. https://doi.org/10.1002/tea.20400

Nevid, J.S., & McClelland, N. (2013). Using action verbs as learning outcomes: Applying Bloom's taxonomy in measuring instructional objectives in introductory psychology. *Journal of Education and Training Studies, 1*(2), 19-24. http://dx.doi.org/10.11114/jets.v1i2.94

Organisation for Economic Co-operation and Development [OECD]. (2015). *Using log-file data to understand what drives performance in PISA (case study), in students, computers and learning: Making the connection*. OECD Publishing. https://doi.org/10.1787/9789264239555-en

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passaged-based tests. *Journal of Technology, Learning, and Assessment, 2*(6), 1–45. https://files.eric.ed.gov/fulltext/EJ905028.pdf

Rabbani, S., & Herman, T. (2017). Increasing Formulate and Test Conjecture Math Competence and Self Confidence in Using the Discovery Learning Teaching Math. *PrimaryEdu: Journal of Primary Education, 1*(1), 119-128. http://dx.doi.org/10.22460/pej.v1i1.488

Rios, J.A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of diferential nonefortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education, 33*(4), 263-279. http://dx.doi.org/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. https://psycnet.apa.org/doi/10.1111/j.1745-3984.1997.tb00516.x

Setzer, J.C., Wise, S.L., van de Heuvel, J.R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34–49. https://doi.org/10.1080/08957347.2013.739453

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review, 31*, 100335. https://doi.org/10.1016/j.edurev.2020.100335

Sundre, D.L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and nonconsequential test performance?. *Contemporary Educational Psychology, 29*(1), 6-26. https://psycnet.apa.org/doi/10.1016/S0361-476X(02)00063-2

Swerdzewski, P.J., Harmes, J.C., & Finney, S.J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*(2), 162–188. http://dx.doi.org/10.1080/08957347.2011.555217

Veeravagu, J., Muthusamy, C., Marimuthu, R., & Subrayan, A. (2010). Using Bloom's taxonomy to gauge students' reading comprehension performance. *Canadian Social Science, 6*(3), 205–212. https://doi.org/10.3968/J.CSS.1923669720100603.023

Walkington, C., Clinton, V., & Sparks, A. (2019). The effect of language modification of mathematics story problems on problem-solving in online homework. *Instructional Science, 47*(5), 499-529. https://link.springer.com/article/10.1007/s11251-019-09481-6

Wang, M. (2017). *Characteristics of item response time for standardized achievement assessments* [Doctoral dissertation]. University of Iowa.

Wang, T., & Hanson, B.A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement, 29*(5), 323-339. https://doi.org/10.1177/0146621605275984

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement, 41*(2), 115-129. https://doi.org/10.1177%2F0146621616676791

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. https://doi.org/10.1207/s15324818ame1902_2

Wise, S.L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52-61. https://doi.org/10.1111/emip.12165

Wise, S.L., & DeMars, C.E. (2005). Low examinee effort in low-stakes assessment: problems and potential solutions. *Educational Assessment, 10*(1), 1-17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S.L., Kingsbury, G.G., Thomason, J., & Kong, X. (2004, April 13-15). *An investigation of motivation filtering in a statewide achievement testing program.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

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. https://doi.org/10.1207/s15324818ame1802_2

Wise, S.L., & Ma, L. (2012, April 13-17). *Setting response time thresholds for a CAT item pool: The normative threshold method*. In annual meeting of the National Council on Measurement in Education, Vancouver, Canada (pp. 163-183). https://www.nwea.org/resources/setting-response-time-thresholds-cat-item-pool-normative-threshold-method/

Wolgast, A., Schmidt, N., & Ranger, J. (2020). Test-taking motivation in education students: Task battery order affected within-test-taker effort and importance. *Frontiers in Psychology, 11*, 1–16. https://doi.org/10.3389/fpsyg.2020.559683

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

Yang, C.L., O Neill, T.R., & Kramer, G.A. (2002). Examining item difficulty and response time on perceptual ability test items. *Journal of Applied Measurement, 3*(3), 282-299.

Zenisky, A.L., & Baldwin, P. (2006). *Using item response time data in test development and validation: Research with beginning computer users*. Center for educational assessment report No, 593. Amherst, MA: University of Massachusetts, School of Education.

Zhao, W. (2020). *Identification and validation of disengagement measures based on response time: An application to PISA 2012 digital math items* [Master's thesis]. University of Oslo.

Zhang, T., Xie, Q., Park, B.J., Kim, Y.Y., Broer, M., & Bohrnstedt, G. (2016). *Computer familiarity and its relationship to performance in three NAEP digital-based assessments* (AIR-NAEP Working Paper No. 01-2016). American Institutes for Research.