Large-scale Assessments in Education

RESEARCH Open Access

Examining successful and unsuccessful time management through process data: two novel indicators of test-taking behaviors

Elena C. Papanastasiou[1]* and Michalis P. Michaelides[2]

*Correspondence:
Papanastasiou.e@unic.ac.cy

[1] Department of Education, School of Education, University of Nicosia, 46 Makedonitissas Avenue, 2417 Nicosia, Cyprus
[2] Department of Psychology, School of Social Sciences and Education, University of Cyprus, Nicosia, Cyprus

Abstract

Test-taking behavior is a potential source of construct irrelevant variance for test scores in international large-scale assessments where test-taking effort, motivation, and behaviors in general tend to be confounded with test scores. In an attempt to disentangle this relationship and gain further insight into examinees' test-taking processes, researchers can now utilize process and timing data to obtain a more comprehensive view of test-taking behaviors, such as test-taking effort. The purpose of this study is to propose and evaluate two novel response-based, standardized indicators of test-taking behaviors that utilize a combination of examinee response and process (timing) data to better understand and describe test-taking effort in ILSAs. These indices were empirically estimated with USA data from two booklets from e-TIMSS 2019 in mathematics for grade 4. In addition, their predictive validity was examined with respect to achievement estimates. Their network of associations with other relevant variables such as motivation, interest in the subject, as well as across subjects were also examined to test their intra-individual stability in e-TIMSS.

Keywords: Response time, Effort, Process data, Computerized assessments, International large-scale assessments (ILSAs)

Introduction

When examinees participate in assessments of their knowledge and skill and obtain scores demonstrating their proficiency in the assessed domain, it is assumed that they have invested effort to perform well; otherwise, scores will not reflect their true level of ability and will not be valid indicators of their proficiency (Baumert & Demmrich, 2001; Wise, 2015). Metanalyses have indeed shown that effort during test-taking is positively correlated with test scores (Silm et al., 2020; Wise & DeMars, 2005). Lack of interest, motivation and effort while taking a test, as well as variations in test-taking strategies that examinees use pose a threat to the validity of test outcomes (Papanastasiou, 2015, 2020; Papanastasiou & Stylianou-Georgiou, 2022). Especially, low-stakes assessment programs like the Trends in International Mathematics and Science Study (TIMSS), where no consequences are posed to test-takers for low performance, are less likely to elicit high test-taking effort (Rutkowski & Wild, 2015).

According to Lundgren and Eklof (2020), test-taking motivation is a specific motivation to maximize performance on a test. To achieve this goal when taking a test, examinees will have to expend effort and regulate the necessary skills, knowledge, time and resources. Empirical studies on test-taking effort have originally approached test-taking effort via self-reports. However, behavioral indicators, primarily automatically-recorded item response times from computerized tests, have been shown to be less prone to response biases, less intrusive, and available at the item-level (Eklöf, 2010). Time spent on reading, processing and giving an answer to an item is considered as a reliable behavioral measure of engagement with the test. Much of this research has been initiated by Schnipke and Scrams (1997), followed by Steven Wise and colleagues who developed the use of (a) rapid guessing, i.e. providing a response in a very short time interval, as a manifestation of disengaged behavior when responding to a test item; and (b) response time effort as an aggregate indicator of effortful behavior on the whole test (Wise, 2017; Wise & Kong, 2005).

Recently, researchers have shown interest in student effort and engagement with international large-scale assessments. Using methods to identify rapid guessers, Michaelides et al. (2020), and Pools and Monseur (2021) have shown that response time effort is correlated with performance in PISA and weakly correlated with achievement motivation and enjoyment variables. Guo and Ercikan (2020), Michaelides and Ivanova (2022), and Rios and Soland (2022) have also looked at cross-country differences that exist in rapid guessing.

Implementation of response time measures to identify rapid guessing behavior (as a dichotomous variable) at the item level requires the selection of a threshold. Examinees who provide a response at a time below the stated threshold are characterized as rapid guessers, not engaging in solution behavior (Wise & DeMars, 2006). Proposed ways to determine a threshold for rapid guessing include a fixed time point, common for all items (Wise et al. 2010), or judgmental decisions based on item length or the inspection of the response time frequency distribution (Wise & Kong, 2005; Setzer et al., 2013). Other approaches incorporate performance on the item depending on response time (Guo et al., 2016), modeling with mixture models or IRT (e.g. Ulitzsch et al., 2020), and normative methods based on a proportion of the average time expended on an item (Wise & Ma, 2012). Comparisons about an optimal threshold identification approach have been inconclusive (Wise, 2019), and there is no consensus on a preferred method as there are strengths and weaknesses for each one (cf., Rios & Deng, 2021; Soland et al., 2021). Simpler methods such as the 5-s rule for all items are easy to implement and provide thresholds for all items but are criticized for higher misclassification errors. For example, a proficient examinee may respond rapidly but thoughtfully to an easy item and could be misclassified as a disengaged rapid guesser if he or she provided a response in less than 5 s; or a disengaged examinee who may glance over an item before moving slowly to the next item could be identified as a non-rapid guesser. Unavoidably, trying to reduce the possibility of false-positive results by changing the threshold, increases the possibility of false-negatives (Wise, 2017). Moreover, studies have predominantly looked at multiple-choice items, although some initial proposals have been recently put forth for omitted and constructed responses (Ivanova et al., 2020; Wise & Gao, 2017). More sophisticated methods that take performance into account appear more valid, but rest on distributional assumptions and often do not converge or do not provide thresholds for all items (cf., Soland et al., 2021; Ulitzsch et al., 2020, 2022).

Further information about response events is available in digital assessments from log files. Examining timing data from log files alone, does not always provide adequate indication of an examinee's test-taking behavior. The time that a student might need to respond to a test item depends on various factors, including those of the examinee's overall ability, examinee test-taking behaviors or strategies, item characteristics (e.g. idem difficulty, item length, auxiliary visual material) as well as any interaction of these factors. Two students who spent very little time on a test item, might have done so for numerous reasons. One student might have spent very little time because the question was very easy for them, while another student might have done so because they did not want to spend any effort on a question that was too difficult for them. Consequently, timing data can be more informative when examined in relation to other variables.

The purpose of the current study is to present and evaluate two novel indicators of examinee test-taking behaviors, that utilize a combination of examinee response and timing data, to better understand and describe test-taking effort. To calculate the proposed indicators, the first step includes the calculation of the $MedianT_i$, which corresponds to the median amount of response time for answering each of the multiple-choice items i, $i = 1,...,K$ that were administered in a test booklet. At a second stage, a deviation score is calculated for each student j who was administered item i by subtracting the median amount of response time for item i from the students' response time T_{ij} for the same item. Based on these deviation scores, a cumulative indicator is calculated for each student for each of the new indicators as follows:

1) For items i that were answered incorrectly by person j in less time than the median response time, the absolute value of this time difference was added to the *Unsuccessful Time Management indicator* (UTM) for the examinee as follows:
If item i was answered incorrectly by person j, then

$$UTM_j = \sum_{i=1}^{K} |\min\left(0, T_{ij} - MedianT_i\right)| \tag{1}$$

Since such items have been answered incorrectly, it is likely that the students made less than adequate effort to answer them correctly since they provided a response in less time than the median.

2) For items i that were answered correctly by person j in less time than the median response time, the absolute value of this time difference was added to the *Successful Time Management indicator (STM)* for the examinee.
If item i was answered correctly by person j, then

$$STM_j = \sum_{i=1}^{K} |\min\left(0, T_{ij} - MedianT_i\right)| \tag{2}$$

Since such items have been answered correctly, it is likely that the students were either already proficient on the specific content and thus did not need additional time to respond to them, or were just lucky in a rapid guess.

Based on these indicators, the research questions of this study, that examine the indicators of "Successful Time Management" (STM) and "Unsuccessful Time Management" (UTM), are the following:

1. What are the distributions of the STM and the UTM indicators in test booklets from the USA sample for TIMSS grade 4 mathematics and science?
2. How do these indicators differ among students in different benchmark levels and in different responder classification categories?
3. To examine their validity, to what extent do these indicators correlate with test performance, and motivational characteristics? Is there intra-individual stability by comparing the indicators across the Math and Science e-TIMSS assessments?
4. To what extent do students with STM and students with UTM behavior exhibit extreme rapid guessing (response in less than three seconds) at the item level?

Such indicators can be used for various purposes. For example, they could be used to obtain a more detailed picture of the students' test-taking behaviors as well as describe the effort they put on the test, conditioning on the accuracy of their responses. In addition, by studying their association with other correlates of effort, it may be possible to identify test design features that can be improved. These indicators will also enrich the field of measurement by moving beyond the examination of rapid responses identified in relation to thresholds that classify students in rapid guessing or not rapid guessing groups (Wise & DeMars, 2005; Wise & Kong, 2005). These scores, which are on a continuous scale of easily interpretable time units (seconds), represent the amount of fluency and efficiency of examinees in the case of STM, or the lack thereof in the case of UTM, while responding to the items in the course of a test session. Finally, they hold the potential to help strengthen the validity of low-stakes tests such as the ones administered by the IEA where student motivation is a potential concern (Baumert & Demmrich, 2001). On a more applied level, educators and policy makers could also utilize such results in the future, to examine factors that can improve student engagement during test-taking.

Methods

The population of the study included grade 4 students from the USA. The sample that was utilized for the analyses in the current study, included the students who were administered Booklets 7 and 8 in e-TIMSS 2019. Booklet 7 was randomly selected as a booklet which started with mathematics items, while Booklet 8 started with science items. This sample included 1250 students, of which 49.44% were female. The average age of the students was 10.26 years (SD = 0.42). The variables from TIMSS used for the current study were obtained from the grade 4 student achievement data files, as well as the student context data files. The information obtained from the student achievement data files were the examinee item responses on multiple-choice items graded as correct or incorrect, the five plausible values (PV) in mathematics, the timing of students on each mathematics multiple-choice item that was administered to them, the examinee benchmark levels, along with a special process variable from the e-TIMSS dataset, titled mathematics (or science) responder classification. The grade 4 responder classification variable categorized students based on the patterns of not-reached items (Fishbein et al., 2021) into one of three distinct categories; so responders are classified based on whether they have reached all items on the test, whether they have run out of time before completing the test, and whether they stopped responding while they still had time to complete the test. From the student context data files, two motivational scales were obtained: Students liking mathematics scale, Student confidence in mathematics scale, and the corresponding scales for the science test. The student confidence in variable was created based on nine items measuring confidence in mathematics and science, separately for each subject. The students like learning scale was created based another nine questionnaire items corresponding to each subject from the student background questionnaire.

The data for the study included five benchmark levels per student in mathematics to correspond to the five PVs in the subject, as well as five benchmarks for each student for science. To be able to present the results of the STM and UTM indicators by benchmark levels, a decision was made to identify the median benchmark for each student, for each subject. Therefore, the median benchmark was specifically created for each student to avoid presenting results separately for each PV.

Finally, for each examinee, we characterized an item response as extreme rapid guessing if it was provided within 3 s of the item appearing on the screen, under the assumption that TIMSS items cannot be answered even with partial effort by 4th-graders in such a brief time interval. Then, we counted the number of items on which the extreme rapid guessing behavior appeared—an indicator similar to but opposite than Response Time Effort (Wise & Kong, 2005).

The analyses were mostly performed with descriptive statistics and inferential statistics. All analyses incorporating plausible values (PV) were conducted using the International Database Analyzer (version 5.0.23) developed by the International Association for the Evaluation of Educational Achievement (IEA), and utilized student weights in the analyses. This specialized software tool facilitated accurate handling and interpretation of PVs, ensuring robustness in the findings. Additionally, the analyses were replicated across two key academic subjects: mathematics and science.

Results

The descriptive statistics of the STM and UTM indicators which have been created based on the USA multiple-choice e-TIMSS items for grade 4, are presented below. According to Table 1, the percentage of students who utilized less time than average on at least one of their items for mathematics was 89.52% for the STM indicator, and 79.60% for the UTM indicator. The corresponding percentages for science were equal to 89.52% and 82.80%. Although slight differences are observed between booklets 7 and 8, in general, there tends to be a higher percentage of students engaging in STM compared to UTM for both subjects. In terms of the magnitude of these differences, the medians of the STM tended to be higher than the median of the UTM indicator for both booklets. This resulted in an overall median of 27.96 s for STM in mathematics, 21.58 s in UTM for mathematics, and 30.39 s and 16.74 s respectively for science.

Table 1 Descriptive statistics of Successful and Unsuccessful Time Management indicators in seconds

	Mathematics		Science	
	STM	UTM	STM	UTM
Overall %	89.52	79.60	89.52	82.80
Median (sec)	27.96	21.58	30.39	16.74
25th Percentile (sec)	10.17	3.22	8.66	3.68
75th Percentile (sec)	55.89	50.55	58.91	37.82
Maximum (sec)	242.53	335.84	236.07	282.47
Booklet 7				
%	89.41	76.73	91.49	88.60
Median (sec)	21.75	15.21	37.10	23.19
25th Percentile (sec)	6.74	1.43	11.93	8.53
75th Percentile (sec)	43.21	40.32	68.62	46.35
Maximum (sec)	186.33	201.14	190.17	282.47
Booklet 8				
%	89.63	82.46	87.56	77.00
Median (sec)	35.23	28.95	23.43	10.41
25th Percentile (sec)	14.24	5.16	6.62	0.79
75th Percentile (sec)	66.50	65.65	51.88	28.24
Maximum (sec)	242.53	335.84	236.07	242.91

Table 2 Correlation coefficients between STM and UTM indicators across subjects

	STM math	UTM math	STM science	UTM science
UTM math	− 0.05			
STM science	0.42**	− 0.03		
UTM science	0.01	0.40**	0.04	
Time of last response	− 0.49**	− 0.43**	− 0.58**	− 0.30**

** p < 0.01

The correlation between the STM and UTM indicators for mathematics equals -0.05 (p = 0.088), and for science it equals − 0.03 (p = 0.281), suggesting no association between the two indicators (Table 2). However, the correlation between the mathematics and science STM variables equaled 0.42 (p < 0.001), and 0.40 (p < 0.001) for UTM. These results clearly indicate that there are similarities in the patterns of time use between academic subjects. The correlations between the STM and UTM variables and the time of last response were all negative, strong in size for STM, moderate for UTM, and statistically significant. The largest correlation was observed between the STM and the time of last response in science (r = − 0.58, p < 0.001), while the smallest was between UTM in science and time of last response in science (r = − 0.30, p < 0.001). This finding verifies the validity of the indicators since it would be expected that examinees who ended the test sooner would have higher levels of STM and UTM; it does not however discriminate between the two indicators.

A series of tests related to construct validity were conducted, by relating them to other variables to understand these indicators further. Such variables are those of median

Table 3 Descriptive statistics of the median of the STM and the UTM indicators by median benchmark

Benchmark level	Mean Plausible Value at benchmark level	% of examinees at benchmark level	Median STM	Median UTM
Mathematics				
1	367.11	5.20	17.60	76.46
2	442.28	20.64	15.36	55.05
3	513.54	34.64	24.23	26.05
4	582.87	28.32	38.87	10.72
5	654.30	11.20	65.24	0.00
Science				
1	360.93	5.52	9.70	70.68
2	443.13	17.68	9.77	31.62
3	516.59	32.96	19.78	20.23
4	585.33	31.28	39.81	11.12
5	654.01	12.56	69.39	7.61

Table 4 Descriptive statistics of the STM and the UTM indicators by mathematics responder classification

Mathematics Responder Classification	N	%	Mean PV	Successful time management	Unsuccessful time management
Mathematics					
Reached all items	1173	93.84	532.79	30.24	22.47
Ran out of time	29	2.32	480.11	.00	18.45
Stopped responding	45	3.60	456.32	6.66	10.01
Science					
Reached all items	1159	92.72	541.50	33.15	17.04
Ran out of time	39	3.12	464.50	0.01	15.78
Stopped responding	49	3.92	449.41	0.00	7.44

benchmark, mathematics responder classification, and examinee achievement, as represented by the five plausible values (PV). Table 3 presents the breakdown of the two indicators by median benchmark level. For higher benchmark levels, the successful time management indicator is higher, while the unsuccessful time management indicator is lower. Students in higher benchmark levels tend to have higher levels of STM and lower levels of UTM for both, mathematics and science.

The variable of Mathematic Responder Classification placed students into four categories, based on their overall timing behavior during the test. This variable included the categories of (a) Reached all items; (b) Ran out of time; (c) Stopped responding; and (d) Could not be classified. In the current sample, only three students were placed in the category “Could not be classified” and were therefore not included in the analyses. Based on this classification, the majority of the students who managed to reach all items on the test were also the ones with the largest median STM and UTM variables (Table 4). Most likely, this occurred due to their attempts to go through the test without many delays, to make sure that they would manage to reach the end of the test. These were also the students with the highest average achievement in terms of their Plausible Values (PVs).

Overall, however, the percentage of students who were classified in the other categories was very small which made it not possible to reach any robust conclusion regarding these cases of students, except that their STM indicator was at or near zero which implies a less than optimal test-taking strategy.

With the IEA Database Analyzer we examined the correlation between each of the two indicators with the plausible values (PV). The results for mathematics indicated that the correlation between the Successful Time Management indicator and the five PVs equaled r = 0.35 (se = 0.04), while the correlation between the Unsuccessful Time Management indicator and the PVs equaled r = -– 0.53 (se = 0.02). In science, the corresponding correlations equaled r = 0.41 (se = 0.03) with STM and r = -– 0.44 (se = 0.04) with UTM. Based on this result, it appears as though higher achieving students tend to be more frequently engaged with increased levels of STM, and with lower levels of UTM; they tended to have more unused time on their correct answers (thus, most likely being an indicator of mastery of the test content), and with less unused time for their incorrect answers (meaning that on items they did not do well, they were not responding hastily).

Table 5 presents the breakdown of the indicator-achievement relationship, broken down by benchmark. This was examined in order to examine whether the types of relationships between ability (as indicated by the PVs) and the two indicators differ among the different benchmark levels. As presented in Table 5, the correlations between the relevant variables were quite small. Most likely this has occurred due to the restriction of range of the achievement levels within each benchmark. The only correlation that was statistically significant at the 0.05 level was at benchmark level 3, between STM and the PV in science. Within this benchmark, the students who had higher levels of achievement, also had higher levels of Successful Time Management, by answering questions correctly in less time than average. In mathematics, none of the correlations were statistically significant at any benchmark level.

The STM and UTM indicators were also correlated with student motivational variables to further examine their validity. The selected variables were those of Students like learning each subject, and Students confident in the subject (Table 6). Of the two motivational variables, the correlation of Students like learning with STM was very small and statistically not significant for both subjects (0.06 and 0.02 with STM). The correlation

Table 5 Correlations between Achievement Levels by Benchmark and the Successful and Unsuccessful Time Management Indicators

Benchmark	Mathematics		Science	
	Successful time management	Unsuccessful time management	Successful time management	Unsuccessful time management
1	0.01	– 0.15	0.00	– 0.14
2	– 0.20	– 0.14	0.01	– 0.08
3	0.15	– 0.11	0.10*	– 0.09
4	0.15	– 0.19	0.14	– 0.13
5	0.08	– 0.15	0.13	– 0.10
Overall sample	0.07*	– 0.16*	0.10*	– 0.13*

* $p \leq 0.05$

Table 6 Correlations of STM and UTM with motivational variables

	STM	Students like learning subject SCL	Students confident in subject/SCL
Mathematics			
Unsuccessful time management	− 0.05	− 0.12	− 0.24
Students like learning /SCL	0.06	–	0.57
Student confidence/SCL	0.24	0.57	–
Science			
Unsuccessful time management	0.04	− 0.12	− 0.12
Students like learning /SCL	0.02	–	0.62
Student confidence/SCL	0.12	0.62	–

was larger between Student confidence and STM, which equaled 0.24, ($p < 0.001$) in mathematics and 0.12 ($p < 0.001$) in science. This result is not surprising, since self-confidence tends to be more strongly aligned with performance compared to enjoyment with the subject (e.g. Michaelides et al., 2019). The correlation between UTM and liking the two subjects was equal to − 0.12 ($p < 0.01$) for both subjects. However, the correlation between UTM and being confident in mathematics was equal to − 0.24 ($p < 0.001$) for mathematics and − 0.12 for science ($p < 0.01$).

The total number of examinees who responded in less than three seconds was estimated as a proxy to the lack of response time effort. In mathematics, there were 45 who responded to at least one item in less than 3 s, while in science there were 55. Among the small number of examinees who had at least one very rapid response, the correlation between the number of extreme rapid guesses with UTM in mathematics was 0.30 ($p = 0.05$). As would be expected, the more examinees engaged in extreme rapid guessing, the more likely they would accrue time because of rapid incorrect response behavior. The corresponding correlation in the science data was 0.14 and non statistically significant. The number of extreme rapid guesses was also non significantly correlated with STM in both subjects.

Discussion

By using classical test theory, a large proportion of assessment researchers, educators, and psychometricians have focused on correct, incorrect, and partially correct answers as indicators of examinee proficiency. Recent technical and methodological advancements in the area of computerized large-scale testing, however, have provided us with opportunities to better understand the testing process through the utilization of process data (Papanastasiou & Eklöf, 2020). Process data provide additional sources of information obtained by examinees during the test-taking process, and they hold the potential to revolutionize the field of testing. However, this field of study is relatively recent. Moreover, due to the fact that process data have only recently started to be collected by the IEA, few efforts have been made to combine process data with additional test-taking behaviors. Since no unified and easily understandable indicator exists that combines test-taking behaviors with process data, this study aimed to create two novel indicators of test-taking behaviors that combine accuracy and timing data in order to describe test-taking effort. These indicators that are easy to calculate and comprehend, can easily be generalized to any other study that is administered electronically, and for which process data are available.

An additional originality of these indicators is that their estimation is dependent on the time that students spent on each item, while controlling for the correctness of their response to that item. Consequently, by incorporating the accuracy of a response in the estimation procedure, the misclassifications that were likely to occur with other time-based rapid-guessing indicators, are avoided in the current approach. For example, a highly competent student who might have correctly answered a question very quickly, should not automatically be considered (misclassified) as a rapid guesser. Moreover, by interpreting the timing data based on whether a response was correct or incorrect, and by comparing the response time to the median, the possibility of having future examinees take advantage of such behaviors is eliminated. For example, it would be difficult for examinees to know in advance whether their response time was above or below the median or for them to know for sure whether their response was correct or incorrect in an attempt to demonstrate either high levels of STM or low levels of UTM accordingly. As a result, these indicators are less susceptible to manipulation by examinees, which is a great concern related to the use of process data (Bennett, 2018).

The medium size correlation that was observed between STM and UTM suggests that students who utilized less time than average on incorrect answers, also did so to some extent on correct answers as well. The fact that these occurrences mostly occurred with the students who managed to complete all test items in the allocated time, might be an indication of a test-taking strategy, to make sure that they had enough time to complete the test. However, the correlation between the two indicators was not large enough to universally claim that utilizing less time than average is purely based on a strong "speededness" trait. This is further supported by the fact that the behavior of responding in less time than average was related to other explanatory variables. For example, students who spend less time than average in correct answers tend to be in higher benchmark levels, indicating that this might have occurred since they had mastered the item content and did not need much time to respond to such items correctly. Also, students who spend less time than average in incorrect answers tend to be in the lower benchmark levels, which could be an indicator of making less effort on the test. This was further verified by the result that student confidence in mathematics was more highly correlated with STM rather than with UTM.

Overall, although narrow in range, STM was positively correlated with test performance, it tended to occur with students who were in the higher benchmark levels, and who also had more confidence in mathematics. This further verifies that this indicator could be considered as an indication of mastery of the content by the examinee. In contrast, UTM occurred more frequently and to a larger extent than STM. This indicator occurred more frequently with students in the lower benchmark levels, and it was not correlated with confidence in mathematics. Also, the fact that this indicator was larger for the students who stopped responding to the test, and was moderately correlated with the extreme rapid guessing frequency, further supports that for these students, this indicator is related to their lack of effort on the test.

Therefore, using less time than average on a test occurs for various reasons. Although to some extent using less time on test items might be an indication of a test-taking behavior that ensures that all items can be completed in the allotted time, this alone does not describe the full situation. One student might have used less time than average because they had clearly mastered the item content and did not need much time to answer it correctly, while another student might have used the same amount of time because they did not put much effort into the question, and eventually answered it incorrectly. As a result, by examining timing data in relation to whether an item was answered correctly or incorrectly in less time than average, can provide us with more detailed information regarding test-taker behavior. Such information can be used to describe examinees in IEA studies, beyond merely looking at their proficiency level. These novel indicators can also describe the *ways* in which each student took the test. For example, it will be possible to differentiate students within a country who mostly responded carelessly to many items, and omitted many other items, from students in another country with similar levels of proficiency, who utilized all of their available time, and viewed the difficult items many times in order to answer them. This is especially useful for international studies which are low-stakes, and in which student motivation and test-taking effort are a potential concern (Baumert & Demmrich, 2001). Educators and policy makers could also utilize such results to examine factors that can improve student engagement overall during test-taking.

These indicators might also differentiate the students who managed to obtain high scores with a high level of the STM indicator (since they managed to respond to most questions in a much lower time than average, without being careless rapid guessers), from other students of similar scores who were persistent, utilized all of their available time, and viewed the difficult items many times in order to answer them correctly. On the other side of the continuum, examinees with large UTM were students who answered many items rapidly and incorrectly, so this may be a way to identify those with a general disengagement with the test content.

Finally, these unified indicators can also be used to demonstrate the degree of validity of the IEA studies, since they can be used to describe examinee behaviors in more detail, without automatically assuming that all responses are thoughtful, or that all rapid responses are always rapid guesses, and indications of careless behaviors. This is further supported by the American Educational Research Association et. al (2014) which stated that test-taking efforts need to be taken into consideration as important validity factors when interpreting scores from low-stakes assessments.

Beyond the results presented in the current study, further research should be performed, to examine these indicators in more detail. For example, how do these variables perform in other subject areas in other studies? Would similar results be obtained from the data for grade 8 students or from students in other countries? What are the examinee characteristics, or country variables that could help explain the variations that exist in the magnitude of these indicators? Finally, additional research should also be performed to determine how these indicators can be calculated in polytomous and Problem Solving and Inquiry (PSI) items in TIMSS, since the current study has only examined exaninee timing on multiple-choice items.

Conclusions

The potential of process data in reshaping the way we perceive and evaluate testing processes cannot be overstated. The introduction of the STM and UTM indicators, which effectively combine accuracy and timing data, presents a promising way forward. They allow for a deeper, more nuanced understanding of test-taking behaviors, especially in relation to speed (responding in less time than the median response time) and accuracy (whether an item is correct or incorrect), and can be adapted across various testing environments. These indicators could provide essential insights to differentiate between students who answer questions quickly due to mastering content, and those who do so due to lack of effort. So STM appears to signify mastery of the test content and is more prevalent among confident students in higher benchmark levels. In contrast, UTM seems to signify a lack of test-taking effort, appearing more frequently in lower benchmark students and those disengaging from the test. These differentiated insights of test-taking behaviors can have profound implications regarding the use of timing data which challenge the assumption that rapid responses are only indicative of careless behaviors. As a result, these indicators can enhance the validity of study findings by considering examinee behaviors in more clarity and in more depth.

However, more research is needed to fully understand these indicators. Questions remain about how these variables perform across different subjects, age groups, and cultural contexts, and how they could be applied to different types of test items. Despite these limitations and the need for further research, this study represents an important step forward in the understanding of test-taking behaviors. It opens up a rich new dimension of test analysis that goes beyond the identification of careless responding, offering a far more sophisticated understanding of examinee behaviors and test-taking strategies, reliability, and usefulness of computerized large-scale testing.

Abbreviations

IEA	International Association for the Evaluation of Educational Achievement
PV	Plausible values
PSI	Problem solving and inquiry
STM	Successful time management
TIMSS	Trends in International Mathematics and Science Study
UTM	Unsuccessful time management

Acknowledgements
The authors would like to thank Mark Reckase who has provided feedback in relation to this manuscript.

Author contributions

EP had the original conceptualization of the idea, although it was refined by both authors. Both authors were involved in the analysis of the data, the write up, the revisions, and the approval of the final manuscript.

Funding
No funding sources were obtained in relation to this research.

Availability of data and materials

The dataset of TIMSS 2019 is available in the IEA Study Data Repository https://www.iea.nl/data-tools/repository/timss.

Declarations

Ethics approval and consent to participate

Since this is a secondary data analysis based on publicly available data, there was no need to obtain ethics approval of the current study.

Consent for publication

All authors have provided consent for publication.

Competing interests

The authors declare that they have no competing interests.

Received: 31 May 2023 Accepted: 1 February 2024
Published online: 13 February 2024

References

American Educational Research Association, & National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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, 16*(3), 441–462. https://doi.org/10.1007/BF03173192

Bennett, R. E. (2018). Educational assessment: What to watch in a rapidly changing world. *Educational Measurement: Issues and Practice, 37*(4), 7–15. https://doi.org/10.1111/emip.12231

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

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

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*, 173–183. https://doi.org/10.1080/08957347.2016.1171766

Ivanova, M. G., Michaelides, M. P., & Eklöf, H. (2020). How does the number of actions on constructed-response items relate to test-taking effort and performance? *Educational Research and Evaluation, 26*(5–6), 252–274. https://doi.org/10.1080/13803611.2021.1963939

Michaelides, M. P., Brown, G. T. L., Eklöf, H., & Papanastasiou, E. (2019). *Motivational profiles in TIMSS mathematics: Exploring student clusters across countries and time*. IEA Research for Education and Springer Open. https://doi.org/10.1007/978-3-030-26183-2

Michaelides, M. P., & Ivanova, M. (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.

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

Papanastasiou, E. C. (2015). Psychometric changes on item difficulty due to item review by examinees. *Practical Assessment, Research and Evaluation, 20*, 3. https://doi.org/10.7275/jcyv-k456

Papanastasiou, E. C. (2020). Do non-responses speak louder than words? Examining patterns of item non-response in TIMSS 2015. *International Journal of Quantitative Research in Education, 5*(2), 157–172. https://doi.org/10.1504/IJQRE.2020.111459

Papanastasiou, E. C., & Eklof, H. (2020). Editorial. *International Journal of Quantitative Research in Education, 5*(2), 157–172. https://doi.org/10.1504/IJQRE.2020.111459

Papanastasiou, E. C., & Stylianou-Georgiou, A. (2022). Modelling test-taking metacognition and achievement. *Assessment in Education: Principles, Policy & Practice, 29*(1), 77–94. https://doi.org/10.1080/0969594X.2022.2053945

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. https://doi.org/10.1186/s40536-021-00104-6

Rios, J. A., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-Scale Assessments in Education, 9*(1), 1–25. https://doi.org/10.1186/s40536-021-00110-8

Rios, J. A., & Soland, J. (2022). An investigation of item, examinee, and country correlates of rapid guessing in PISA. *International Journal of Testing*. https://doi.org/10.1080/15305058.2022.2036161

Rutkowski, D., & Wild, J. (2015). Stakes matter: Student motivation and the validity of student assessments for teacher evaluation. *Educational Assessment, 20*(3), 165–179. https://doi.org/10.1080/10627197.2015.1059273

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*(3), 213–232. https://doi.org/10.1111/j.1745-3984.1997.tb00516.x

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

Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education, 9*, 8. https://doi.org/10.1186/s40536-021-00100-w

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika, 87*(2), 593–619. https://doi.org/10.1007/s11336-021-09817-7

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research, 55*(3), 425–453. https://doi.org/10.1080/00273171.2019.1643699

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010). An investigation of the relationship between time of testing and test-taking effort. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, CO, Denver.

Wise, S. L., & Ma, L. (2012). Setting response time thresholds for a CAT item pool: The normative threshold method. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, Canada, Vancouver.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*(3), 237–252. https://doi.org/10.1080/08957347.2015.1042155

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. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education, 32*(4), 325–336. https://doi.org/10.1080/08957347.2019.1660350

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., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19–38. https://doi.org/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. https://doi.org/10.1080/08957347.2017.1353992

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

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.