


RESEARCH

Open Access



Identification and cross-country comparison of students' test-taking behaviors in selected eTIMSS 2019 countries

Xiaying Zheng^{1*} , Fusun Sahin², Ebru Erberber¹ and Frank Fonseca¹

*Correspondence:
xzheng@air.org

¹ American Institutes
for Research, 1400 Crystal Drive,
10th Floor Arlington, Arlington,
VA 22202, USA

² Curriculum Associates, 153
Rangeway Road, North Billerica,
Billerica, MA 01862, USA

Abstract

Unfavorable test-taking behaviors, such as speededness and disengagement, have long been a validity concern for large-scale low-stakes assessments. Understanding the presence and extent of such behaviors is important for ensuring the validity of inferences based on test scores. This study examined test-taking behaviors using item response time (RT), a process data-derived variable from the TIMSS 2019 database. Analyses compared the United States to three other countries (England, Singapore, and the United Arab Emirates) that administered the digital version of TIMSS (eTIMSS) 2019 in English at grade 8. Test-taking behaviors were identified within each country and compared within and across countries. Specifically, to identify distinct types of test-taking behaviors, mixture modeling was employed on RT and item scores from Booklet 1, Part 1, of the eTIMSS 2019 eighth-grade assessment. The results indicated that each country had several latent classes of students with different pacing trajectories and performance. The test-taking behaviors of these latent classes were labeled as *Steady*, *Disengaged* or *Very disengaged*, *Speeded* or *Very speeded*, and *Efficient and high-performing*. Most of the students in each country had a *Steady* pace (medium to high sum score; steady RT throughout the test): 71% in England, 74% in both Singapore and the United Arab Emirates, and 84% in the United States. *Disengaged* or *Very disengaged* students (low sum score; short RT) were identified in each country but were more prevalent in England and the United Arab Emirates (above 20% in both) than in the United States and Singapore (both below 10%). The study also revealed small percentages of *Speeded* or *Very speeded* students (low to medium sum score; long RT at first but very short RT toward the end) in England, the United Arab Emirates, and the United States (1%, 5%, and 6%, respectively) but not in Singapore. A unique class of *Efficient and high-performing* students (high sum score; short RT) was identified only in Singapore (24%). This study demonstrated that mixture modeling is a useful technique for identifying distinct test-taking behaviors and highlighted the presence and extent of unfavorable test-taking behaviors within each selected country using data from Booklet 1, Part 1, of the eTIMSS 2019 eighth-grade assessment.

Keywords: Process data, Log data, Finite mixture modeling, Disengagement, Rapid-guessing, Speededness, Large-scale assessments, TIMSS, Digital assessments, International comparisons

Introduction

Two long-standing concerns in large-scale assessments are examinees not sufficiently engaging with items (i.e., disengagement) and running out of time before responding to all items (i.e., speededness). Both behaviors are concerning because examinees' achievement scores under the conditions of speededness and disengagement typically do not represent their true ability, diminishing the validity of arguments based on the scores (Lu & Sireci, 2007; Wise, 2015; Yamamoto, 1995) and leading to potential biases in parameter estimates (Oshima, 1994; Rogers & Swaminathan, 2016). In the case of international large-scale low-stakes assessments, such as TIMSS, speededness and disengagement may also impact the validity of performance comparisons across countries and years.

Some studies have compared the percentage of disengaged examinees across countries (e.g., Debeer et al., 2014; Rios & Guo, 2020) and years (Kuang & Sahin, 2021). However, none have examined disengagement, speededness, and other test-taking behaviors simultaneously in a cross-country analysis. Identifying multiple test-taking behaviors simultaneously paints a fuller picture of how students spend their testing time and how test-taking behaviors relate to performance. Digitally based assessments allow for the collection of log data; that is, the accumulation of examinees' interactions with the testing screen and associated timestamps. Using these timestamps, it is possible to compute response times (RTs), which are then used to identify disengaged and speeded examinees (Lu & Sireci, 2007; Wise, 2017).

Recent studies have examined disengagement at the item level using either RT and scores (Goldhammer et al., 2017; Guo et al., 2016) or RT and response behaviors derived from process data (Sahin & Colvin, 2020). An alternative to *item*-level detection is *test*-level detection, where researchers analyze RTs over item sequences (i.e., pacing trajectories) using latent models, such as mixture modeling or growth modeling, to detect disengagement (Zheng, 2019; Zheng et al., 2018) or speededness (Bolt et al., 2002; Kahrman et al., 2013).

Our study is one of the first to compare speededness and disengagement simultaneously and also one of the first to use RT data from TIMSS 2019. This was the cycle in which TIMSS began the transition from a paper-and-pencil assessment to a computer-based assessment, introducing a digital version called "eTIMSS". Specifically, we used data from four education systems: England (ENG), Singapore (SGP), the United Arab Emirates (UAE), and the United States (USA). Our objectives were, first, to examine the presence and extent of unfavorable test-taking behaviors using the test-level mixture models within selected countries and, second, to investigate the commonality and specificity of the test-taking behaviors across the countries.

Data

This study compared eTIMSS 2019 eighth-grade mathematics data from the USA and three other countries: SGP, ENG, and UAE. The comparison countries were selected for the following reasons: (1) Like the USA, they administered the digital version, not the paper version, of TIMSS (i.e., eTIMSS) in 2019,¹ (2) like the USA, they administered the

¹ In 2019, TIMSS began the transition from a paper-and-pencil assessment to a computer-based assessment by introducing a digital version called "eTIMSS." Over half of the TIMSS 2019 participants opted for the digital version, while the remaining countries administered TIMSS in the paper-and-pencil format (paperTIMSS), as in previous assessment cycles. For the list of countries participated in the 2019 cycle of eTIMSS and paperTIMSS, please refer to Exhibit 2.2 of the *TIMSS 2019 User Guide for the International Database* (Fishbein et al. 2021).

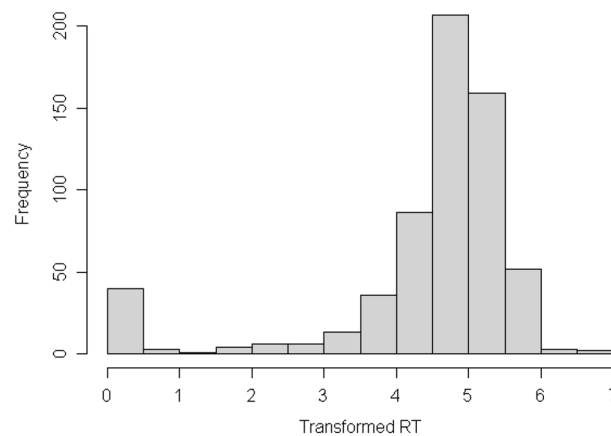


Fig. 1 Histogram of the transformed RT for an example item* responded to by USA eighth-graders * This is the last mathematics item in the eighth-grade TIMSS 2019 Booklet 1, Part 1. Note: Transformed $RT = \log(RT + 1 \text{ s})$. Sample size for USA is 618. Of the 624 eighth-grade USA students who were administered Booklet 1, 6 of them did not have valid RT data for any of the 32 items examined in the study. Source: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2019

eighth-grade assessment in the English language,² and (3) compared to the USA, they had varying mathematics performance at the eighth-grade in TIMSS 2019. Specifically, SGP had performance higher than the USA's; ENG had performance similar to the USA's; and the UAE had performance lower than the USA's.³

The study used item scores and response time variables (i.e., total time spent by the student on each item screen) for one part of a test booklet from the student achievement data files in the TIMSS 2019 public-use international database (Fishbein et al., 2021).

TIMSS uses a matrix sampling approach that packages mathematics and science items into 14 booklets, with each student completing just one booklet. As shown in Exhibit 4.2 of the TIMSS 2019 Assessment Frameworks Mullis & Martin. (2017), each booklet consists of two parts (Part 1 and Part 2), and each part contains two blocks of items (either two mathematics blocks or two science blocks). In half of the 14 booklets, the two mathematics blocks come first, and then the two science blocks, and in the other half the order is reversed. Students are allotted 45 min in the eighth-grade assessment to complete each part of the booklet. Booklets are distributed such that approximately equal proportions of students respond to each booklet and the students completing each booklet are approximately equivalent in terms of student ability Mullis & Martin. (2017).

To the extent possible, within each block the distribution of items across the TIMSS content and cognitive domains matches the overall item pool distribution Mullis & Martin. (2017). This important design feature of TIMSS made either part of any booklet equally suitable for examination in this study. However, to limit the scope, this study focused only on one of the seven booklets that started with two mathematics blocks;

² Exhibit 5.3 of the *TIMSS 2019 Technical Report* (Martin et al., 2020) lists the target languages used for the TIMSS 2019 eighth-grade assessment. The USA, ENG, and SGP administered the test in English. The UAE administered the test in English and Arabic, although only about a quarter of the UAE students took the test in Arabic.

³ Figure 1b of the *TIMSS 2019 U.S. Highlights Web Report* (National Center for Education Statistics, 2021) presents average scores of eighth-grade students on the TIMSS 2019 scale by country.

specifically, blocks ME01 and ME02 that are administered in Part 1, Booklet 1, of the 2019 TIMSS eighth-grade assessment. These blocks consisted of 31 mathematics items totaling 32 score points. Additional file 1: Table S1 lists the characteristics of these 31 items (e.g., item type, content domain, cognitive domain, item label) in the order in which they were administered during the first 45 min of Booklet 1 of the eighth-grade assessment. Examinees' responses were scored using the scoring syntax provided in the international database where "Omitted" and "Not Reached" responses were recoded as incorrect.⁴ For the 31 mathematics items examined, the database included 28 RT variables because several items shared the same screen.

Methods

This research used finite mixture models (see, e.g., Everitt & Hand, 1981; McLachlan & Basford, 1988; Titterton et al., 1985) to identify unique groups of examinees. Finite mixture modeling is a model-based clustering method that has been used to identify unobserved (latent) class memberships based on observed characteristics. In our study, the observed input variables were examinees' RTs and item scores. If \mathbf{y} is a vector of the 28 RT and 31 score variables, the joint distribution of the observed \mathbf{y} and the latent class membership can be expressed as:

$$f(Y = \mathbf{y}) = \sum_{k=1}^K \pi_k f_k(\mathbf{y} | \mu_k, \Sigma_k), \quad (1)$$

where K is the total number of latent classes enumerated in the mixture model, π_k is the probability for class k , and f_k is a function of observing \mathbf{y} for class k with means of μ_k and covariance structure of Σ_k . Various Σ_k could be tested to identify the model with the best fit. As the RT variables were continuous and the item scores were ordinal, we only tested covariances of the RT variables. The covariances between RTs and item scores were not considered due to the complexity involved in parameterizing them as well as the lack of one-to-one correspondence for all item scores and RTs.

Maximum likelihood estimation was used iteratively with increasing numbers of classes to determine the best-fitting model. Due to the clustering of students within classrooms, the design-based sandwich estimator (Asparouhov & Muthén, 2006; Rabe-Hesketh & Skrondal, 2006) was used to estimate robust standard errors, which were adjusted for clustering and stratification. Each model was evaluated using fit indices, such as entropy (Celeux & Soromenho, 1996) and the Bootstrapped Likelihood Ratio Test (BLRT; McLachlan & Peel, 2000). Values for entropy range from 0 to 1, with higher values indicating better separation of latent classes. A general rule of thumb is that an entropy over 0.8 would indicate good distinctions of latent classes (see, e.g., Clark, 2010; Nagin, 2005). The BLRT compares the fit of a model with K classes to one with $K-1$ classes. If these tests are significant, the model with the higher number of classes is favored. Each model was also evaluated on its interpretability based on the observed

⁴ In TIMSS 2019, not-reached items were treated as incorrect responses, except during the item calibration step of the IRT scaling. During calibration, not-reached items were considered to have not been administered (Fishbein et al. 2021).

characteristics of the classes. The mixture model was run separately for each country using one to five classes.

The item-level RTs needed to be transformed because they were highly skewed. In RT modeling, natural logarithm transformations are commonly used on timing data. In our data, some students did not interact with certain items (e.g., if they ran out of time before reaching an item). The RT for these items is 0, which cannot be transformed with a natural logarithm. To accommodate these cases, the RT was transformed by taking the natural logarithm of examinees' RT (in seconds) plus 1. If an examinee did not spend time on an item, their transformed RT would be 0.

Figure 1 presents an example of the transformed RT distribution for the last item in the USA sample. Of the 618 eighth-graders, approximately 40 did not interact with this item. Therefore, the distribution is slightly zero-inflated.

Following mixture modeling, we examined the validity of the classifications by conducting cross tabulations to explore the association between the classifications and selected student contextual variables. Specifically, we focused on two questions from the U.S. national version of the student questionnaire: Question #32, with variable name BSXG32, which asked about examinees' effort while taking the test; and Question #33, with variable name BSXG33, which asked about examinees' perception of the importance of the test.⁵ These two questions directly relate to students' test-taking behaviors. We hypothesized that students classified as *Disengaged* in the USA would report lower effort on the test and lower perceived importance of the test.

However, these two questions were not included in the international version of the student questionnaire. As a result, data on these questions were not available for the other three countries included in our study (SGP, ENG, and the UAE). To address this limitation, we selected a student questionnaire variable that was available in the TIMSS 2019 international database for all eTIMSS countries: Question 1Ba, with variable name BSBE01BA, which asked whether students had difficulty typing during the test.⁶ This variable served as a proxy of students' familiarity with using computers, as some students might have been disengaged due to experiencing difficulties with the computer-based testing platform. We hypothesized that a greater percentage of disengaged students would report having trouble with typing in the test. Additionally, we examined students' gender distribution (using ITSEX variable) within the identified classes, as previous research (Wise & DeMars, 2010) has shown that male students are typically overrepresented in the disengaged class. We hypothesized that the same pattern would emerge in the identified classes.

Results

Mixture modeling results

Through testing different mean and covariance structures for the transformed RTs in the mixture model, we found that including the class-specific variance-covariance matrices led to nonconvergence due to the large numbers of parameters to be

⁵ Questions #32 and #33 in the U.S. version of the TIMSS 2019 of the eighth-grade student questionnaire, which is available at https://nces.ed.gov/timss/pdf/T19_GR8_StudentQ_USA_Questionnaire.pdf.

⁶ Question #1Ba in the TIMSS 2019 eighth-grade student questionnaire-eTIMSS supplement, which is available at <https://timssandpirls.bc.edu/timss2019/questionnaires/index.html>.

Table 1 Fit indices for 2 to 5-class mixture models for the USA, ENG, UAE, and SGP

Fit index		USA	ENG	UAE	SGP
BLRT p-values	2-class vs. 1-class	< 0.01	< 0.01	< 0.01	< 0.01
	3-class vs. 2-class	< 0.01	< 0.01	< 0.01	< 0.01
	4-class vs. 3-class	< 0.01	< 0.01	< 0.01	< 0.01
	5-class vs. 4-class	< 0.01	< 0.01	< 0.01	NA
Entropy	2-class	0.941	0.945	0.938	0.969
	3-class	0.965	0.970	0.947	0.968
	4-class	0.974	0.957	0.960	0.937
	5-class	0.979	0.962	0.948	NA

estimated. Therefore, we constrained the variance of the transformed RT of a specific item to be equal across classes but allowed the variances of transformed RTs to vary across items within a class. The covariances of transformed RTs were not included in the model for convergence consideration.

The BLRT p-values and the entropies for the mixture models are presented in Table 1. The entropies were high (> 0.9 , above the recommended threshold of 0.8) for all models, indicating very good separations between the latent classes within each model. Thus, entropy itself did not sufficiently differentiate between the models. As a result, more weight was placed on the BLRT p-values which explicitly compared the fits of two solutions. As the number of classes increased, the BLRT p-values pointed to 5-class solutions for the USA, ENG, and the UAE using a significance level of 0.05. For SGP, the 5-class model did not converge; instead, a 4-class solution converged and gave the best model fit. The final results adopted 5-class solutions for the USA, ENG, and UAE and a 4-class solution for SGP.

Figures 2, 3, 4, 5 present the results for the USA, ENG, UAE, and SGP, respectively. Using median transformed RT, the two upper panels show the pacing trajectories of the classes across the items in each of the two blocks examined. The bottom-left panel presents the cumulative RT in minutes across the item sequence in both blocks. The bottom-right panel shows the percentage distribution and average sum score (SS) of each class found in the sample. Table 2 summarizes the results for all four countries.

For the USA sample (Fig. 2), we identified five classes and labeled them as *Disengaged*, *Very speeded*, *Speeded*, *Steady and high-performing*, and *Steady but low-performing*. About 9% of the sample was identified as *Disengaged* (shown in black). These students consistently spent relatively less time across all items (median total RT was about 20 of the 45 min allotted) and had low mean sum scores (about 5 of a possible 32 points). Based on students' timing patterns and low performance, we concluded that they might not have fully considered each item.

Two classes were interpreted as speeded; both are characterized by a large amount of time spent on the items in the first block and little to no time spent on the last few items in the second block. The *Very speeded* class (magenta) began to run out of time earlier than the *Speeded* class (purple) (at the 21st vs. the 25th item, respectively) and had a lower average sum score (10 vs. 12 points, respectively). The cumulative RT of the *Very speeded* group plateaued earlier than the *Speeded* group's cumulative RT.

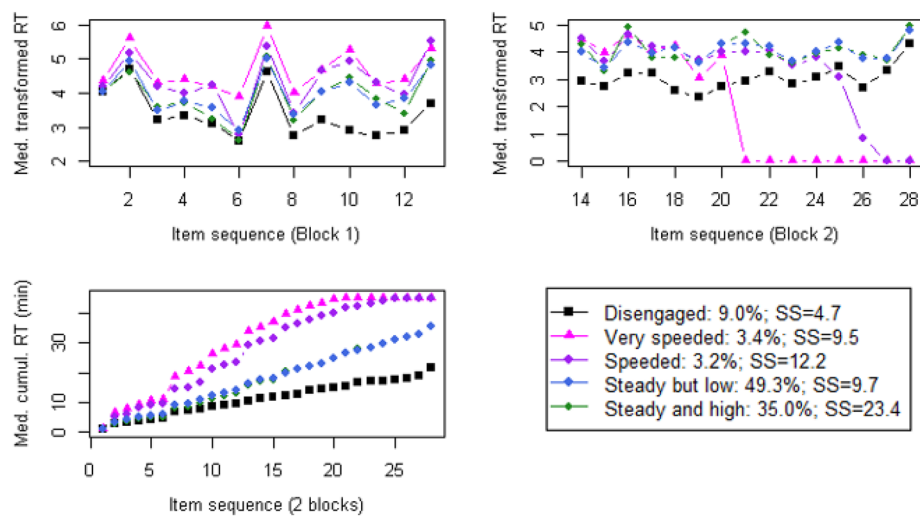


Fig. 2 Pacing trajectories of student classes identified in USA sample Note: Pacing trajectories are depicted for the 28 RT variables for the 31 mathematics items in the eighth-grade TIMSS 2019 Booklet 1, Part 1. Transformed $RT = \log(RT + 1 \text{ s})$. SS is the average sum score. Sample size for the USA is 618. Of the 624 eighth-grade USA students who were administered Booklet 1, 6 students did not have valid RT data for any of the 32 items examined in the study. The median cumulative response times for the two steady groups (green and blue) overlap, making only one of them (blue) visible on the graph. Source: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2019.

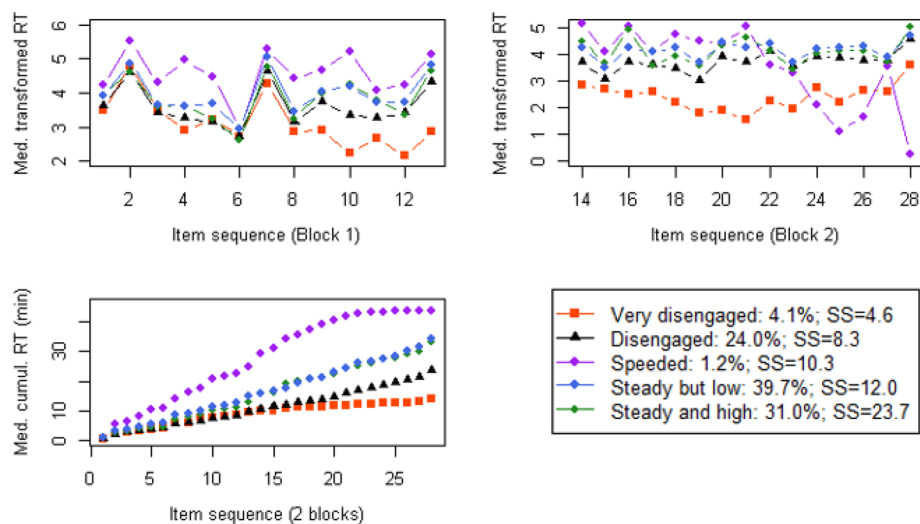


Fig. 3 Pacing trajectories of student classes identified in ENG sample NOTE: Pacing trajectories are depicted for the 28 RT variables for the 31 mathematics items in the eighth-grade TIMSS 2019 Booklet 1, Part 1. Transformed $RT = \log(RT + 1 \text{ s})$. SS is the average sum score. Sample size for ENG is 242. The median cumulative response times for the two steady groups (green and blue) overlap, making one of them (blue) more visible in the graph. Source: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2019.

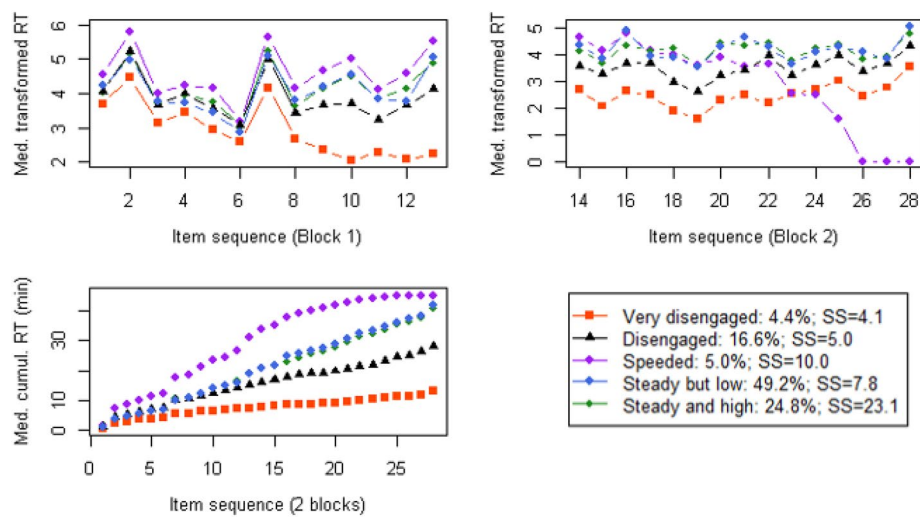


Fig. 4 Pacing trajectories of student classes identified in UAE sample Pacing trajectories are depicted for the 28 RT variables for the 31 mathematics items in the eighth-grade TIMSS 2019 Booklet 1, Part 1. Transformed $RT = \log(RT + 1 \text{ s})$. SS is the average sum score. Sample size for the UAE is 1595. Of the 1599 eighth-grade USA students who were administered Booklet 1, 6 students did not have valid RT data for some or all of the 32 items examined in the study. The median cumulative response times for the two steady groups (green and blue) overlap, making one of them (blue) more visible in the graph. Source: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2019.

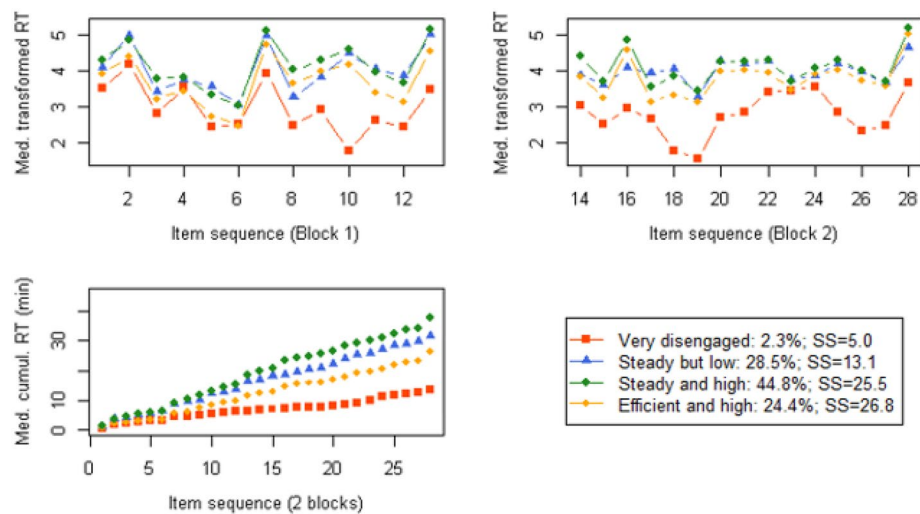


Fig. 5 Pacing trajectories of student classes identified in SGP sample NOTE: Pacing trajectories are depicted for the 28 RT variables for the 31 mathematics items in the eighth-grade TIMSS 2019 Booklet 1, Part 1. Transformed $RT = \log(RT + 1 \text{ s})$. SS is the average sum score. Sample size for SGP is 348. Of the 350 eighth-grade SGP students who were administered Booklet 1, 2 students did not have valid RT data any of the 32 items examined in the study. Source: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2019.

Table 2 Latent class distributions, mean sum scores, and median total response time (RT) of the selected countries in the eighth-grade TIMSS 2019 Booklet 1, Part 1

Performance	Latent Class	% Distributions				Mean Sum Score				Median Total RT (min)			
		SGP	USA	ENG	UAE	SGP	USA	ENG	UAE	SGP	USA	ENG	UAE
Low	Very disengaged	2	-	4	4	5	-	5	4	14	-	14	13
	Disengaged	-	9	24	17	-	5	8	5	-	22	23	28
Low to Medium	Very speeded	-	3	-	-	-	10	-	-	-	45	-	-
	Speeded	-	3	1	5	-	12	10	10	-	45	44	45
	Steady	29	49	40	49	13	10	12	8	32	37	34	42
High	Steady	45	35	31	25	26	23	24	23	38	36	33	41
	Efficient	24	-	-	-	27	-	-	-	26	-	-	-

Disengaged or Very disengaged students are those with low sum scores and relatively short response times. Speeded or Very speeded students are those with low to medium sum scores and long response times at first but very short response times toward the end. Steady-paced students are those with medium to high sum scores and steady response times throughout the test. Efficient students are those with high sum scores and relatively short response times. Sample sizes are as follows: 618 in the USA, 242 in ENG, 1595 in the UAE, and 348 in SGP. Source: International Association for the Evaluation of Educational Achievement (IEA), Trends in International Mathematics and Science Study (TIMSS), 2019

The remaining two majority classes had an almost identical, steady pace and spent 35–40 min of the 45 min allotted, although their performance differed greatly. The *Steady and high-performing* group (green) scored 23 points on average, while the *Steady but low-performing* group (blue) scored about 10 points.

For the ENG and UAE samples (Figs. 3, 4, respectively), we identified the *Speeded*, *Disengaged*, *Steady and high-performing*, and *Steady but low-performing* classes (as in the USA), but not the *Very speeded* class. Instead, we observed a *Very disengaged* group, shown in red (about 4% of each country's sample), which spent only about 15 min on all the items out of the 45 min allotted.

For the SGP sample (Fig. 5), we observed the *Very disengaged*, *Steady and high-performing*, and *Steady but low-performing* groups, but no speeded group. Instead, SGP had a unique *Efficient and high-performing* class, shown in orange (24.4% of the sample), characterized by a relatively fast responding pace (about 25 min total out of the 45 min allotted) and relatively high score (27 points, on average).

Table 2 summarizes class memberships and characteristics across all four countries. In all countries, most students belonged to the two steady-pacing groups, but we consistently identified some disengaged students, even in the high-performing SGP. The proportion of disengaged students was relatively small in the USA and SGP (lower than 10%), but over 20% in ENG and the UAE. Speeded examinees were found in all countries but SGP, which had a unique *Efficient and high-performing* class not observed elsewhere.

Relating class membership to student contextual variables

The crosstabulations of the two USA-specific questions with the latent classes are shown in Additional file 1: Table S2. For the question on effort on the test, a considerably higher percentage of *Disengaged* USA students (41%) reported trying “not as hard on TIMSS as on other tests” compared to students in the other classes (less than 30%). This confirmed our interpretation of the *Disengaged* class as having lower level of effort on the test. In addition, we noted that a comparatively larger percentage of high-performing *Steady* students (61%) reported trying “about as hard as on other tests” while a comparatively lower percentage tried “harder” or “much harder”, which is in line with the expectation

that the students in the high-performing *Steady* class did not require as much effort due to their higher performance. For the question on the perceived importance of the test, a considerably higher percentage of *Disengaged* USA students (29%) reported that it was “not very important to do well” compared to the students in the other classes (less than 13%). This further supports our hypothesis that disengaged students placed less importance on the test than other students.

Regarding the question on experiencing difficulty with typing, the results for the four countries are shown in Additional file 1: Table S3. In the USA, UAE, and SGP, higher percentages of *Very disengaged* or *Disengaged* students reported experiencing difficulty with typing compared to students in the other classes. This supports our hypothesis that disengagement may be related to typing difficulties, which we considered as a proxy for lack of familiarity with the digital testing platform. However, in ENG, a higher percentage of students experiencing typing difficulty was observed in the high-performing *Steady* class (45%), not in the *Very disengaged* or *Disengaged* class (35%).

Lastly, the results for the gender variable are presented in Additional file 1: Table S4 for the four countries. In the USA, UAE, and SGP, a significantly higher percentage of male students were disengaged compared to female students, consistent with previous research. Again, ENG stood out as the only exception, with a higher proportion of female students being disengaged. This finding suggests the presence of unique contextual factors in ENG that may have contributed to this outcome, warranting further investigation. Overall, the response patterns of the identified latent classes mostly confirmed our hypotheses and provided additional validity evidence for their interpretations.

Discussion

Classes discovered

In this study, the classes discovered were *Very disengaged*, *Disengaged*, *Very speeded*, *Speeded*, *Steady but low-performing*, *Steady and high-performing*, and *Efficient and high-performing*. Compared to previous studies, this study identified more fine-grained disengaged and speeded test-taking behaviors. Specifically, we differentiated between *Disengaged* and *Very disengaged* students as well as between *Speeded* and *Very speeded* students. Even though Wise and Kong's (2005) response time effort (RTE) index considered disengagement on a continuum, it did not differentiate between *Very disengaged* and *Disengaged* behaviors. The RTE index reflects the number of items on which a student is identified as providing non-effortful responses out of the total number of items administered. If a student has an RTE index value of 0, the lowest possible value, the suggestion is that the student rushed through all the items, thus showing *Very disengaged* behavior. On the other hand, if a student has an RTE index value of 1, the highest possible value, the student did not rush through any of the items and is said to be fully engaged. The shortcoming of the RTE index is that it does not have an established cut-off point to distinguish *Very disengaged* from *Disengaged* students. In addition, there is no established RTE-like index to measure speededness, let alone to differentiate *Speeded* from *Very speeded* students.

Another limitation of many previous studies that detect test-taking behaviors is that some efficient test takers can be misclassified as speeded or disengaged, particularly when only response times are used for detection. For example, high-performing students

may not find the test challenging and, therefore, spend little time on many items. Consequently, in the absence of scores, they may be misclassified as speeded or disengaged. To overcome the potential misclassifications, in this study, we used both response times and item scores. This joint model allowed us to differentiate between speeded, disengaged, and efficient students. We provided further validity evidence for the classifications via examining selected student contextual background variables as described in the Results section.

It should be noted that the two adjacent mathematics blocks (ME01 and ME02) of this study appeared in the first part of Booklet 1 of the eighth-grade eTIMSS 2019 assessment. The same two blocks also appeared in booklets 14 and 2, respectively, but the second part of those booklets. When the blocks are in a different position or timed together with some other blocks, the mixture model results may differ due to the change in the test-taking experience. Consequently, we advise caution in generalizing analysis results from booklet 1 without further evidence from other booklets.

Country comparisons

This study revealed distinct differences among countries in the existence and prevalence of various classes of test-taking behaviors. For example, the *Disengaged* and *Very disengaged* classes were more prevalent in England and the United Arab Emirates than in the United States or Singapore. Compared to the United States, the percentages of *Disengaged* and *Very disengaged* students were higher both in England, a country with eTIMSS performance similar to that of the United States, and the United Arab Emirates, a country with eTIMSS performance lower than that of the United States. Even in Singapore, a high-performing country, we were still able to identify a small *Very disengaged* group of students. These results suggest that the existence or prevalence of *Disengaged* students may not be directly related to country performance. We hypothesize that observing disengagement may have been related to the low-stakes nature of TIMSS as in other international large-scale assessments. Additionally, we found an *Efficient and high-performing* group only in Singapore, which is the highest performing country in our sample.

Methodological considerations

It should be noted that mixture modeling is a probability-based exploratory technique. Although entropy was high in all models, some examinees could have been misclassified. Thus, although we could make plausible interpretations about the classes at the group level, one should cautiously interpret these classifications at the individual student level.

We conducted the analyses separately for each country as opposed to analyzing a pooled dataset from the four countries. One reason for doing this was to be able to identify small but unique clusters within each country we examined. In general, working with a pooled dataset would be disadvantageous when the expectation is of small groups with distinct behaviors. Because both disengagement and speededness were expected to be nondominant testing behaviors (typically reported for less than 10% of the test takers), working with smaller but meaningful partitions of the data—in this case, split by countries—allowed us to identify these behaviors with more precision. Furthermore, we might not have identified the *Efficient and high-performing* class found in Singapore if

we had combined the data from all countries and identified clusters within this pooled dataset.

In this study, we examined only the first part of one of the seven eighth-grade TIMSS 2019 test booklets that started with two mathematics blocks, drawing on data from 4 out of the 27 education systems that administered TIMSS 2019 digitally at grade 8. Similarly, we analyzed only one type of log-data-derived variable (i.e., item RT). Future studies can examine data from more countries, from the fourth-grade assessment, from the other six eighth-grade booklets starting with two mathematics blocks, or seven eighth-grade booklets starting with two science blocks. Future studies could also use additional log data-derived variables, such as frequency of item visits, to see if other distinct behaviors can be identified. These studies might find that students from other countries or in fourth-grade would interact differently with booklets containing different items.

Conclusions

This study is one of the first to examine distinct test-taking behaviors using response time, a process data variable included in the TIMSS 2019 database. As a result of this study, we discovered that even though most students in a country followed a steady testing pace, almost every country examined had students who demonstrated *Speeded*, *Very speeded*, *Disengaged*, and *Very disengaged* behavior. Also, an *Efficient and high-performing* group was found only in Singapore, which had the highest performance in our sample. Thus, future studies may consider studying the prevalence of efficient test-taking behavior, particularly in other high-performing countries.

The prevalence of the *Disengaged* group was not linearly related to the achievement of the countries in the study. For example, a *Very disengaged* student group was found in 2% of the sample in Singapore (one of the highest performing countries in TIMSS 2019 at eighth-grade), and in 4% of the sample both in England (a country that performed above the TIMSS scale center-point, 500, in TIMSS 2019 at eighth-grade) and the United Arab Emirates (a country performed below the TIMSS scale center-point, 500, in TIMSS 2019 at eighth-grade). Similarly, the *Disengaged* group was found in 9% of the United States, 17% in the United Arab Emirates, and in 24% of the England.

In addition, there was little difference in the percentage of students identified as *Speeded* across countries (between 1 and 5%) except in Singapore, where we did not observe any students in that class. A *Very speeded* class was observed in the United States sample (3%), again suggesting that there is not a linear relationship between achievement and speededness at the country level.

This study demonstrated that mixture modeling is a useful technique for identifying various test-taking behaviors simultaneously. Previously, it was used to study only one of the unfavorable testing behaviors in a two-cluster model that classified examinees into either “speeded” or “not speeded” groups (e.g., Schnipke & Scrams, 1997) or “disengaged” or “not disengaged” groups (e.g., Wang et al., 2018). This study also demonstrated that mixture modeling can differentiate between levels of speededness and disengagement. Identifying unfavorable testing behaviors simultaneously and differentiating the degree of these behaviors can help testing programs ensure the validity of inferences based on test scores. For example, testing programs can flag students who show these unfavorable behaviors and embed warning systems in the testing platform to prevent

these behaviors on the fly. In addition, low-stakes testing programs may consider incorporating incentives, as appropriate, for students to show their best effort.

Abbreviations

BLRT	Bootstrapped Likelihood ratio test
ENG	England
eTIMSS	Digital version of the Trends in International Mathematics and Science Study
Item RT	Item response time
RT	Response time
RTE	Response time effort
SGP	Singapore
SS	Sum scores
TIMSS	Trends in International Mathematics and Science Study
UAE	United Arab Emirates
USA	United States

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40536-023-00179-3>.

Additional file 1: Table S1. Item Information for the 31 mathematics items included in the study. **Table S2.** Percentage of U.S. students in the latent classes reporting on “how hard they tried on the test” and “how important it was for them to do well on the test” in the TIMSS 2019 eighth-grade Booklet 1. **Table S3.** Percentage of students reporting “difficulty typing on the test” in the latent classes within the selected countries in the TIMSS 2019 eighth-grade Booklet 1. **Table S4.** Percentages of male and female students in the latent classes within the selected countries in the TIMSS 2019 eighth-grade Booklet 1.

Acknowledgements

The authors are grateful to Dr. Markus Broer and Dr. Ting Zhang at the American Institutes for Research (AIR) for their reviews and comments in the preparation of this manuscript.

Author contributions

Not applicable.

Funding

This research was supported by the National Center for Education Statistics (NCES) within the U.S. Department of Education.

Availability of data and materials

The TIMSS datasets analyzed in this study are publicly available on the TIMSS website (<https://timss2019.org/international-database/>).

Declarations

Competing interests

No competing interests.

Received: 9 December 2022 Accepted: 24 July 2023

Published online: 01 September 2023

References

- Asparouhov, T., & Muthén, B. (2006). Multilevel modeling of complex survey data. In *Proceedings of the joint statistical meeting in Seattle* (pp. 2718–2726).
- Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement*, 39(4), 331–348.
- Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2), 195–212.
- Clark, S. L. (2010). *Mixture modeling with behavioral data* Doctoral dissertation. California: University of California.
- Debeer, D., Bucholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA Reading Assessment. *Journal of Educational and Behavioral Statistics*, 39, 502–523.
- Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. Chapman and Hall.
- Fishbein, B., Foy, P., & Yin, L. (2021). TIMSS 2019 user guide for the international database. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <https://timssandpirls.bc.edu/timss2019/international-database/>
- Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments in Education*, 5(1), 18.

- 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.
- Kahraman, N., Cuddy, M. M., & Clauser, B. E. (2013). Modeling Pacing Behavior and Test Speededness Using Latent Growth Curve Models. *Applied Psychological Measurement*, 37(5), 343–360.
- Kuang, H., & Sahin, F. (2021). *Is item disengagement different across years? Comparisons between PISA 2018 and 2015*. Paper presented at the annual meeting, conducted virtually, of the National Council on Measurement in Education.
- Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement: Issues and Practice*, 26(4), 29–37.
- Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and procedures: TIMSS 2019 technical report*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <https://timssandpirls.bc.edu/timss2019/methods>
- McLachlan, G. J., & Basford, K. E. (1988). *Mixture models Inference and applications to clustering* (Vol. 38). M. Dekker.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. Wiley.
- Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: <http://timssandpirls.bc.edu/timss2019/frameworks/>
- Nagin, D. S. (2005). *Group-based modeling of development*. Harvard University Press.
- National Center for Education Statistics (NCES). (2021). *TIMSS 2019 U.S. highlights web report* (NCES 2021–021). U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics. Available at <https://nces.ed.gov/timss/results19/index.asp>.
- Oshima, T. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement*, 31, 200–219.
- Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society Series a: Statistics in Society*, 169(4), 805–827.
- Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education*, 33(4), 263–279.
- Rogers, H. J., & Swaminathan, H. (2016). *The effect of unmotivated examinees on field test item calibrations*. Paper presented at annual meeting of the National Council on Measurement in Education, Washington, DC.
- Sahin, F., & Colvin, K. F. (2020). Enhancing response time thresholds with response behaviors for detecting disengaged examinees. *Large-Scale Assessments in Education*, 8(5), 1–24.
- 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.
- Titterton, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Wiley.
- Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting aberrant behavior and item preknowledge: A comparison of mixture modeling method and residual method. *Journal of Educational and Behavioral Statistics*, 43(4), 469–501.
- Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education*, 28(3), 237–252.
- Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice*, 36(4), 52–61.
- Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, 15(1), 27–41.
- Wise, S. L., & Kong, X. J. (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
- Yamamoto, K. (1995). Estimating the effects of test length and test time on parameter estimation using the HYBRID model (Report No TR-95–2). *Educational Testing Service*, 1995(1), 39.
- Zheng, X., Beverly, T., & Kim, Y. Y. (2018). Identifying rapid-guesser using growth mixture models. In M. Broer (Chair) Students' use of response time, testing behavior, and performance in digitally-based assessments. Paper presented at the meeting of the National Council on Measurement in Education. New York City.
- Zheng, X. (2019, April). Identifying rapid-guesser using growth mixture models. In M. Broer (Chair), Testing strategies, extended time accommodation, and speededness, using process data in NAEP. Symposium conducted at the annual meeting of the National Council on Measurement in Education, Toronto, Canada.

Publisher's Note

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