International Journal of Assessment Tools in Education
2023, Vol. 10, Special Issue, 224–242
https://doi.org/10.21449/ijate.1368600
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The use of on-screen calculator as a digital tool in technology-enhanced items

Ozge Ersan[1*], **Burcu Parlak**[1]

[1]Ministry of National Education, General Directorate of Measurement, Evaluation and Examination Services, Ankara. Türkiye

ARTICLE HISTORY
Received: Sep. 29, 2023
Accepted: Dec. 18, 2023

Keywords:
Technology-enhanced items,
Calculator use,
Mathematics achievement,
Trends in mathematics and science study (TIMSS),
Problem solving and inquiry tasks.

Abstract: In this study, the effect of using on-screen calculators on eighth grade students' performance on two TIMSS 2019 Problem Solving and Inquiry Tasks items considered as examples of technology-enhanced items administered on computers was examined. For this purpose, three logistic regression models were run where the dependent variables were giving a correct response to the items and the independent variables were mathematics achievement and on-screen calculator use. The data of student from 12 countries and 4 benchmarking participants were analyzed and some comparisons were made based on the analyses. The results indicate that using on-screen calculators is positively associated with higher odds of giving correct responses for both items above and beyond students' mathematics achievement scores. The results of this study promote the inclusion of on-screen calculator as a digital tool in technology-enhanced items that require problem solving.

1. INTRODUCTION

Item types used in the assessments have evolved to be richer in technological features following the widespread use of computerized testing. These new types of items differ from the conventional multiple-choice (MC) and constructed-response (CR) items in terms of technological innovations. To define technology-enhanced (TE) items, Parshall et al. (2010) provided seven facets that each of these facets can vary at different levels of innovations in the items. These facets are: (a) assessment structure, (b) response action, (c) interactivity, (d) media inclusion, (e) fidelity, (f) complexity, and (g) scoring method.

The assessment structure describes how a TE item is formatted. A taxonomy for assessment structure for e-assessment items are described in the literature (Scalise & Gifford, 2006, p. 9). The structure of TE items can vary from the most constrained (multiple-choice) form to the least constrained (presentation/portfolio) forms, and the items in between were referred to as intermediate constraint items (selection/identification, reordering/rearrangement, substitution/correction, completion, construction). Response action indicates how item responses were collected such as by mouse clicks, keyboard typing, or voice recording. Interactivity refers to how test takers interact with the item such as running a science simulation

*CONTACT: Ozge ERSAN ozge.ersan09@gmail.com Milli Eğitim Bakanlığı, Ölçme Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü, Ankara. Türkiye

e-ISSN: 2148-7456 /© IJATE 2023

or using item tools such as magnifier, highlighter or ruler for the item. Media inclusion indicates that a graphic, picture, short animation, or a sound clip may be added to the item stem or response options. Fidelity refers to the realistic and accurate representation of a scenario, task, graph, or picture. Complexity of an item indicates how each facet of innovations are combined during item development phase such as item structure, number of response options, number of supporting materials, multiple forms of response actions, as well as the design of item interface. Last, scoring method indicates a strategy for translating all inputs of the test taker into a quantitative score relevant to the measured construct (Parshall et al., 2010).

A special case of TE items, scenario based items (a.k.a. task based simulations, performance tasks), are integrated item set developed around a common scenario. The common scenario or each item relevant to the scenario may include a passage, a video clip, an animation, a graph, or a small simulation run by the test taker. Since scenario based items are generally developed to represent real life problems and tasks, they offer a potential for high fidelity in contributing to the validity of use and interpretation of test scores (Russell & Moncaleano, 2019; Sireci & Zenisky, 2006). Despite advantages scenario based items offer, there are also measurement challenges they may pose. Developing scenario based items is more challenging and expensive when compared to stand-alone items, as a result they tend to be fewer in item pools posing validity threat for repeated item exposure and memory effects (Bryant, 2017; Impara & Foster, 2006; Sireci & Zenisky, 2006). Furthermore, complex structure of scenario based items may require detailed consideration if partial credit scoring is required or what kind of scoring rule should be applied (Betts et al., 2022; Clyne, 2015; Lorié, 2016).

1.1. Technological Innovations in Trends in Mathematics and Science Study

TIMSS is an international assessment administered every four years starting from 1995 that measures mathematics and science achievements of fourth and eighth grade students. A transition from paper based assessment to digital assessment started in 2019 which is called eTIMSS 2019. Along with the digital transition, technological features were added to the items in the existing pool or new TE items were developed accordingly (Martin et al., 2020). Innovations in eTIMSS 2019 were also observed in new interactive item types called Problem Solving and Inquiry (PSI) tasks which were technology-enhanced scenario based items. By using PSI tasks, IEA aimed to extend the coverage and enhance the measurement of the TIMSS mathematics and science assessment frameworks benefitting from the features of computerized assessments, especially in the applying and reasoning cognitive domains. PSI tasks simulate real world and laboratory situations where students can apply and combine their content knowledge, skills, reasoning, and interpretation of a given situation by solving a mathematics problem, running a scientific experiment or running multiple steps of a simulation. PSI tasks involve visually attractive, interactive scenarios that require students follow a series of tasks or TE items with various response actions (e.g., number pad, drag and drop, graphing tools, and free drawings) in an adaptive and responsive way that would bring them toward a final solution or product (Mullis et al., 2021).

With increased fidelity in eTIMSS 2019 items and PSI tasks, a ruler and calculator were also available to the students at eighth grade as part of the on-screen interface. The on-screen calculator included the four basic functions (+, −, ×, ÷) and a square root key. Since a standardized ruler and calculator was available on the test screen, students were not allowed to bring their own rulers and calculators (Mullis & Martin, 2017).

1.2. Tool Use in Technology-Enhanced Items

Technological innovations of computerized items also include tools offered with the item such as magnifier, digital pen for highlighting or taking notes on a digital scratchpad, ruler, or a calculator. Some tools may be compulsory to use for the test taker to be able to correctly respond

to an item (Salles et al., 2020); they can also be available for all the items or test takers as universal tools across the entire test (WIDA, n.d.). Although examining how such tools contribute to the test taking experience in paper-pencil or classroom assessments has a long history of research, studies for digital tools in computerized assessments are limited.

Process data collected during the test administration now offer information regarding the students' use of tools. Process data may include information for which tool is used, frequencies of each tool using, or patterns of tool using to understand the test-taking strategies of students, to collect evidence for suspicious test-taking activities or to collect evidence for fairness issues. Analyzing process data regarding the use of digital tools can also contribute to item and test development processes as they provide clues for how to ease test-taking processes, eliminating construct-irrelevant variances and increasing the fidelity and validity of the item. For instance, Salles et al., (2020) showed that test takers who responded to a mathematical item with a graph correctly tended to use a digital pen for taking notes on the graph. Another study on computer based office simulation tests showed that successful test takers tended to use notepad and spreadsheets helping computation more efficiently (Ludwig & Rausch, 2022).

1.3. On-Screen Calculators as a Technological Innovation

A group of researchers who are against the use of calculators in mathematics classes before high school stated that calculation skills and understanding of mathematics concepts may be negatively affected by the use of calculators during learning (e.g., Dick, 1988; Hopkins, 1992; Plunkett, 1978). Another perspective of research indicates that calculators can ease the learning process as they still would require the students to improve their mental computation strategies anyway. Similarly, when the student is solving a problem and faces a complex calculation in the middle of a problem, the student's work flow does not need to be interrupted due to hand calculation (Sparrow et al., 1994; Vasquez & McCabe, 2002; Williams, 1987).

Parallel to the idea of using calculators during the teaching and the learning of mathematics, researchers debated the use of calculators during assessments as well. The National Council of Teachers of Mathematics (NCTM) states that when on-screen calculators are used appropriately, calculators can positively contribute to students' fluency in numbers, operations, and estimation skills (2015). According to early research findings conducted by Hopkins (1992), numbers in problems can be made more compatible with realistic situations, making the use of calculators more appropriate. Additionally, calculators can increase motivations in students' test taking (Ellington, 2003).

Test developers and other test score users should be aware that the frequency of calculator use may have an effect on students' performance in assessments (Tarr et al., 2000). Additionally, the availability of on-screen calculators should be determined depending on item types and complexity level of the items (Cohen & Kim, 1992; Loyd, 1991). For some item types, an on-screen calculator should not be provided depending on the construct being measured and for some items the calculators may not be needed. For instance, Walcott and Stickles (2012) conducted a study using eighth grade level NAEP data that included two types of items — problem solving items and noncomputational mathematics concept items— where they studied the effect of calculator use and item types. The results showed that students who used calculators had significantly better performance on problem solving items when compared to students who did not use calculators. On the other hand, calculators were not used by the majority of the students for noncomputational mathematics concept items and the ones used consistently performed worse on the test.

In summary, research shows that the calculator can improve students' fluency in numbers, operation and estimation skills that may contribute to the development of complex problem-solving and higher-order thinking skills as well as increase motivations in the students' test taking. Additionally, computerized assessments can control the calculator effect providing the

same on-screen calculators to all students suitable to the given item type and grade level. Yet, test developers should be aware that calculator use should not change the measured construct and therefore an on-screen calculator may only be available for specific items (Wolfe, 2010). Finally, further validity research is needed to examine the extent to which frequency of calculator use affects test scores to ensure equity across cultures or education systems and whether the on-screen calculator contributes to students' mathematics performance.

1.4. Purpose of the Study

As a digital tool, an on-screen calculator for mathematics items including PSI tasks in eTIMSS 2019 was available to the students. In PSI tasks, while there were two successive items administered that were essentially developed to be calculator neutral, calculators can also help problem solving process.

Preliminary analysis results of calculator use relevant to these items were reported in eTIMSS 2019 PSI report (Mullis et al., 2021). According to the report, around 88% and 84% of students used the on-screen calculator for the first and second items respectively among the students who answered the items correctly.

Therefore, preliminary findings imply that availability and use of calculators may be helpful for responding to TE items correctly; however, further research is needed to examine the extent to which use of on-screen calculators contributes to student responses above and beyond mathematics proficiency. If a significant contribution is observed, this finding will provide some evidence for item and test development endeavors in terms of making on-screen calculators available as part of innovations in TE items. To serve this purpose, the following research question was investigated: "To what extent does an on-screen calculator available for two TE items in eTIMSS 2019 PSI tasks explain eighth graders' probability of giving correct responses above and beyond their mathematics achievements?"

2. METHOD

2.1. Data Sources and Variables

The data of eTIMSS 2019 study conducted by the International Association for the Evaluation of Educational Achievement (IEA) was used in this study. This data are available to public use on IEA's website (Fishbein *et al*., 2021).

In the eighth grade level of eTIMSS 2019 mathematics item pool, there were a total of 208 stand-alone computerized items and 25 PSI items presented under three PSI tasks. There were a total of 16 booklets each of which was administered to a single student. The booklets 1-14 consisted of stand-alone eTIMSS items and booklets 15-16 contained PSI items. In each PSI booklet, there were a total of four tasks, two of them mathematics PSI tasks and other two were science tasks administered in two sessions where each session took around 45 minutes. The mathematics tasks were *Building*, *Robots*, *and Dinosaur Speed* of which the Building task was combined with *Robots* and presented/administered in a single session (Mullis *et al*., 2021). In the task of *Building,* one item was a multiple-choice item and the remaining eight of them were constructed-response items. Similarly, *Robot* included four constructed-response items and *Dinosaur* included one selected-response and sixteen constructed-response items.

In this study, two items ("Water Tank A" [MQ12B05A] and "Water Tank B" [MQ12B05B]) in *Building* task were studied, both were constructed-response items (Mullis *et al*., 2021, p. 110). Item response theory item parameters in *Building* tasks vary between 0.617 and 1.779 for discrimination, 0.467 and 2.084 for difficulty parameters. For "Water Tank A" and "Water Tank B" items, item parameters were 1.390 and 1.472 for discrimination, 0.771 and 0.816 for difficulty parameters respectively (Fishbein *et al*., 2021). Omit rates of items in *Building* also varied between 0.7% and 17.6%, the omitting rates for "Water Tank A" and "Water Tank B"

were 7.9% and 10.2% respectively. The omitted responses in *Building*, after excluding students who omitted all the items in the task, were recoded as "incorrect".

In this research, the study variables were student responses [incorrect(0)-correct(1)] to "Water Tank A" and "Water Tank B" items of *Building* task, a dichotomous variable showing whether the student used calculator or not during response generation for "Water Tank A" and "Water Tank B" items [not used (0)-used(1)] and first plausible values calculated for students' mathematics achievement across item pools scaled to a distribution with an international mean of 500 and a standard deviation of 100.

2.2. Sample

The TIMSS program employs a complex sampling method to increase the representation of the student population in each participated country. TIMSS uses stratified two-stage cluster random sample design in which a sample of schools drawn at first stage and one or more intact classes of students drawn from the sampled schools at second stage taking into account the stratification of schools depending on each participated countries' territorial-demographic characteristics (e.g., regions of the country, public-private schools, urban-rural areas). One apparent benefit of sampling the intact classes rather than individuals is easing the data collection process in terms of time and resources, and another benefit is that TIMSS pays particular attention to students' curricular and instructional experiences, and these are typically organized on a classroom basis (Martin, et al., 2020).

Students from each anticipated country and benchmarking participants were planned to include for this study first. However, there were students excluded from the analyses of this study. First, students who did not have access (not reached) to all the items in the given test were excluded. Similarly, students who did not answer all the items in the *Building* task were excluded. Students who were considered as noneffortful respondents were excluded from the analysis. Finally, some countries and benchmarking participants were excluded due to not having enough students in each cell of the levels of the variables (Table A in Appendix). Sample size of students who were administered *Building* task in eTIMSS 2019 cycle were presented in Table 1. How nonefforful respondents were decided are clarified in next paragraphs.

Table 1. *Number of eighth grade students included in analysis from each country.*

Country	Original Sample Size in TIMSS Dataset	Final Sample Size for Analysis
Chinese Taipei	665	644
Georgia	356	314
Hong Kong SAR	434	411
Hungary	588	572
Korea, Rep. of	503	479
Lithuania	453	445
Norway	446	402
Qatar	448	422
Russian Federation	423	408
Türkiye	523	513
United Arab Emirates	2792	2629
United States	1083	1043
Ontario, Canada	433	418
Quebec, Canada	368	356
Abu Dhabi, UAE	1060	973
Dubai, UAE	681	662
Total	11256	10691

Common approach for deciding noneffortful respondents is utilizing item response time information collected during testing. However, response times for each item were not available in eTIMSS2019 data, rather screen times were available. Unfortunate for the analysis, some screens contain several items. Therefore, response times for each screen were examined in this study.

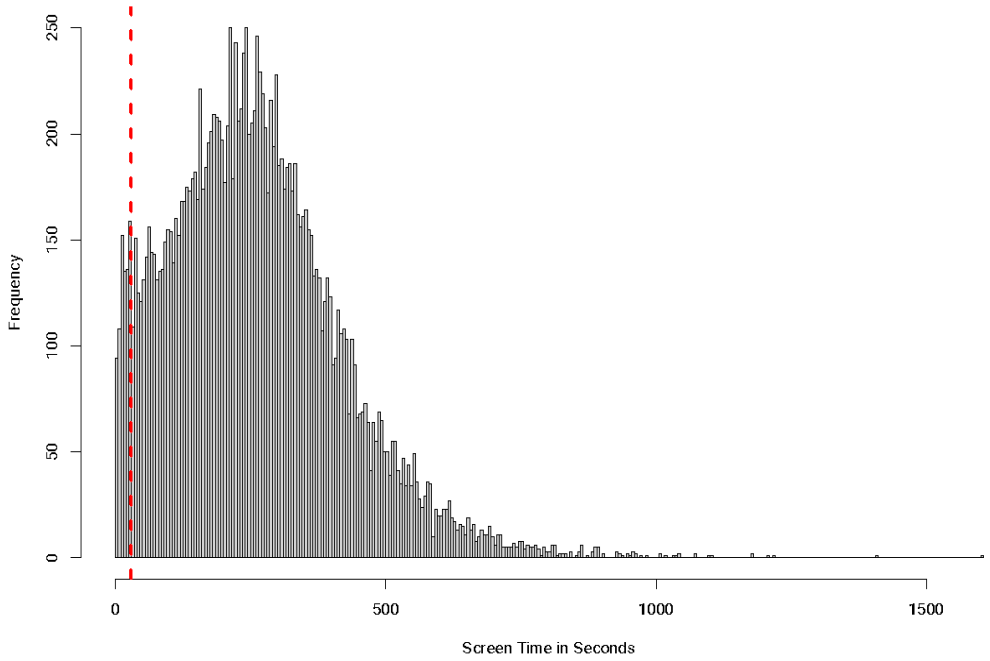
Table 2 shows screen time distributions for each screen consisting of items of the *Building* task. Screens completed less than a second implies noneffortful responding clearly. Previous researchers developed several methods to set a threshold of response time for filtering noneffortful respondents (e.g., Setzer et al., 2013; Ulitzsch et al., 2023; Wise, 2017; Wise & Gao, 2017). Among these methods, thresholds were set by using 3 or 5 seconds as common threshold across the items or calculating 10% of mean response time for an item with a maximum 10 seconds limitation (Wise et al., 2004; Wise & Ma, 2012). Though, the items in those studies were traditional item types (e.g., multiple-choice items) and response time distributions were available for each item.

Table 2. *Screen time distributions including items for "Building" task (in seconds).*

Screen	Min.	5th Quantile	25th Quantile	Mean	Median	75th Quantile	95th Quantile	Max.
Screen 2-Building Size	0.29	13.86	28.41	54.62	43.51	66.83	130.06	1194.93
Screen 3-Roof	0.21	13.20	37.70	71.59	59.12	91.44	165.94	771.99
Screen 4-Constructing the Walls	0.18	48.18	113.76	187.92	167.57	235.50	390.58	1550.70
Screen 5-Painting the Walls	0.29	31.28	98.39	170.10	152.46	220.39	365.54	1056.55
Screen 6-Water Tank	0.18	30.92	143.12	259.34	244.20	350.55	545.48	1600.26

The items in this study were part of a more complex problem solving task. There were three items on Screen 6, two of which were constructed-response items. Additionally, reading the instructions and the items in Screen 6 that require calculations can make the screen response time longer on average compared to other screens (Table 2). Considering 10 second-threshold in previous research contained relatively longer and complex items (Setzer et el., 2013), 30 seconds for a total of 3 items in Screen 6 were used as a threshold for screen response time. Screen time distribution given in Figure 1 also showed a "bump" on response time frequency occurred during the first 30 seconds that may be a sign of noneffortful responses of students (Schnipke, 1995). Therefore, students who spent less than 30 seconds on Screen 6 were excluded for eliminating noneffortful respondents.

Figure 1. *Screen time spent by the students showing the thresholds of 30 seconds.*



2.3. Data Analysis

The study has a cross-sectional design where strength of associations between dependent and independent variables were examined. Data analysis was conducted on R programming language environment (R Core Team, 2022, v.4.2.2) by modifying the relevant intsvy R package functions (Caro & Biecek, 2022, v.2.6).

2.3.1. *Sources of uncertainty and sampling variances*

The eTIMSS 2019 item pool contains 171 items with additional 29 PSI items in the fourth grade level item pool. Similarly, there were 206 items and 25 PSI items in the eighth grade level item pool. However, administering the entire item pool to each student would resulted in a burden of testing time. Instead, TIMSS uses matrix-sampling assessment design where each student is administered only a subset of items comparable through a common core of items. Based on the matrix-sampling approach, items were divided into 16 booklets where each item appeared in two booklets that allowed linking between booklets (Martin et al., 2017).

Matrix-sampling approach eases the testing process but it costs some variance and uncertainty in parameter estimates. One source of uncertainty is generalizing analysis results obtained from a student sample to the population of students called sampling variance, and second source of uncertainty is estimating students achievement scores from a sample of items called imputation variance (Foy & LaRoche, 2020).

2.3.1.1. Sampling Variance. The data were collected from national samples of students drawn once; therefore, how well the sample represents the target population is a crucial aspect of the analysis findings. As a result, sampling variance that also implies how well the sample represents the target population was computed and included during the analysis. The approach used for computing sampling variance in TIMSS 2019 was Jackknife Repeated Replication [(JRR), Foy & LaRoche, 2020].

2.3.1.2. Imputation Variance. In addition to sampling variance, as stated earlier, another variability is observed due to the fact that the student achievement is estimated by a subset of items instead of the entire item pool due to matrix-sampling assessment design. Students' achievement scores were generalized to the entire item pool by five plausible values (PV1-PV5)

computed by an imputation model. As a result, variation due to imputation procedures is observed in student achievement scores.

In summary, total variance in student achievement scores is obtained by summing JRR sampling variance and imputation variance; overall standard error for achievement estimations of each country is the square root of total variance computed for each country.

2.3.2. *Logistic Regression Models*

In order to answer the research questions, three binary logistic models were run in all of which the dependent variables P were probability of giving a correct response to the item. Models were given as follows:

$$\text{Model1: } log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_{PV1}$$

$$\text{Model2: } log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_{PV1} + \beta_{Calculator}$$

$$\text{Model3: } log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_{PV1} + \beta_{Calculator} + \beta_{PV1}{:}\beta_{Calculator}$$

In each of these models, $log\left(\frac{P}{1-P}\right)$ represents the natural logarithm of odds ratio (OR) of giving correct response, β_0 represents the intercept, β_{PV1} represents regression coefficient for first plausible value (PV1), and $\beta_{Calculator}$ represents the difference between reference group (non-calculator users) and focal group (calculator users) in the dependent variable.

In Model 1, PV1 was included as an independent variable only. In Model 2, a dichotomous variable that indicates the status of calculator use was added as another independent variable. In Model 3, two independent variables and their interaction effects were included. Since dependent variable was a dichotomous variable, estimated regression coefficients were associated to the change in log-odds of giving correct response with one-unit change in β_{PV1} and $\beta_{Calculator}$ and in their interaction term when controlling the effect of other independent variables.

For each logistic regression model, nested models were compared by chi-square difference tests (Tables 3-4). Additionally, McFadden R^2 as an approximation of the proportion of variance explained by independent variables (Smith & McKenna, 2013) and Akaike Information Criterion [(AIC); Bozdogan, 1987) were computed. These statistics were reported in Tables 3-4 and used for model comparison.

Ignoring the sampling and imputation variances while running logistic regression models can lead to bias in the estimation of standard errors and confidence intervals that may also cause incorrect interpretation of the results. Therefore in this study, total student weights (TOTWGT) and Jacknife replication values (JKZONE, JKREP) and first plausible values (PV1) were used to take into account the sampling variances and uncertainties.

3. RESULTS

Model comparison results for each fitted logistic regression models for “Water Tank A” and “Water Tank B” items were provided under M1-M3 columns where each represents model 1 through model 3 in Table 3 and Table 4. As seen in these tables, chi-square difference tests were examined and observed that for all the countries and benchmarking participants, model 2 had better model-data fit when compared to model 1. Similarly, adjusted McFadden R^2 values and AIC values showed that model 2 had a better fit with a higher proportion of explained variance and lower AIC values respectively when compared to model 1.

Next, model 2 and model 3 were compared. Accordingly, chi-square tests for “Water Tank A” showed that adding the interaction effect in model 3 provided a significant improvement for

Hong Kong SAR, Norway, Qatar, Türkiye, UAE, United States, Quebec-Canada and Abu Dhabi-UAE when compared to model 2 (α= 0.05). Similarly, chi-square tests for "Water Tank B" showed that adding the interaction effect in model 3 provided a significant improvement for Georgia, Hong Kong SAR, Republic of Korea, Norway, Qatar, Türkiye, UAE, United States, Ontario-Canada and Abu Dhabi-UAE when compared to model 2 (α= 0.05). AIC values were also lower for these countries specified above for both items, though adjusted McFadden R^2 values did not seem provide a larger proportion of variance explained in model 3 when compared to model 2. These findings suggest that using digital calculators are positively associated with higher odds of giving correct responses for both items above and beyond students` mathematics achievement scores conditional on students' mathematics achievement scores; however, odds-ratio coefficients vary across the status of calculator use for some of the countries.

Table 3. *Logistic regression model comparison statistics for "Water Tank A".*

Country	Chi-Square Test		Adjusted McFadden R^2			AIC		
	M1-M2	M2-M3	M1	M2	M3	M1	M2	M3
Chinese Taipei	< 0.001	0.058	0.31	0.37	0.37	630.71	573.48	572.54
Georgia	< 0.001	0.323	0.37	0.43	0.43	165.03	147.70	148.19
Hong Kong SAR	< 0.001	0.013	0.31	0.32	0.33	399.50	394.37	389.54
Hungary	< 0.001	0.896	0.34	0.45	0.44	606.10	508.64	510.57
Korea, Rep. of	< 0.001	0.793	0.37	0.45	0.45	404.80	352.94	354.83
Lithuania	< 0.001	0.643	0.33	0.36	0.36	392.67	374.38	376.28
Norway	< 0.001	0.027	0.23	0.30	0.31	445.82	408.63	402.50
Qatar	< 0.001	0.011	0.37	0.44	0.45	328.75	293.85	288.10
Russian Fed.	< 0.001	0.120	0.33	0.35	0.34	404.14	393.85	394.18
Türkiye	< 0.001	0.017	0.41	0.50	0.51	263.21	220.95	217.25
UAE	< 0.001	0.023	0.31	0.33	0.33	2335.08	2270.70	2263.37
United States	< 0.001	0.014	0.31	0.32	0.32	1097.31	1076.01	1074.34
Ontario, Canada	< 0.001	0.967	0.25	0.27	0.27	436.04	425.43	427.45
Quebec, Canada	< 0.001	0.047	0.19	0.21	0.22	390.93	378.93	376.92
Abu Dhabi, UAE	< 0.001	0.009	0.32	0.33	0.34	727.07	717.69	711.10
Dubai, UAE	< 0.001	0.716	0.28	0.31	0.30	671.04	651.40	653.47

Note1. UAE: United Arab Emirates.
Note2. ChiSquare difference test was evaluated at α=0.05 level.
Note3. Model 2 was adopted for Chinese Taipei, Georgia, Hungary, Republic of Korea, Lithuania, Russian Federation, Ontario-Canada, Dubai-UAE.
Note4. Model 3 was adopted for Hong Kong SAR, Norway, Qatar, Türkiye, UAE, United States, Quebec-Canada and Abu Dhabi-UAE.

To provide a further demonstration, how calculator use was associated with higher probability of giving correct responses conditional on students' mathematics achievement scores were presented with plots. As shown in Figure 2 and Figure 3, the group of students who used calculators in both items had higher probability of giving correct responses when compared to the group of students who did not use calculators having the same mathematics scores on average. The statistically significant interaction effects between calculator use and mathematics scores for the countries who were listed above can be observed in Figure 2 and Figure 3.

Table 4. *Logistic regression model comparison statistics for "Water Tank B".*

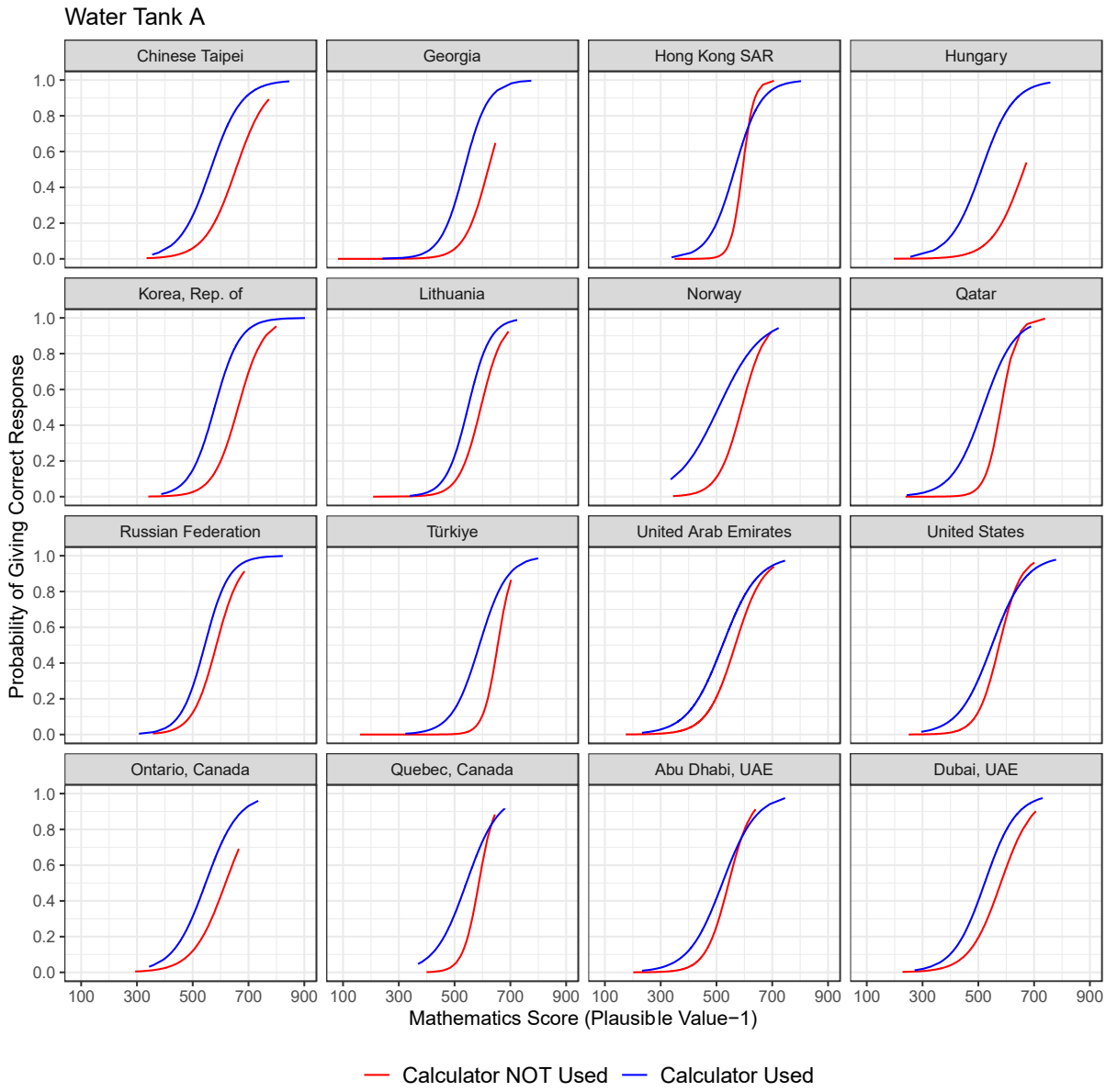
Country	Chi-Square Test		Adjusted McFadden R^2			AIC		
	M1-M2	M2-M3	M1	M2	M3	M1	M2	M3
Chinese Taipei	< 0.001	0.957	0.31	0.37	0.37	619.82	600.83	602.83
Georgia	0.020	0.003	0.37	0.43	0.43	154.22	151.35	141.30
Hong Kong SAR	0.003	0.041	0.31	0.32	0.33	417.55	410.71	407.57
Hungary	< 0.001	0.988	0.34	0.45	0.44	575.59	540.64	542.64
Korea, Rep. of	< 0.001	0.029	0.37	0.45	0.45	436.55	408.39	406.62
Lithuania	< 0.001	0.946	0.33	0.36	0.36	402.04	381.53	383.52
Norway	< 0.001	< 0.001	0.23	0.30	0.31	448.06	397.20	382.77
Qatar	0.001	0.021	0.37	0.44	0.45	310.68	299.97	293.40
Russian Fed.	0.002	0.693	0.33	0.35	0.34	365.27	355.74	357.65
Türkiye	0.004	0.012	0.41	0.50	0.51	240.30	234.91	230.84
UAE	< 0.001	< 0.001	0.31	0.33	0.33	2110.97	2060.75	2047.92
United States	< 0.001	< 0.001	0.31	0.32	0.32	1022.31	994.48	980.68
Ontario, Canada	0.001	0.033	0.25	0.27	0.27	472.51	461.64	457.90
Quebec, Canada	0.001	0.097	0.19	0.21	0.22	373.40	363.48	362.18
Abu Dhabi, UAE	0.001	< 0.001	0.32	0.33	0.34	693.35	683.52	673.70
Dubai, UAE	< 0.001	0.559	0.28	0.31	0.30	633.39	612.09	613.63

Note1. UAE: United Arab Emirates.
Note2. ChiSquare difference test was evaluated at α=0.05 level.
Note3. Model 2 was adopted for Chinese Taipei, Hungary, Lithuania, Russian Federation, Quebec-Canada, Dubai-UAE.
Note4. Model 3 was adopted for Georgia, Hong Kong SAR, Republic of Korea, Norway, Qatar, Türkiye, UAE, United States, Ontario-Canada and Abu Dhabi-UAE.

As seen in the Figures, for the participating countries and benchmarking that show a significant interaction effect, regression coefficients between student' mathematics scores and odds of giving correct response were not equal across the calculator users or non-users. Therefore, students who did not use the on-screen calculator and who had a score of 600 or higher had similar or even higher probabilities of giving the correct responses when compared to the ones who did not use the calculator. The authors note that the statistical coefficients are also a function of sample size and observed significant interaction effect may be due to relatively larger sample size in countries such as United Arab Emirates, United States or Abu Dhabi. Additionally, the prediction of probabilities for giving correct responses are limited to the range of the predictor data on x-axes.

How the findings of this study are consistent with the findings of the previous research were discussed in the next section. The impact of current study findings to the educational measurement literature and implications for the computerized item development were presented in the Discussion section.

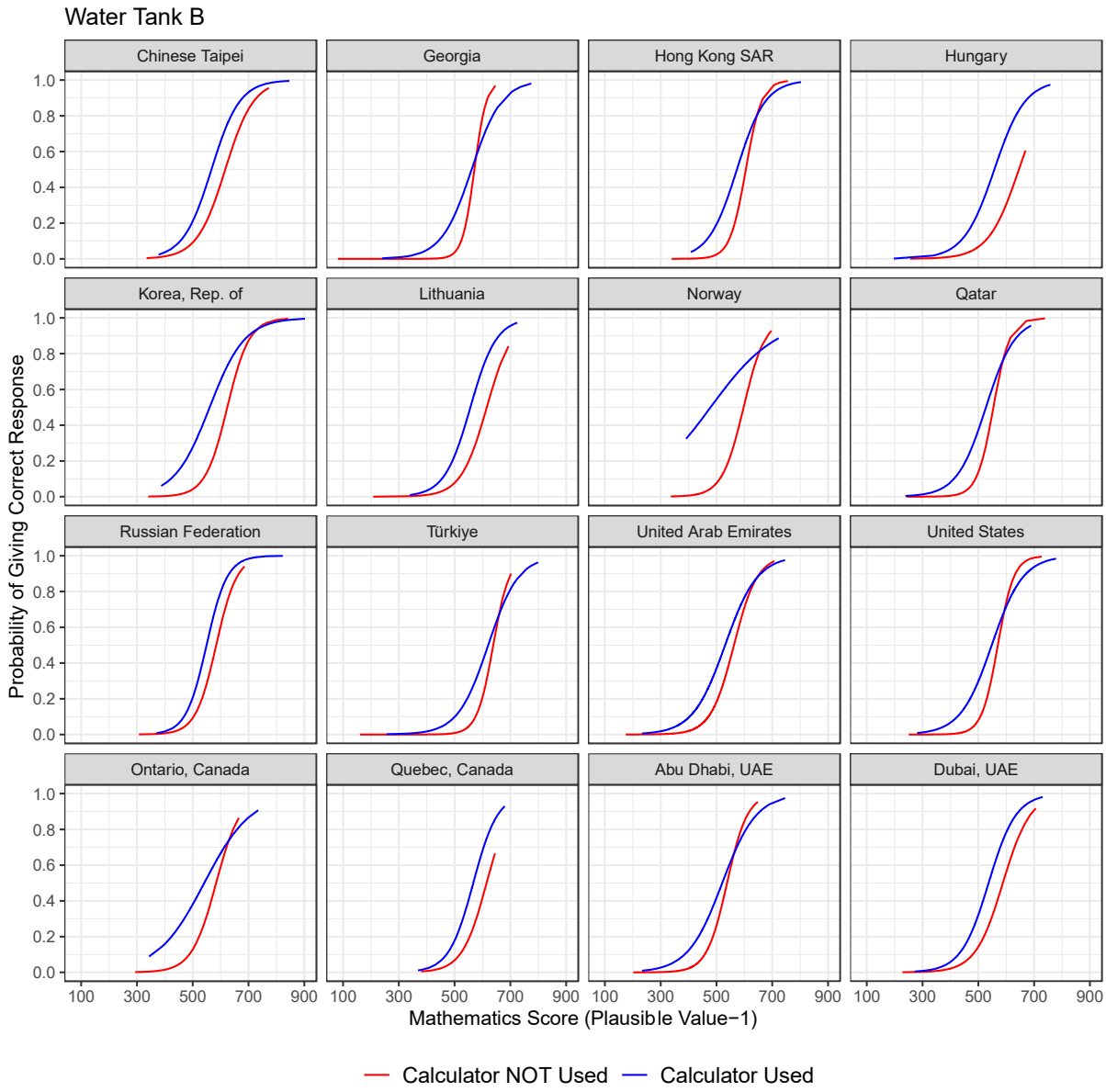
Figure 2. *Predicted probabilities of giving correct response to "Water Tank A" conditional on mathematics achievement and calculator use based on adopted logistic regression model.*



Note1. Model 2 was adopted for Chinese Taipei, Georgia, Hungary, Republic of Korea, Lithuania, Russian Federation, Ontario-Canada, Dubai-UAE.

Note2. Model 3 was adopted for Hong Kong SAR, Norway, Qatar, Türkiye, UAE, United States, Quebec-Canada and Abu Dhabi-UAE.

Figure 3. *Predicted probabilities of giving correct response to "Water Tank B" conditional on mathematics achievement and calculator use based on adopted logistic regression model.*



Note1. Model 2 was adopted for Chinese Taipei, Hungary, Lithuania, Russian Federation, Quebec-Canada, Dubai-UAE.
Note2. Model 3 was adopted for Georgia, Hong Kong SAR, Republic of Korea, Norway, Qatar, Türkiye, UAE, United States, Ontario-Canada and Abu Dhabi-UAE.

4. DISCUSSION and CONCLUSION

Discussions on the use of calculators have become a research topic in recent years at the point of designing it as a tool that can be used during learning and assessments, even as a digital tool that students can use on screen for computerized tests. Early research findings showed that calculator use can improve computational skills of students with average ability and have no adverse effects on the computational skills of the low and the high ability students (Brolin & Bjork, 1992: Hembree & Dessart, 1986; Hembree & Dessart, 1992.

Additionally, studies reveal that the use of calculators supports students during assessments. To solve a problem, the students must understand the problem, decide which problem-solving strategy is appropriate, carry out the strategy, and determine the solution. Therefore, calculators can contribute complex computing processes while students can spend more time on thinking

and developing a strategy (NCTM, 2015; Sparrow et al., 1994; Vasquez & McCabe, 2002). Previous studies in which large-scale assessment data were used showed that students who used calculators for mathematics problem solving items had significantly higher test scores than the students who did not use them (Mullis et al., 2021: Walcott & Stickles, 2012).

Current study results are parallel to the literature that promotes the use of calculators as a supportive tool during assessments. Current study findings showed that students who used the on-screen calculator had significantly higher probability of giving correct response above and beyond their mathematics achievements. As a result, it is suggested that the test and item developers should consider adding the on-screen calculator tool to the item as part of the innovations in TE items if test specifications and construct being measured allow. With the lights of the current study findings, more structured research is needed to collect further validity evidence regarding on-screen calculators.

The research findings also suggest that for some of the participating countries and benchmarkings, the interaction effect between student' mathematics scores and calculator use status was significant. This means that the odds of giving correct response were not equal across the calculator users or non-users in some of the countries. This observation may be related to the countries' education programs and students' familiarity and being used to the calculators in solving the mathematics problems. For instance, previous research indicated that the majority of the eighth grade students in participating European countries were allowed to use calculators approximately half or more than half of lessons to solve complex problems, do routine computations, and check answers (Eurydice, 2011). Considering the European students' potential familiarity with calculators, the proportion of students who answered the items correctly among the students who come from the European countries and did not use the calculators is extremely small is not surprising (Table A in Appendix). Similarly, even though Singapore is a high achieving country, the proportion of students who answered the items correctly is small among the students who did not use the calculator that may be related to the students' familiarity with using calculators starting from fifth grade (Koay, 2006; Mullis et al., 2016). Though, why a significant interaction effect was found in only some of the countries require further review and research.

This study is not without limitations. In this study, two items given under PSI tasks of TIMSS study were studied and the role of calculator use in other items in eTIMSS 2019 could not be studied due to the fact that such process data were available only for those two items in publicly available data. Yet, the findings of this study serve as preliminary findings and the content and context of the study can be expanded with more detailed process data regarding the use of calculators or other digital tools (e.g., ruler) with TIMSS data or any other TE items data.

There were 27 countries and benchmarking participants in eTIMSS 2019; however, data analysis was completed with students from only 16 countries and benchmarking participants in this study. The reason for this situation was that there were not enough students in each cell of the study variables (Table A in Appendix). Future research can examine if there are specific characteristics of these excluded countries that are relevant to using calculators during mathematics classrooms and assessments. Additionally, as prediction plots in Figure 2 and Figure 3 indicate, calculator use does not impact the probability of giving correct response at a fixed rate for some countries, rather high ability students may not need to use them as their problem solving processes. Therefore, future research can also examine what characteristics of their education system are associated with such findings for these countries that may inform the item and test development processes for country-specific assessments or cross-cultural assessments due to fairness.

Acknowledgments

The authors would like to thank the blind reviewers for their valuable feedback and suggestions. We also would like to thank to Emine Özdemir for the help in editing and proofreading the manuscript.

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.

Contribution of Authors

Ozge Ersan: Literature review, research design, data processing, data analysis, interpretations, writing and revising the manuscript. **Burcu Parlak:** Literature review, research design, interpretations, writing and revising the manuscript.

Orcid

Ozge Ersan https://orcid.org/0000-0003-0196-5472
Burcu Parlak https://orcid.org/ 0000-0001-7515-7262

REFERENCES

Betts, J., Muntean, W., Kim, D., & Kao, S. (2022). Evaluating different scoring methods for multiple response items providing partial credit. *Educational and Psychological Measurement*, *82*(1), 151–176. https://doi.org/10.1177/0013164421994636

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370. https://doi.org/10.1007/BF02294361

Brolin, H., & Bjork, L-E (1992). Introducing calculators in Swedish schools. In J.T. Fey & C. R. Hirsch (Eds.), *Calculators in mathematics educationi* (pp. 226–232). Reston, VA: National Council of Teachers of Mathematics.

Bryant, W. (2017). Developing a strategy for using technology-enhanced items in large-scale standardized tests. *Practical Assessment, Research & Evaluation*, *22*(1).

Caro D.H. & Biecek P. (2022). *intsvy: International Assessment Data Manager*. R package version 2.6. https://CRAN.R-project.org/package=intsvy

Clyne, C.M. (2015). *The effects of different scoring methodologies on item and test characteristics of technology-enhanced items* [unpublished doctoral dissertation]. University of Kansas, Lawrence, Kansas. https://kuscholarworks.ku.edu/bitstream/handle/1808/21675/Clyne_ku_0099D_14314_DATA_1.pdf?sequence=1

Cohen, A.S. & Kim, S. (1992). Detecting calculator effects on item performance. *Applied Measurement in Education*, *5*(4), 303-320. https://doi.org/10.1207/s15324818ame0504_2

Dick, T. (1988). The continuing calculator controversy. *Arithmetic Teacher,* 37–41.

Ellington, A.J. (2003). A meta-analysis of the effects of calculators on students' achievement and attitude levels in precollege mathematics classes. *Journal for Research in Mathematics Education*, *34*(5), 433–463. https://doi.org/10.2307/30034795

Eurydice (2011). *Mathematics education in Europe: common challenges and national policies.* http://keyconet.eun.org/c/document_library/get_file?uuid=e456b461-d3cd-4bd5-aabc-2cae2d4bfaf9&groupId=11028

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 User Guide for the International Database (2nd ed.).* TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation

of Educational Achievement (IEA). https://timssandpirls.bc.edu/timss2019/international-database

Foy, P., & LaRoche, S. (2020). Estimating standard errors in the TIMSS 2019 results. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 14.1–14.60). TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Hembree, R., & Dessart, D.J. (1986). Effects of hand-held calculators in precollege mathematics education: A meta-analysis. *Journal for Research in Mathematics Education*, *17*(2), 83–99. https://doi.org/10.2307/749255

Hembree, R., & Dessart, D.J. (1992). Research on calculators in mathematics education. In J.T. Fey & C.R. Hirsch (Eds.), *Calculators in mathematics education: 1992 NCTM Yearbook* (pp. 23–32). Reston, VA: The National Council of Teachers of Mathematics.

Hopkins, M.H. (1992). The use of calculators in assessment of mathematics. In T. Fey & C. R. Hirsch (Eds.), *Calculators in mathematics education: 1992 NCTMYearbook* (pp. 158–166). Reston, VA: The National Council of Teachers of Mathematics.

Impara, J.C., & Foster, D. (2006). Question and test development strategies to minimize test fraud. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp. 91–114). Lawrence Erlbaum Associates.

Koay, P.L. (2006). Calculator use in primary school mathematics: A Singapore perspective. *The Mathematics Educator*, *9*(2), 97-111.

Lorié, W. (2016). Automated scoring of multicomponent tasks. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (p. 627–658). IGI Global. https://doi.org/10.4018/978-1-4666-9441-5.ch024

Loyd, B. H. (1991). Mathematics test performance: The effects of item type and calculator use. *Applied Measurement in Education*, *4*(1), 11–22.

Ludwig, S., & Rausch, A. (2022). The relationship between problem-solving behaviour and performance – Analysing tool use and information retrieval in a computer-based office simulation. *Journal of Computer Assisted Learning*, 1-27. https://doi.org/10.1111/jcal.12770

Martin, M.O., Mullis, I.V.S., & Foy, P. (2017). TIMSS 2019 Assessment Design. In I.V.S. Mullis, & M.O. Martin (Eds.), *TIMSS 2019 Assessment Frameworks* (pp. 79–92). TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Martin, M.O., von Davier, M., & Mullis, I.V.S. (2020). *Methods and Procedures: TIMSS 2019 Technical Report.* TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA).

Mullis, I.V.S., Martin, M.O. (2017). *TIMSS 2019 Assessment Frameworks*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). https://timss2019.org/psi/

Mullis, I.V.S., Martin, M.O., Fishbein, B., Foy, P., & Moncaleano, S. (2021). *Findings from the TIMSS 2019 problem solving and inquiry tasks*. TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). https://timss2019.org/psi/

Mullis, I.V.S., Martin, M.O., Goh, S., & Cotter, K. (Eds.) (2016). *TIMSS 2015 Encyclopedia: Education Policy and Curriculum in Mathematics and Science.* Retrieved from Boston College, TIMSS & PIRLS International Study Center, Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). http://timssandpirls.bc.edu/timss2015/encyclopedia/

National Council of Teachers of Mathematics (2015). *Calculation Use in Elementary Grades.* https://www.nctm.org/Standards-and-Positions/Position-Statements/Calculator-Use-in-Elementary-Grades

Parshall, C.G., Harmes, J.C., Davey, T., & Pashley, P.J. (2010*).* Innovative items for computerized testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Elements of adaptive testing* (pp. 215–230). Springer. https://doi.org/10.1007/978-0-387-85461-8

Plunkett, S. (1978). Decomposition and all that rot. *Mathematics in Schools, 8*(3), 2–5.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Russell, M., & Moncaleano, S. (2019) Examining the use and construct fidelity of technology-enhanced items employed by K-12 testing programs. *Educational Assessment, 24*(4), 286–304. https://doi.org/10.1080/10627197.2019.1670055

Salles, F., Dos Santos, R., & Keskpaik, S. (2020). When didactics meet data science: process data analysis in large-scale mathematics assessment in France. *Large Scale Assessment in Education 8*(7). https://doi.org/10.1186/s40536-020-00085-y

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: a framework for constructing "intermediate constraint" questions and tasks for technology platforms. *Journal of Technology, Learning, and Assessment*, *4*(6).

Schnipke, D.L. (1995). *Assessing speededness in computer-based tests using item response times* [Unpublished doctoral dissertation]. Johns Hopkins University, Baltimore, MD.

Setzer, J.C., Wise, S.L., van den 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

Sireci, S.G. & Zenisky, A.L. (2006). Innovative item formats in computer-based testing: In pursuit of improved construct representation. In S.M. Downing & T.M. Haladyna (Eds.), *Handbook of test development* (p. 329–348). Lawrence Erlbaum Associates.

Smith, T.J., & McKenna, C.M. (2013). A comparison of logistic regression pseudo R^2 indices. *General Linear Model Journal, 39*(2), 17–26. http://www.glmj.org/archives/GLMJ_2014v39n2.html

Sparrow, L., Kershaw, L., & Jones, K. (1994). *Issues in primary mathematics education: calculators: research and curriculum implications.* Perth, Australia: Mathematics, Science & Technology Education Centre, Edith Cowan University.

Tarr, J.E., Uekawa, K., Mittag, K.C., & Lennex, L. (2000). A comparison of calculator use in eighth-grade mathematics classrooms in the United States, Japan, and Portugal: Results from the Third International Mathematics and Science Study. *School Science and Mathematics*, *100*(3), 139–150. https://doi.org/10.1111/j.1949-8594.2000.tb17249.x

Ulitzsch, E., Domingue, B.W., Kapoor, R., Kanopka, K. and Rios, J.A. (2023). A probabilistic filtering approach to non-effortful responding. *Educational Measurement: Issues and Practice*. Advanced online publication. https://doi.org/10.1111/emip.12567

Vasquez, S., & McCabe, T.W. (2002). The effect of calculator usage in the learning of basic skills. *Research and Teaching in Developmental Education*, *19*(1), 33–40.

Walcott, C., Stickles, P.R. (2012). Calculator Use on NAEP: A look at fourth- and eighth-grade mathematics achievement. *School Science and Mathematics, 112*(4), 241–254. https://doi.org/10.1111/j.1949-8594.2012.00140.x

WIDA (n.d.). (2023) *2022-2023 Accessibility & accommodations Manual.* https://wida.wisc.edu/sites/default/files/resource/Accessibility-Accommodations-Manual.pdf

Williams, D. (1987). Using calculators in assessing mathematics achievement. *Arithmetic Teacher, 34*(2), 21-23.

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

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., Kingsbury, G.G., Thomason, J., & Kong, X. (2004, April). *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., & Ma, L. (2012, April). *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, Vancouver, Canada.

Wolfe, E.W. (2010). What impact does calculator use have on test results? *Test, Measurement & Research Services Bulletin, 14,* 1–6.

APPENDIX

Table A. *Number of Students Depending on Their Use of Calculator and Giving Correct Responses.*

Country	Water Tank A			Water Tank B		
	Calc. Users	n	Number of Students Correctly Responded	Calc. Users	n	Props. of Students Correctly Responded
Chile	0	205	1	0	229	4
	1	173	60	1	149	51
Chinese Taipei	0	275	60	0	319	106
	1	369	253	1	325	229
England	0	135	1	0	178	6
	1	245	122	1	202	107
Finland	0	211	0	0	256	4
	1	311	102	1	266	107
France	0	130	0	0	175	3
	1	244	85	1	199	93
Georgia	0	203	11	0	220	15
	1	111	42	1	94	33
Hong Kong SAR	0	72	13	0	105	20
	1	339	217	1	306	193
Hungary	0	198	14	0	230	20
	1	374	257	1	342	205
Israel	0	160	2	0	206	9
	1	239	99	1	193	103
Italy	0	135	7	0	162	10
	1	255	112	1	228	101
Korea, Rep. of	0	216	36	0	235	60
	1	263	182	1	244	171
Lithuania	0	185	37	0	204	31
	1	260	128	1	241	120
Malaysia	0	197	6	0	201	13
	1	695	257	1	691	258
Norway	0	234	50	0	250	50
	1	168	101	1	152	95
Portugal	0	120	3	0	145	4
	1	256	102	1	231	89
Qatar	0	208	11	0	228	21
	1	214	85	1	194	74
Russian Federation	0	118	26	0	141	28
	1	290	172	1	267	164
Singapore	0	43	8	0	58	13
	1	579	462	1	564	469
Sweden	0	103	2	0	137	3
	1	247	118	1	213	95
Türkiye	0	364	13	0	385	22

	1	149	52	1	128	38
United Arab Emirates	0	1124	166	0	1335	198
	1	1505	664	1	1294	548
United States	0	254	31	0	314	38
	1	789	396	1	729	373
Ontario, Canada	0	97	18	0	111	22
	1	321	168	1	307	175
Quebec, Canada	0	65	13	0	89	14
	1	291	171	1	267	138
Moscow, Russian Fed.	0	71	6	0	94	9
	1	357	247	1	334	236
Abu Dhabi, UAE	0	488	65	0	564	80
	1	485	182	1	409	161
Dubai, UAE	0	197	43	0	244	54
	1	465	280	1	418	237

Note. 0: did not use calculator, 1: used calculator.