psychology international

Article

Examining Position Effects on Students' Ability and Test-Taking Speed in the TIMSS 2019 Problem-Solving and Inquiry Tasks: A Structural Equation Modeling Approach

Joyce Xinle Liu [1,*], Okan Bulut [2] and Matthew D. Johnson [3]

[1] Measurement, Evaluation, and Data Science, University of Alberta, Edmonton, AB T6G 2G5, Canada
[2] Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB T6G 2G5, Canada; bulut@ualberta.ca
[3] Department of Human Ecology, University of Alberta, Edmonton, AB T6G 2G5, Canada; matthew4@ualberta.ca
* Correspondence: xinle4@ualberta.ca

Abstract: Position effects occur when changes in item positions on a test impact the test outcomes (e.g., item parameters or test scores). Previous studies found that position effects may vary by the testing context and conditions and thus affect each test-taker differently. With the increasing adoption of digital assessments involving innovative item types that are focused on problem-solving skills, it is also essential to study position effects within this context. This study incorporates item-level scores and screen-level response time data from the Trends in International Mathematics and Science Study (TIMSS) 2019 problem-solving and inquiry tasks for Grade 4 students to examine position effects on students' ability and test-taking speed. This dataset included responses from 27,682 students from 36 countries. A structural equation modeling approach was employed to model ability and test-taking speed within the same model. The results indicated a small but statistically significant booklet effect on students' overall mathematics and science ability. The impact of block position changes within the booklets seemed to be greater than the impact of a reordering of subjects tested in the two sessions. The results also showed that when an item block was placed earlier in a test session, students spent more time on the items and performed better. The implications of these findings are discussed.

Keywords: booklet; position effect; problem-solving; large-scale assessment; structural equation modeling

check for updates

Citation: Liu, J.X.; Bulut, O.; Johnson, M.D. Examining Position Effects on Students' Ability and Test-Taking Speed in the TIMSS 2019 Problem-Solving and Inquiry Tasks: A Structural Equation Modeling Approach. *Psychol. Int.* **2024**, *6*, 492–508. https://doi.org/10.3390/psycholint6020030

Academic Editor: Clemens Draxler

Received: 10 February 2024
Revised: 14 March 2024
Accepted: 15 March 2024
Published: 2 April 2024

1. Introduction

Digital assessments are on the rise, with many countries around the world making the transition from paper-based to computer-based assessments for at least some of their school- or national-level examinations. International large-scale assessments in education, such as the Program for International Student Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), made the transition to a digital format in 2015 and 2019, respectively [1,2]. To fully take advantage of the digital platform, test developers usually incorporate new innovative item types (e.g., technology-rich items) in these assessments to enhance test-taking engagement and potentially improve the measurement quality of intended constructs. In addition, various types of process data are often captured in the background (e.g., item response times and event log data) to help uncover greater insights into students' test-taking process [3].

In eTIMSS 2019—the digital version of TIMSS 2019—in addition to the usual 14 student booklets that are included in the paper-based version of TIMSS, two additional booklets (Booklets 15 and 16) were developed comprising innovative problem-solving and inquiry (PSI) tasks. These tasks were designed around real-life scenarios and incorporated various interactive elements to engage the students and capture their responses [3]. In each of the two booklets, the tasks were identical but placed in different orders to counterbalance potential position effects on item statistics and achievement [4]. Upon analysis of data from the PSI tasks, Mullis et al. [3] noted that there were differences between students' completion rates for each block of tasks in the two booklets. For example, the completion rate was generally higher when a task was presented earlier in a test session. Further analysis revealed that among those students who did not complete all the items, a higher proportion of students stopped responding rather than running out of time on the test [3]. This finding suggests that items' positions on a test might have impacted students' use of time during the test, their test-taking motivation (or effort), and their performance.

Previous studies on position effects in large-scale assessments have mainly focused on its impact on item parameters, such as item difficulty, to address the concern of fairness (e.g., [5–9]). Several more recent studies have also examined how position effects could vary in different subject domains (e.g., [10,11]), for different item types (e.g., [11,12]), or given different student characteristics such as ability levels (e.g., [11,13]) or gender (e.g., [14]). Other studies have explored the relationship between position effect and test-taking effort (e.g., [15,16]) or the relationship between ability and speed, including potential applications of response time in measuring or predicting achievement [17–20]. However, only a few studies have examined the effects of item position on test-taking speed. Given the increasing adoption of digital assessments involving innovative item types, it is also essential to study position effects within this context. In this study, we make use of response data from the eTIMSS 2019 Grade 4 Mathematics and Science PSI tasks and examine the associations between block positions, students' test-taking speed, and their ability. Findings from this study could offer insight into the interplay of these variables in a computer-based test with technology-enhanced items and potentially help to inform future test development practices.

2. Theoretical Framework

In large-scale educational assessments such as PISA and TIMSS, booklet designs are typically used for test assembly and administration [21]. As such, each student is administered a particular booklet that contains a subset of all items that are used in the assessment, organized into item blocks. The same block of items usually appears in more than one booklet, so that items can be linked and calibrated on a common scale [8]. Item blocks are intentionally distributed so that the same item block will appear at different positions in different booklets. This approach helps enhance the test security [13] and counterbalance position effects on item statistics [21,22]. The eTIMSS 2019 PSI booklets used a similar counterbalancing booklet design, but in this case, there were only two booklets, each containing all five PSI items (see Table 1).

Table 1. eTIMSS 2019 PSI booklet design.

Booklet	Session 1		Session 2	
	Block Position 1	**Block Position 2**	**Block Position 3**	**Block Position 4**
Booklet 15	M1	M2	S1	S2
Booklet 16	S2	S1	M2	M1

Note: M1 and M2 are mathematics item blocks. S1 and S2 are science item blocks. There were 5 PSI tasks in total, 3 for mathematics (2 in M1, 1 in M2), and 2 for science (1 each in S1 and S2). Table adapted from [3].

Researchers have shown significant interest in item position effects, driven by the prevalent use of test designs where students encounter the same items at different points during the assessment. This phenomenon applies to booklet designs and computerized adaptive tests or multistage adaptive tests, where item and testlet positions cannot be fully controlled [6,23]. Numerous studies have explored how items' position influences item parameters, particularly item difficulty, employing various modeling approaches.

Researchers have often advocated for the review and potential removal of items displaying substantial position effects to enhance test fairness [6,23].

Generally, two types of position effects have been reported in the literature [24]: a positive position effect (i.e., when an item becomes easier when administered at later positions, see for example [10]) and, more frequently, a negative position effect (i.e., when an item becomes more difficult when administered at later positions, see for example [11]). Kingston and Dorans [23] and Ong et al. [12] found that the susceptibility to position effects appears to be item-type-specific. In particular, they found that longer items with higher reading demands were more susceptible to item position effects. Demirkol and Kelecioğlu [11] found stronger negative position effects in reading items compared to mathematics items using PISA 2015 data from Turkey. On the other hand, Hohensinn et al. [8] did not find any significant position effects for mathematical or quantitative items given unspeeded conditions (i.e., when sufficient time was given to complete all items). This supported Kingston and Dorans' [23] earlier findings and led the researchers to suggest that "position effects should be examined for every newly constructed assessment which deals with booklet designs" (p. 508). Debeer and Janssen [13] conducted an empirical study using PISA 2006 data and found that position effects could differ for individuals with different latent abilities (students with a higher ability tend to be less susceptible to position effects). Weirich et al.'s [16] study partly supported this finding and further demonstrated that changes in test-taking effort may also moderate position effects throughout a test.

In the context of eTIMSS 2019, Fishbein et al. [22] acknowledged the presence of position effects occurring in the PSI booklets, especially for mathematics. PSI item blocks appearing in the second half of a test session were more difficult and had more not-reached responses than item blocks appearing in the first half [22]. The actual completion rates for each task also varied based on block position [3]. These findings suggest that there could have been a booklet effect on students' overall achievement and their performance on individual items. In this case, the availability of response time data also presents a unique opportunity to examine the booklet effect on students' use of time during the test as an indicator of their test-taking speed.

Figure 1 shows a theoretical model demonstrating the relationship between items, booklets, and response times. The model defines two latent variables: ability, with item-level scores as its indicators, and speed, with screen-level response times as its indicators (item-level response times were not available for the PSI tasks in TIMSS 2019). Booklet is a binary variable in this context, and its effect on ability and speed will be examined. In the model, it is also possible to examine the booklet effect on ability and speed across individual items and screens throughout the test. This addition could offer greater insight, especially when viewed in conjunction with individual item characteristics.

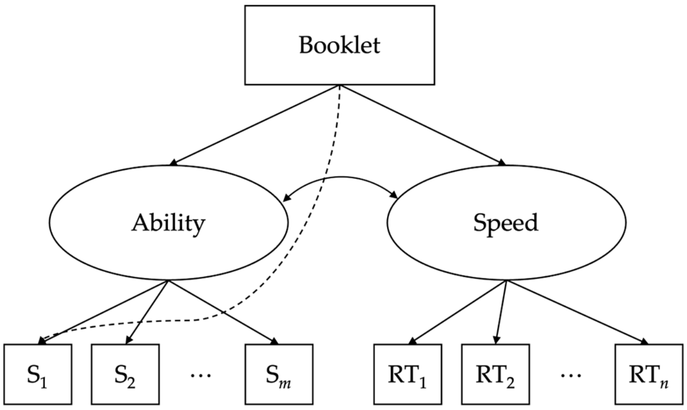

Figure 1. Theoretical model for examining booklet effects on ability and test-taking speed. S_i represents item-level scores, and RT_j represents screen-level response times.

Ability and speed are commonly associated with each other (e.g., [18,19,25,26]). There are generally two perspectives on the relationship between speed and ability. One perspective is that spending more time on an item (i.e., working more slowly) increases the probability of answering the item correctly, whereas speeding up reduces the expected response accuracy. This phenomenon is commonly referred to as the within-person "speed–ability trade-off" [19,27]. On the other hand, a person with stronger ability in a domain could exhibit faster speed due to greater skill and fluency [28]. Goldhammer [19] pointed out that most assessments are a mixture of speed and ability tests, as they typically have a time limit and include items of varying difficulty, so it can be very difficult to separate these measures. Goldhammer et al. [28] closely examined the relationship between the time spent on a task and task success using large-scale assessment data from the computer-based Programme for the International Assessment of Adult Competencies (PIAAC) and found that the time spent on task effect is moderated by the task difficulty and skill. Notably, the researchers found that task success is positively related to time spent on task for more difficult tasks, such as problem-solving, and negatively related to more routine or easier tasks. These findings suggest that the relationship between speed and ability is complex and could vary in different contexts. In Figure 1, the relationship between speed and ability is left as a correlation, as there is no theoretical basis to say that either one causes the other.

Position, ability, and speed have all been modeled in different ways through various studies that examined different combinations of these ideas. For speed, a well-known approach to model response times is the lognormal model introduced by van der Linden [29]. This model is based on item response theory (IRT) and has been extended in various ways to incorporate other variables, such as with a multivariate multilevel regression structure [30] and with structural equation modeling (SEM) [31]. For a detailed overview of modeling techniques involving response times, see De Boeck and Jeon's [32] recent review. For position effects, researchers often employed IRT-based methodologies such as Rasch or 2PL models, incorporating random or fixed position effects (e.g., [9,33]), or explanatory IRT approaches based on generalized linear mixed models (e.g., [8,11,12,16]). Bulut et al. [6] introduced a factor analytic approach using the SEM framework, which allows for the examination of linear position effects and interaction effects in the same model and provides added flexibility for assessments with more complex designs. In this study, an SEM approach was employed to allow us to model position, ability, and test-taking speed within the same model. Due to the way in which response times were captured (at the screen level rather than at the item level), it was not appropriate to use an IRT-based approach.

The following hypotheses, derived from a thorough literature review, can offer insights into the PSI tasks in TIMSS 2019. First, a negative correlation is anticipated between speed and ability, owing to the problem-solving nature of PSI tasks—implying that heightened speed may correspond to diminished ability. Second, a shift in booklet order from 15 to 16 is predicted to be associated with an elevation in science ability but a reduction in mathematics ability. This expectation arises from the alteration in the subject sequencing. Third, the impact of booklet changes is expected to manifest across all four item blocks, with a potentially heightened influence on items in blocks M1 and S2 due to the more substantial positional change between Block Position 1 and Block Position 4.

The current study aims to contribute to the existing literature in several ways. First, previous research examining position effects typically used item data from more traditional forms of assessment (e.g., multiple-choice items). In this study, position effects are studied in the context of a computer-based assessment with technology-rich items, which could offer valuable insights, especially as more PSI-type items are planned to be incorporated in future cycles of eTIMSS [34]. Second, few studies have incorporated response times into research on position effects (e.g., [35]). Since response times are routinely captured in digital assessments, tapping into this data source would add value to current discussions.

3. Methods

3.1. Data Source

This study used response data from the eTIMSS 2019 Grade 4 PSI booklets. eTIMSS, the digital version of TIMSS, was taken by students from 36 participating countries in 2019. PSI tasks were placed in Booklets 15 and 16 and administered to approximately 12% of all students who participated in eTIMSS 2019. In the eTIMSS 2019 administration, each student was randomly assigned one booklet to complete, followed by a 30 min questionnaire [36]. At the Grade 4 level, five PSI tasks (three mathematics and two science tasks, each comprising between six and twelve items) were grouped into two mathematics and two science blocks and presented to students in two separately timed sessions of 36 min each with a 15 min break in between (see Table 1) [3]. The two PSI booklets consisted of the same tasks and item blocks, arranged in different orders.

The Grade 4 PSI dataset included responses from 27,682 students from 36 countries. The students had a mean age of 10.14 years (*SD* = 0.57 years) and were evenly split between males (50.6%) and females (49.4%). Half (50%; 13,829) of the students completed Booklet 15, and the rest completed Booklet 16. The two booklets were similar regarding students' demographic characteristics (see Table 2), which is expected, given that all booklets that were used in eTIMSS were administered according to a rotated design [22]. A separate check was carried out to confirm that the representation by country was also similar across the two booklets.

Table 2. Demographic summary of students across two PSI booklets.

Booklet	*N* of Students	*N* of Countries	Age	Gender
Booklet 15	13,829	36	*M*: 10.15 y *SD*: 0.56 y	F: 49.9% M: 50.1%
Booklet 16	13,853	36	*M*: 10.14 y *SD*: 0.57 y	F: 48.8% M: 51.2%

Note: *M*: mean; *SD*: Standard Deviation; F: female; M: male.

3.2. Measures

Two sets of measures were derived from the PSI dataset: one for scores and another for response times on each of the 5 PSI tasks. The TIMSS International Database [4] contained students' responses to all the individual PSI items, coded as fully correct, partially correct, incorrect, not reached, or omitted/invalid. For this study, all items were scored using the same methodology as that used by TIMSS for achievement scaling. Omitted items were given a score of zero, and not-reached items were treated as missing. Furthermore, some items were excluded from the data, as not all PSI items were included in achievement scaling for TIMSS (e.g., items exhibiting poor psychometric properties and science items with post-clue scores [22]). These data preparation procedures yielded a total of 29 mathematics and 18 science PSI items. Table 3 shows the complete list of items and the maximum possible score for each item.

Response times for each task were derived using screen times captured in the original dataset. Screen time refers to the total time a student spends on a particular screen, and each screen could contain between one and three items. There was a total of 17 screens containing mathematics items and 17 screens containing science items (see Table 4). It was observed in the data that some students spent a disproportionate amount of time on specific screens, which could have resulted from disengaged behavior (i.e., the student stopped responding midway through the test) or from early completion of the test and staying on the same screen until the test ended. The screen time would not accurately represent the student's speed in these cases. Thus, it was necessary to determine a reasonable threshold to remove outliers from the data. As the number of items on each screen was not the same, and items may vary in difficulty and demand, the outlier threshold for each screen should not be the same.

In this study, the transformation approach suggested by Cousineau and Chartier [37] was adopted to identify response time outliers for each screen. This method was found to work well for response time data, yielding low bias in the data cleaning process [38]. To identify the outliers (i.e., responses with very high or low response times), the following transformation was first applied to the response times for each screen:

$$y = \sqrt{\frac{x - X_{min}}{X_{max} - X_{min}}} \tag{1}$$

where x is the untransformed response time, X_{min} is the minimum response time (out of all students) on a given screen, and X_{max} is the maximum response time on that screen. This transformation normalizes the data and bounds the data into the range of [0, 1]. Following this step, z scores were computed. In this study, screen response times associated with a z-score that was either larger than 3 or smaller than -3 were removed. This application of Cousineau and Chartier's [37] method removed between 0.4% and 1.7% of the response time data for each screen.

3.3. Data Analysis

This study followed an SEM approach to examine booklet effects on students' ability and speed in the context of a PSI assessment. Descriptive and correlation analyses were first conducted on all the observed variables (booklet, 47 score indicators, and 34 speed indicators) to check that distributional assumptions were met and that there were no multicollinearity issues. For the SEM analysis, ability indicators (item scores) were treated as categorical (ordinal) variables due to the way in which they were scored (i.e., correct, partially correct, incorrect). In contrast, speed indicators (screen response times) were treated as continuous variables. All analyses were conducted using Mplus 8.10 [39]. The weighted least-squares mean- and variance-adjusted (WLSMV) estimator was used to handle the categorical indicators for ability, and the rest of the model was estimated using the default maximum likelihood (ML) estimator.

The theoretical SEM model shown In Figure 1 was first fitted to the data and assessed for model fit. Alternative models were tested for global and local fit before arriving at the final structural regression models. The final models were analyzed in two stages. In the first stage, model parameters were estimated without the dashed paths from the booklet variable to the individual items or screens. In the second stage, these paths were added to examine the booklet effect on individual items and screens. To evaluate the model fit, aside from the chi-square test, the following indices were used: root-mean-square error of approximation (RMSEA), comparative fit index (CFI), Tucker–Lewis Index (TLI), and standardized root-mean-square residual (SRMR). The cutoff values suggested by Hu and Bentler [40] were referenced, namely, CFI and TLI greater than 0.95, RMSEA smaller than 0.06, and SRMR smaller than 0.08 indicate a relatively good fit for models analyzed using ML. Xia and Yang [41] cautioned against using a universal set of cutoff values for analyses conducted with ordered categorical variables. In particular, they noted that fit indices under WLSMV estimation tended to show better model–data fit compared to ML fit indices for the same misspecified model. Hence, in this study, the suggested cutoff values were used to diagnose the model fit, but not to serve as the sole justification for the acceptance of a model.

Regarding missing data, the proportion of missing scores across the PSI items ranged from 0.7% to 19.4%, while the proportion of missing response times ranged from 2.2% to 15.2%. Greater missingness typically occurred in the last few items of a task due to running out of time (see Table 3). For the speed part of the model, missing data were handled through the full-information ML estimation in Mplus, which estimates model parameters directly from available data without deleting cases or imputing missing values [42]. For the ability part of the model, missing data were handled using pairwise deletion through the WLSMV estimator [39].

Table 3. Descriptive statistics for ability indicators (mathematics and science item scores).

Block	Task	Item	*M*	*SD*	Max Score	% Data Present
M1	Penguins	MA01	0.29	0.45	1	99.3%
		MA02A	0.63	0.48	1	99.2%
		MA02B	0.44	0.50	1	99.0%
		MA02C	0.55	0.50	1	99.0%
		MA03A	0.41	0.49	1	98.9%
		MA03B	0.55	0.50	1	98.7%
		MA04A	0.57	0.80	2	98.6%
		MA04B	0.40	0.49	1	98.2%
		MA05A	0.33	0.47	1	98.0%
		MA05B	0.55	0.50	1	97.5%
		MA06A	0.49	0.50	1	97.0%
		MA06B	0.20	0.40	1	96.5%
	Robots-4	MR01A	0.69	0.46	1	96.0%
		MR01B	0.60	0.49	1	95.1%
		MR02A	0.28	0.45	1	94.8%
		MR02B	0.44	0.50	1	88.2%
		MR03	0.32	0.46	1	87.6%
		MR04	0.56	0.84	2	83.8%
M2	School Party	MP01A	0.38	0.49	1	97.1%
		MP01B	0.43	0.71	2	96.1%
		MP02	0.46	0.50	1	95.6%
		MP03	1.26	0.90	2	95.2%
		MP04	0.39	0.75	2	94.5%
		MP05A	0.62	0.48	1	92.0%
		MP05B	0.13	0.34	1	91.2%
		MP06A	0.13	0.34	1	89.0%
		MP06B	0.21	0.41	1	87.5%
		MP07A	0.14	0.35	1	84.0%
		MP07B	0.10	0.30	1	80.6%
S1	Farm Investigation	SF01	0.70	0.84	2	97.6%
		SF02	0.44	0.50	1	95.6%
		SF03	0.53	0.50	1	94.5%
		SF04	0.51	0.50	1	92.7%
		SF05	0.51	0.50	1	90.0%
		SF06	0.61	0.49	1	88.3%
		SF07A	0.58	0.49	1	86.2%
		SF07B	0.11	0.32	1	86.9%
		SF08	0.64	0.48	1	85.4%
		SF09	0.72	0.79	2	84.1%
S2	Sugar Experiment	SS01	0.56	0.50	1	99.3%
		SS02	0.41	0.49	1	99.2%
		SS03	0.66	0.88	2	98.9%
		SS04	0.73	0.80	2	96.6%
		SS05	0.68	0.74	2	96.0%
		SS07	0.82	0.74	2	91.2%
		SS08	0.39	0.49	1	90.8%
		SS09	0.47	0.50	1	88.4%

Note: The first letter of the item name refers to the subject (M—mathematics; S—science). The second letter is an abbreviation of the task name. If an item has multiple parts, e.g., A, B, C, it means that they appeared on the same screen. *M*: mean; *SD*: Standard Deviation.

Table 4. Descriptive statistics for speed indicators (response time in seconds for each screen).

Block	Task	Item	*M*	*SD*	Min	Max	% Data Present	% Outliers Removed
M1	Penguins	MA01_S	56.20	33.38	0.26	207.86	97.6%	1.4%
		MA02_S	102.92	50.02	3.54	326.07	97.2%	1.7%
		MA03_S	67.86	36.39	0.72	235.88	97.3%	1.4%
		MA04_S	119.52	72.04	0.12	443.71	97.5%	0.9%
		MA05_S	115.66	69.22	0.09	434.35	96.9%	1.1%
		MA06_S	102.87	70.69	0.14	419.68	96.2%	1.1%
	Robots-4	MR01_S	76.39	37.51	2.57	242.22	94.8%	1.4%
		MR01_S	133.97	82.72	0.13	490.76	94.8%	0.6%
		MR02_S	103.25	62.71	0.10	384.53	93.0%	0.7%
		MR02_S	164.79	87.76	1.18	555.92	90.7%	1.2%
M2	School Party	MP01_S	134.21	81.35	0.16	499.81	95.8%	1.0%
		MP02_S	36.56	22.69	0.09	144.95	94.2%	1.3%
		MP03_S	52.44	31.75	0.09	202.00	93.6%	1.4%
		MP04_S	138.21	84.34	0.18	516.02	93.2%	1.0%
		MP05_S	106.83	59.96	0.67	372.77	90.9%	1.3%
		MP06_S	136.99	88.46	0.11	531.50	89.8%	0.4%
		MP07_S	121.48	80.91	0.13	482.05	86.8%	0.5%
S1	Farm Investigation	SF01_S	117.76	68.37	0.50	421.01	96.1%	1.3%
		SF02_S	109.72	64.82	0.22	396.96	95.3%	1.0%
		SF03_S	113.35	52.20	5.19	340.71	93.3%	1.7%
		SF04_S	106.81	49.84	4.19	324.96	91.8%	1.4%
		SF05_S	66.16	41.68	0.07	251.27	89.6%	1.0%
		SF06_S	29.61	14.46	0.94	93.53	87.8%	1.1%
		SF07_S	100.66	61.59	0.08	374.66	87.7%	0.5%
		SF08_S	35.50	18.81	0.38	119.59	85.5%	0.8%
		SF09_S	90.91	58.13	0.08	349.53	84.8%	0.6%
S2	Sugar Experiment	SS01_S	102.39	69.85	0.09	421.17	97.8%	1.3%
		SS02_S	92.10	57.67	0.13	348.79	97.8%	1.0%
		SS03_S	156.22	94.97	0.17	577.84	97.5%	1.0%
		SS04_S	143.47	86.48	0.10	527.48	96.9%	0.7%
		SS05_S	98.37	65.93	0.10	399.17	95.8%	0.8%
		SS07_S	113.12	64.87	0.22	400.78	93.8%	0.6%
		SS08_S	59.79	40.42	0.06	241.21	92.1%	0.6%
		SS09_S	38.93	28.87	0.07	174.02	90.7%	1.0%

Note: The first letter of the item name refers to the subject (M—mathematics; S—science). The second letter is an abbreviation of the task name. *M*: mean; *SD*: Standard Deviation.

4. Results

4.1. Descriptive and Correlation Analyses

A preliminary data screening was conducted on all observed variables to check that the assumptions of the SEM had been met. Descriptive statistics for the ability and speed indicators are presented in Tables 3 and 4, respectively. For the part of the model that was estimated using the default ML method, it was important to screen the data for multivariate normality. Here, we adopted the approach suggested by Kline [42] to assess the quantitative measures of skewness and kurtosis in the observed variables. After removing outliers, the distribution of each screen response time variable was found to be approximately normal, with skewness and kurtosis values below 2 and 7, respectively [42]; hence, further transformation was not necessary.

Bivariate correlations were computed for each pair of observed variables. The correlations between item score variables generally ranged from $r = 0.1$ to $r = 0.3$. The correlations between screen response time variables appeared to vary distinctly by PSI task, with the screen response times between some pairs of tasks correlating more strongly than others. The response times for screens of the same task were more closely related (generally ranging from $r = 0.2$ to $r = 0.4$) than those for different tasks. The correlations between item scores and screen response times were mainly close to 0. Overall, the maximum absolute correlation between any two variables was 0.51, and the variance inflation factor for all observed variables ranged between 1.1 and 2.2 (less than 10), which indicated that multicollinearity would not be a concern [43].

4.2. Model Specification and Fit

As a first step in the analysis, the initial theoretical model in Figure 1 was fitted to the data to assess the model fit. It was necessary to rescale the screen response time variables to units of minutes instead of seconds to prevent the problem of an ill-scaled covariance matrix and allow the model to converge [42]. The conceptual structural regression model demonstrated a poor fit to the data. To identify the source of the misfit, the measurement components of the structural regression model were analyzed separately before combining with the booklet variable. Table 5 shows the model fit indices of models tested in this study. It was noted that all tested models yielded a significant exact-fit (χ^2) test, which could be due to the large sample size in this study.

Table 5. Model fit indices for different measurement models and structural regression models.

Model	χ^2	*df*	*p*	RMSEA [90% CI]	CFI	TLI	SRMR
Measurement Model							
1-factor CFA (ability)	43,785.064	1034	<.001	0.039 [0.038, 0.039]	0.931	0.927	0.049
2-factor CFA (ability—math and science)	31,660.815	1033	<.001	0.033 [0.032, 0.033]	0.950	0.948	0.042
1-factor CFA (speed)	97,587.490	527	<.001	0.082 [0.081, 0.082]	0.462	0.427	0.113
2-factor CFA (speed—math and science)	89,237.944	526	<.001	0.078 [0.078, 0.079]	0.508	0.475	0.115
5-factor CFA (speed—5 tasks)	22,238.995	517	<.001	0.039 [0.039, 0.039]	0.880	0.869	0.041
5-factor CFA (speed—5 tasks, modified)	20,438.493	516	<.001	0.037 [0.037, 0.038]	0.890	0.880	0.040
Structural Model							
Original theoretical model	285,358.551	3237	<.001	0.056 [0.056, 0.056]	0.682	0.674	0.094
Booklet on 2-factor CFA (ability)	33,827.461	1078	<.001	0.033 [0.033, 0.033]	0.948	0.945	0.043
Booklet on 5-factor CFA (speed)	22,557.918	545	<.001	0.038 [0.038, 0.039]	0.892	0.882	0.039

Note: RMSEA: root-mean-square error of approximation; CFI: comparative fit index; TLI: Tucker–Lewis Index; SRMR: standardized root-mean-square residual.

For the ability component, a one-factor model with a single ability construct, indicated by all items on the test, was found to fit the data well. However, as TIMSS typically treats mathematics and science achievement as two separate constructs and reports these results separately, it is more appropriate to reflect this in the model using two separate latent constructs. The two-factor ability model showed a good fit to the data, as indicated by the global fit indices. As Mplus does not display standardized or normalized residuals for analyses conducted using the WLSMV estimator, a reference was made to modification indices for an indication of the local fit. No specific inter-item error correlations were suggested, which would result in a significant improvement in χ^2. Thus, the measurement model for ability was retained as such, which also fits well with the theory. TIMSS uses item response theory for achievement scaling, which assumes local independence of item responses given ability. Since the items used in this study were those included in the eTIMSS achievement scaling, they can be assumed to be high-quality items, and there is thus no basis for correlating errors between any pair of items.

For the speed component, it was found that a one-factor model with a single speed construct that was indicated by all screen response times on the test did not fit the data well. It is conceivable that the response time patterns for mathematics items could differ from those for science items, and thus, a two-factor model was also tested. However, the model fit was still poor (see Table 5). The earlier correlation analysis suggested that the response time patterns could be task specific. A five-factor model with separate latent speed variables (one for each task), each indicated by the screen response times for the specific tasks, yielded a substantial improvement in the global model fit. The absolute fit indices (RMSEA and SRMR) indicated a good fit, although the relative fit indices (CFI and TLI) still indicated an insufficient model fit. An inspection of the normalized residuals showed that the local misfit was scattered throughout the model rather than just confined to several pairs of observed variables (this is discussed more in the next section). One possible explanation is that response time data are inherently prone to fluctuations and are difficult to capture accurately in a way that truly represents a student's test-taking speed. One modification was made to the model by correlating the errors on two specific screens (SF01_S and SF02_S), as it was observed that the normalized residual between these two screens was substantively larger than others. A close review of the specific items on these two screens (available in [3]) revealed that the items were more open-ended, with very similar wording and structure, suggesting that the unique variances of these screen response times could be related. Adding this modification improved the model fit slightly, and no other modifications were made, as they would not be theoretically justifiable.

The original theoretical model (see Figure 1) required combining the ability and speed models and assumed that ability and speed could be uniquely measured by item scores and screen response times, respectively (i.e., no cross-loadings). The five-factor speed model yielded the best possible model fit in this study. While not ideal, it could give a reasonable representation of test-taking speed for the purpose of this study. Other attempts to form a single latent speed variable (e.g., using a higher-order latent variable to draw shared variance from the five tasks or specifying indicators at the task level instead of at the screen level) also did not yield a sufficient model fit. Due to the challenges of modeling speed in this context, a combined ability and speed model was not feasible and would not fit the data well. Hence, subsequent analyses of the booklet effect were carried out separately for the ability and speed models.

4.3. Booklet Effect on Ability

The final structural regression model for the booklet effect on ability is shown in Figure 2. The parameter estimates for the overall model are reported in Table 6. The model was also re-run once for each item on the test (including the dashed path) to examine the booklet effect on each item throughout the test. The parameter estimates for the dashed paths are reported in Table 7. Due to the large sample size in this study, most of the parameter estimates were statistically significant. Hence, it was essential to consider the effect size. For the factor loadings, the average variance that was extracted (average of all squared standardized loadings) for the mathematics ability factor was 0.40, and it was 0.30 for the science ability factor. These results showed that the item score variables were good indicators of their factors (based on criteria from [44], who also noted that factor loadings for categorical indicators tend to be lower).

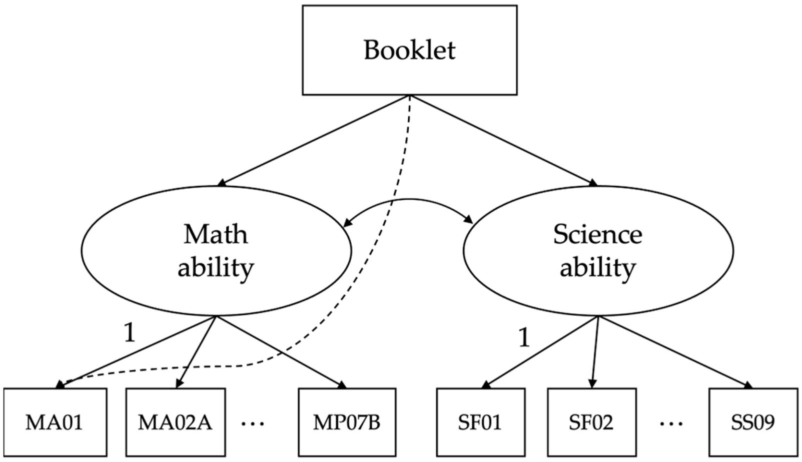

Figure 2. Final structural regression model for booklet effect on mathematics and science ability. Model fit indices: χ^2 (1078) = 33,827.461; p < .001; RMSEA = 0.033; CFI = 0.948; TLI = 0.945; SRMR = 0.040.

Table 6. WLSMV estimates for the structural regression model of the booklet effect in the mathematics (M) and science (S) tasks.

Parameter	Unstandardized		Standardized		Parameter	Unstandardized		Standardized	
	Estimate	SE	Estimate	SE		Estimate	SE	Estimate	SE
Factor loadings					**Factor loadings**				
MA01	1.000	-	0.454	0.007	SF01	1.000	-	0.500	0.006
MA02A	1.469	0.027	0.667	0.006	SF02	1.253	0.019	0.627	0.006
MA02B	1.337	0.025	0.608	0.006	SF03	1.211	0.020	0.606	0.007
MA02C	1.548	0.027	0.703	0.005	SF04	1.221	0.020	0.611	0.007
MA03A	0.988	0.022	0.449	0.007	SF05	1.075	0.019	0.537	0.007
MA03B	1.551	0.028	0.705	0.005	SF06	0.807	0.019	0.403	0.008
MA04A	1.396	0.025	0.634	0.005	SF07A	0.777	0.019	0.389	0.009
MA04B	1.507	0.026	0.685	0.005	SF07B	0.920	0.022	0.460	0.009
MA05A	0.883	0.021	0.401	0.008	SF08	0.916	0.020	0.458	0.009
MA05B	1.515	0.027	0.688	0.005	SF09	1.225	0.019	0.613	0.006
MA06A	1.723	0.029	0.783	0.005	SS01	1.311	0.020	0.656	0.006
MA06B	1.649	0.028	0.749	0.005	SS02	1.151	0.019	0.575	0.006
MR01A	1.090	0.024	0.496	0.007	SS03	1.129	0.019	0.565	0.006
MR01B	1.321	0.026	0.600	0.006	SS04	1.190	0.019	0.595	0.006
MR02A	1.569	0.027	0.713	0.005	SS05	1.192	0.019	0.596	0.006
MR02B	1.523	0.027	0.692	0.005	SS07	0.992	0.018	0.496	0.006
MR03	1.641	0.028	0.745	0.005	SS08	0.914	0.019	0.457	0.008
MR04	1.362	0.025	0.619	0.006	SS09	1.100	0.020	0.550	0.007
MP01A	1.500	0.026	0.682	0.005					
MP01B	1.082	0.022	0.492	0.007	**Direct effects on ability**				
MP02	1.220	0.024	0.555	0.006	Booklet → M	−0.044	0.006	−0.049	0.006
MP03	1.070	0.023	0.486	0.007	Booklet → S	−0.040	0.007	−0.040	0.007
MP04	1.592	0.027	0.723	0.005					
MP05A	1.606	0.029	0.730	0.006					
MP05B	1.275	0.027	0.579	0.008					
MP06A	1.260	0.027	0.573	0.008					
MP06B	1.477	0.027	0.671	0.006					
MP07A	1.434	0.027	0.652	0.007					
MP07B	1.468	0.029	0.667	0.008					

Note: $p < .001$ for all unstandardized estimates with standard errors. Model fit indices: χ^2 (1078) = 33,827.461; $p < .001$; RMSEA = 0.033; CFI = 0.948; TLI = 0.945; SRMR = 0.040.

Table 7. WLSMV estimates for the structural regression model of the booklet effect on performance in individual mathematics (M) and science (S) tasks.

Item	Unstandardized		Standardized		Item	Unstandardized		Standardized	
	Estimate	SE	Estimate	SE		Estimate	SE	Estimate	SE
MA01	0.046	0.015	0.023	0.008	MP05B	0.201	0.019	0.100	0.009
MA02A	−0.058	0.014	−0.029	0.007	MP06A	0.224	0.019	0.111	0.010
MA02B	−0.151	0.014	−0.075	0.007	MP06B	0.281	0.017	0.140	0.008
MA02C	−0.053	0.013	−0.026	0.006	MP07A	0.196	0.019	0.098	0.010
MA03A	−0.060	0.015	−0.030	0.007	MP07B	0.166	0.022	0.083	0.011
MA03B	−0.091	0.013	−0.045	0.007	SF01	−0.068	0.013	−0.034	0.007
MA04A	−0.096	0.013	−0.048	0.006	SF02	0.063	0.014	0.032	0.007
MA04B	−0.131	0.013	−0.065	0.007	SF03	−0.059	0.015	−0.029	0.007
MA05A	−0.003	0.015	−0.001	0.008	SF04	−0.044	0.014	−0.022	0.007
MA05B	−0.132	0.013	−0.066	0.007	SF05	−0.153	0.015	−0.076	0.007
MA06A	−0.178	0.013	−0.088	0.006	SF06	−0.124	0.016	−0.062	0.008
MA06B	−0.215	0.015	−0.106	0.007	SF07A	−0.168	0.016	−0.084	0.008
MR01A	−0.091	0.015	−0.046	0.008	SF07B	−0.079	0.021	−0.040	0.011
MR01B	−0.061	0.014	−0.031	0.007	SF08	−0.281	0.016	−0.139	0.008
MR02A	−0.128	0.014	−0.064	0.007	SF09	−0.013	0.014	−0.006	0.007
MR02B	−0.042	0.014	−0.021	0.007	SS01	−0.049	0.014	−0.024	0.007
MR03	−0.074	0.014	−0.037	0.007	SS02	0.021	0.014	0.011	0.007
MR04	−0.032	0.015	−0.016	0.007	SS03	−0.049	0.014	−0.024	0.007
MP01A	0.140	0.013	0.070	0.007	SS04	0.165	0.013	0.083	0.006
MP01B	0.118	0.014	0.059	0.007	SS05	0.183	0.013	0.091	0.006
MP02	0.132	0.014	0.066	0.007	SS07	0.150	0.013	0.075	0.007
MP03	0.183	0.014	0.091	0.007	SS08	0.132	0.015	0.066	0.008
MP04	0.247	0.015	0.123	0.007	SS09	0.076	0.015	0.038	0.008
MP05A	0.203	0.014	0.101	0.007					

Note: $p < .05$ for all items except the following: MA05A (p = .854), SF09 (p = .353), SS02 (p = .126).

The first part of the analysis focused on the direct effects of a booklet on mathematics and science ability, respectively. The results showed that a change from Booklet 15 to Booklet 16 predicted a slight decrease in both mathematics and science ability. However, the differences were only about 0.04 in the score, so they may not be of practical significance. For the booklet effect at the item level, a consistent pattern could be seen, where the booklet variable was associated with decreased performance for items in blocks M1 and S1 but conversely predicted increased performance for items in blocks M2 and S2. In the assessment, blocks M1 and S1 were administered in the first half of each session in Booklet 15, but in the second half in Booklet 16. On the other hand, blocks M2 and S2 were administered in the second half of each session in Booklet 15 but in the first half in Booklet 16. Our results showed that students performed better when the same item was placed earlier in a test session. Also, the booklet effect appeared stronger for some items than others. In particular, the effect generally seemed stronger for items appearing in the second half of a block.

4.4. Booklet Effect on Speed

The final structural regression model for the booklet effect on speed is shown in Figure 3. The parameter estimates for the overall model are reported in Table 8. As with the ability model, the speed model was re-run once for each screen on the test (including the dashed path) to examine the booklet effect on the response time on each screen. The parameter estimates for the dashed paths are reported in Table 9.

Table 8. Maximum likelihood estimates for the structural regression model of the booklet effect on speed at the mathematics (M) and science (S) task levels.

Parameter	Unstandardized		Standardized		Parameter	Unstandardized		Standardized	
	Estimate	SE	Estimate	SE		Estimate	SE	Estimate	SE
Factor loadings					**Factor loadings**				
Speed MTask1					Speed STask1				
MA01_S	1.000	-	0.439	0.006	SF01_S	1.000	-	0.367	0.006
MA02_S	1.941	0.034	0.568	0.005	SF02_S	1.012	0.021	0.392	0.006
MA03_S	1.347	0.024	0.542	0.005	SF03_S	0.785	0.020	0.378	0.006
MA04_S	3.134	0.053	0.637	0.005	SF04_S	0.852	0.020	0.429	0.006
MA05_S	3.105	0.052	0.656	0.004	SF05_S	0.828	0.018	0.500	0.006
MA06_S	2.773	0.049	0.575	0.005	SF06_S	0.307	0.007	0.534	0.006
					SF07_S	1.594	0.033	0.653	0.005
Speed MTask2					SF08_S	0.386	0.009	0.517	0.006
MR01_S	1.000	-	0.508	0.006	SF09_S	1.293	0.027	0.562	0.005
MR02_S	2.852	0.045	0.657	0.005					
MR03_S	1.976	0.033	0.601	0.005	Speed STask2				
MR04_S	1.947	0.040	0.423	0.006	SS01_S	1.000	-	0.473	0.005
					SS02_S	1.029	0.016	0.589	0.005
Speed MTask3					SS03_S	1.914	0.028	0.665	0.004
MP01_S	1.000	-	0.591	0.005	SS04_S	1.790	0.027	0.684	0.004
MP02_S	0.150	0.003	0.319	0.006	SS05_S	1.163	0.019	0.583	0.005
MP03_S	0.279	0.005	0.423	0.006	SS07_S	1.125	0.019	0.574	0.005
MP04_S	1.049	0.015	0.598	0.005	SS08_S	0.585	0.011	0.479	0.006
MP05_S	0.663	0.010	0.533	0.005	SS09_S	0.332	0.007	0.381	0.006
MP06_S	1.199	0.016	0.655	0.005					
MP07_S	0.920	0.014	0.549	0.005					
Direct effects									
Booklet →									
Speed MTask1	−15.027	0.276	−0.511	0.005					
Speed MTask2	−23.427	0.371	−0.614	0.006					
Speed MTask3	43.479	0.743	0.451	0.006					
Speed STask1	−22.646	0.525	−0.450	0.006					
Speed STask2	39.259	0.614	0.593	0.005					

Note: MTask1 to MTask3 refer to math tasks in the PSI test; STask1 to STask2 refer to science tasks in the PSI test. $p < .001$ for all unstandardized estimates with standard errors. Model fit indices: χ^2 (545) = 22,557.918; $p < .001$; RMSEA = 0.038; CFI = 0.892; TLI = 0.882; SRMR = 0.039.

Table 9. Maximum likelihood estimates for the structural regression model of the booklet effect on speed at individual screen level in mathematics (M) and science (S) tasks.

Screen	Unstandardized		Standardized		Screen	Unstandardized		Standardized	
	Estimate	**SE**	**Estimate**	**SE**		**Estimate**	**SE**	**Estimate**	**SE**
MA01_S	0.267	0.469	0.004	0.007	SF01_S	−5.308	0.898	−0.039	0.007
MA02_S	4.235	0.691	0.042	0.007	SF02_S	4.416	0.857	0.034	0.007
MA03_S	6.338	0.510	0.087	0.007	SF03_S	14.372	0.751	0.137	0.007
MA04_S	6.442	0.982	0.045	0.007	SF04_S	8.194	0.714	0.082	0.007
MA05_S	−7.187	0.912	−0.052	0.007	SF05_S	−6.601	0.571	−0.079	0.007
MA06_S	−15.611	0.932	−0.110	0.007	SF06_S	2.397	0.206	0.083	0.007
MR01_S	4.458	0.729	0.059	0.010	SF07_S	−5.775	0.829	−0.047	0.007
MR02_S	8.610	1.623	0.052	0.010	SF08_S	1.039	0.270	0.028	0.007
MR03_S	−4.115	1.078	−0.033	0.009	SF09_S	−14.056	0.797	−0.121	0.007
MR04_S	−14.938	1.492	−0.085	0.008	SS01_S	−1.198	1.043	−0.009	0.007
MP01_S	−4.090	1.085	−0.025	0.007	SS02_S	−7.340	0.840	−0.063	0.007
MP02_S	−2.018	0.323	−0.044	0.007	SS03_S	5.347	1.316	0.028	0.007
MP03_S	−4.575	0.446	−0.072	0.007	SS04_S	−9.010	1.228	−0.052	0.007
MP04_S	−3.432	1.132	−0.020	0.007	SS05_S	−0.101	0.960	−0.001	0.007
MP05_S	−0.433	0.825	−0.004	0.007	SS07_S	9.754	0.950	0.075	0.007
MP06_S	12.892	1.155	0.073	0.007	SS08_S	3.591	0.630	0.044	0.008
MP07_S	6.903	1.126	0.043	0.007	SS09_S	0.065	0.472	0.001	0.008

Note: $p < .05$ for all items except the following: MA01_S (p = .570), MP05_S (p = .599), SS01_S (p = .251), SS05_S (p = .916), SS09_S (p = .891).

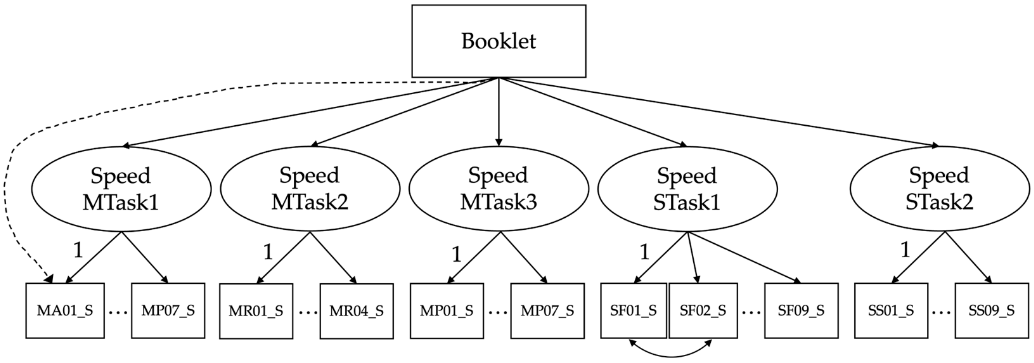

Figure 3. Final structural regression model for booklet effect on task speeds. Correlations between latent variables are not shown to minimize clutter. Model fit indices: χ^2 (545) = 22,557.918; $p < .001$; RMSEA = 0.038; CFI = 0.892; TLI = 0.882; SRMR = 0.039.

Regarding factor loadings, the average extracted variance for the speed factors ranged between 0.24 and 0.33, suggesting that the indicators for speed were fairly good [44]. For direct effects, our results showed that a change from Booklet 15 to 16 predicted an increase in speed for the tasks in blocks M1 and S1, and a decrease in speed for the tasks in blocks M2 and S2. This finding suggested that students tended to spend more time responding to the same task when it was placed in the first half of a test session. The standardized estimates, indicating effect size, suggested that the booklet effect on task-level speeds was non-trivial. The analysis of the booklet effect on individual screen response times showed a more mixed picture within each task, but larger effects tended to show up on the last screens of each task.

5. Discussion

This study examined booklet effects on students' ability and test-taking speed in a digital problem-solving and inquiry assessment in eTIMSS 2019. The two booklets contained the same tasks and items but differed in the position of the various item blocks. The results from the analysis on overall ability suggested a small but statistically significant booklet effect on overall mathematics and science ability, both being slightly lower for Booklet 16. In the booklet design, the order of the subjects and the order of appearance of the item blocks in each test session were switched in Booklet 16. Referring to the IRT item parameters published by TIMSS [4], the average difficulty (b) parameters for the four item blocks were 0.317 (M1), 0.861 (M2), 0.227 (S1), and 0.463 (S2), respectively, meaning that the items in M2 and S2 were generally more difficult than the ones in M1 and S1. In Booklet 16, students were first presented with the more difficult blocks in both test sessions. This could be a possible explanation for the observed booklet effect, which is consistent with previous research (e.g., [45–47]), which found that hard-to-easy item arrangements on a test tended to predict a lower test performance compared to easy-to-hard or random arrangements, particularly when there is an imposed time limit. These studies were typically conducted using traditional pen-and-paper multiple-choice tests.

The results from the analysis at the item level suggested a booklet effect on both ability and speed for the items appearing in the same block. When item blocks were placed in the first half of a test session, students' speed on those items was slower and performance was better. This points to a negative position effect, which is consistent with numerous other studies (e.g., [9,11,13,24]). An intuitive explanation would be that students tended to go through items more carefully and slowly at the start of each test session, but they may feel more tired, less motivated, or rushed for time toward the end of the test. Previous research surrounding item position effects often discussed fatigue effects and practice effects (e.g., [8,10,23,48]), suggesting that performance could decrease as a test progresses due to fatigue or increase due to practice if students become more familiar with the test material [49]. Due to the problem-solving nature of the PSI tasks, the presence of a fatigue effect seems more likely than a practice effect, as each item was crafted to be unique. However, as each test session was only 36 minutes long, another plausible explanation is that students might have felt more rushed for time when they attempted the second item block, affecting their performance. This finding echoes Albano's [5] argument that items with more complex content or wording may be more susceptible to position effects (i.e., perceived as more difficult) when testing time is limited. In a more recent study, Demirkol and Kelecioğlu [11] found negative position effects in the reading and mathematics domains in PISA 2015, with stronger position effects for reading and for open-ended items in mathematics, which are more complex than multiple-choice items in the same domain. Weirich et al. [16] further found that position effects were more pronounced for students whose test-taking effort decreased more throughout a test, but also pointed out that position effects remained, even in students with persistently high test-taking effort. These findings suggest that there could be multiple causes of position effects, and further research could help uncover when and why they occur.

Interestingly, all the key findings in this study pointed towards booklet effects that were unique to each item block. The swapped order of mathematics and science between the two booklets did not seem to have impacted students' performance or speed as much as the ordering of blocks within each test session. This finding suggests that the short 15-minute break between the two test sessions acted almost like a "reset button", which mitigated the position effect and gave students equal time and opportunity to perform in both portions of the assessment. In a study by Rose et al. [50], item position and domain order effects were examined concurrently in a computer-based assessment with mathematics, science, and reading items and were found to interact substantially. However, in this case, the assessment did not incorporate any breaks between the domains. When discussing the speed–ability trade-off, Goldhammer [19] recommended that item-level speed limits be set on assessments to estimate ability levels more accurately. The confounding effect of speed would be removed by ensuring that students have the same amount of time to work on each item. This controlled speed idea was later tested in a more recent study [51]. In practice, it may be challenging to implement this condition due to various technical and logistical issues. However, the results of this study suggest that administering a long assessment in separately timed sessions could be a feasible alternative to improve measurement, especially if each portion is aimed at a different construct.

Limitations and Future Research

It is necessary to acknowledge the limitations of this study. First, even though the results hinted at a possible relationship between students' ability and speed in this context (e.g., a slower speed may be related to a better performance), it was not possible to test this directly in the SEM model due to poor model fit in the combined model. In eTIMSS 2019, the total response time on each screen was captured throughout the assessment. This measured the total time that students spent on each screen, but this may not be the best measure of the actual response time (i.e., the amount of time that students spent engaging with items on each screen). For example, some students may have finished the test early or decided to take a break halfway through and lingered on some screens for longer. It was also unclear whether the screen times included overhead times (e.g., screen loading times), which could vary on different devices and contribute to increased screen times if students visited the same screen multiple times. In this study, response time outliers were removed as best as possible from the two ends of the distribution, but it was still a challenge to model speed with the existing data. More fine-grained response time data, such as those available in PISA 2018 [52], may be helpful for researchers looking to use response time data to model test-taking speed.

Second, the dataset used in this study consisted of students from all the countries who took the eTIMSS 2019 PSI booklets. While this approach provided further insights into booklet effects occurring for all students, there may be country-specific differences that could be analyzed within each country's context. Student motivation, engagement, and exposure to PSI-like items could vary widely in different countries, in addition to the level of ability. As eTIMSS is a low-stakes assessment, the results from this study may not apply to high-stakes assessments, where speed and ability may be more tightly related. As pointed out by Ong et al. [12], results from position effect studies that incorporate examinee variables (e.g., gender, effort, anxiety) tended to vary depending on the features of the testing context (e.g., content, format, and stakes associated with the test). More research is thus needed to reveal how different groups of students may be impacted by position effects in different testing contexts.

Digital assessments incorporating elements of authentic assessment (e.g., scenario-based assessment) and interactive item types are increasingly used to evaluate students' learning. As such, contextual item blocks resembling those seen in the PSI assessment may increasingly replace the typical discrete items that are used in mathematics and science assessments. This study showed that students tended to spend more time and perform better on item blocks when they were placed earlier in a test session. Test developers should be mindful of the potential effects of different orderings of item blocks on students' test-taking process. In practice, the relative difficulty of item blocks and position effects due to blocks appearing earlier or later in a test session should be considered when assembling multiple test forms.

In the PSI section of eTIMSS 2019, each task consists of a set of items that follow a narrative or theme surrounding a real-life context. Even though the items themselves are independent of each other [3], students' response and response time patterns could still be related to the specific tasks. Our findings suggested that in this context, response time patterns could be task specific. More research could be carried out to examine these patterns within a task and between tasks, alongside item-specific features such as the inclusion of interactive elements, to provide insights into students' use of time and performance in such innovative digital assessments. Future research could also examine position effects alongside item-specific and examinee-specific features to better inform test development. In this study, we analyzed data from all the countries that participated in the PSI assessment. A future study could explore country-level variations in the observed position effects and their underlying causes. Lastly, it is also worthwhile to explore how speed could be better modeled using response time data, and how the response time could be better captured in digital assessments, which may allow researchers to draw a link between ability and speed in this context.

Author Contributions: Conceptualization, J.X.L. and O.B.; methodology, J.X.L., O.B. and M.D.J.; validation, J.X.L.; formal analysis, J.X.L.; investigation, J.X.L. and M.D.J.; writing—original draft preparation, J.X.L.; writing—review and editing, J.X.L., O.B. and M.D.J.; visualization, J.X.L.; supervision, O.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study made use of data that are available from the TIMSS 2019 International Database: https://timss2019.org/international-database/ (accessed on 5 September 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mullis, I.V.S.; Martin, M.O.; Foy, P.; Kelly, D.L.; Fishbein, B. *TIMSS 2019 International Results in Mathematics and Science*; TIMSS & PIRLS International Study Centre, Boston College: Chestnut Hill, MA, USA, 2020.
2. OECD. *PISA 2015 Technical Report*; OECD: Paris, France, 2017.
3. Mullis, I.V.S.; Martin, M.O.; Fishbein, B.; Foy, P.; Moncaleano, S. *Findings from the TIMSS 2019 Problem Solving and Inquiry Tasks*; TIMSS & PIRLS International Study Centre, Boston College: Chestnut Hill, MA, USA, 2021.
4. Fishbein, B.; Foy, P.; Yin, L. *TIMSS 2019 User Guide for the International Database*, 2nd ed.; TIMSS & PIRLS International Study Centre, Boston College: Chestnut Hill, MA, USA, 2021.
5. Albano, A.D. Multilevel Modeling of Item Position Effects: Modeling Item Position Effects. *J. Educ. Meas.* **2013**, *50*, 408–426. [CrossRef]
6. Bulut, O.; Quo, Q.; Gierl, M.J. A Structural Equation Modeling Approach for Examining Position Effects in Large-Scale Assessments. *Large-Scale Assess. Educ.* **2017**, *5*, 716. [CrossRef]
7. Hahne, J. Analyzing Position Effects within Reasoning Items Using the LLTM for Structurally Incomplete Data. *Psychol. Sci. Q.* **2008**, *50*, 379–390.
8. Hohensinn, C.; Kubinger, K.D.; Reif, M.; Schleicher, E.; Khorramdel, L. Analysing Item Position Effects Due to Test Booklet Design within Large-Scale Assessment. *Educ. Res. Eval.* **2011**, *17*, 497–509. [CrossRef]
9. Nagy, G.; Nagengast, B.; Becker, M.; Rose, N.; Frey, A. Item Position Effects in a Reading Comprehension Test: An IRT Study of Individual Differences and Individual Correlates. *Psychol. Test Assess. Model.* **2018**, *60*, 165–187.
10. Christiansen, A.; Janssen, R. Item Position Effects in Listening but Not in Reading in the European Survey of Language Competences. *Educ. Assess. Eval. Acc.* **2020**, *33*, 49–69. [CrossRef]
11. Demirkol, S.; Kelecioğlu, H. Investigating the Effect of Item Position on Person and Item Parameters: PISA 2015 Turkey Sample. *J. Meas. Eval. Educ. Psychol.* **2022**, *13*, 69–85. [CrossRef]
12. Ong, T.Q.; Pastor, D.A. Uncovering the Complexity of Item Position Effects in a Low-Stakes Testing Context. *Appl. Psychol. Meas.* **2022**, *46*, 571–588. [CrossRef] [PubMed]
13. Debeer, D.; Janssen, R. Modeling Item-Position Effects Within an IRT Framework: Modeling Item-Position Effects. *J. Educ. Meas.* **2013**, *50*, 164–185. [CrossRef]
14. Sideridis, G.; Hamed, H.; Jaffari, F. The Item Position Effects in International Examinations: The Roles of Gender. *Front. Psychol.* **2023**, *14*, 1220384. [CrossRef]
15. Lindner, M.A.; Lüdtke, O.; Nagy, G. The Onset of Rapid-Guessing Behavior Over the Course of Testing Time: A Matter of Motivation and Cognitive Resources. *Front. Psychol.* **2019**, *10*, 1533. [CrossRef]
16. Weirich, S.; Hecht, M.; Penk, C.; Roppelt, A.; Böhme, K. Item Position Effects Are Moderated by Changes in Test-Taking Effort. *Appl. Psychol. Meas.* **2017**, *41*, 115–129. [CrossRef] [PubMed]
17. Tan, B. Response Time as a Predictor of Test Performance: Assessing the Value of Examinees' Response Time Profiles. Master's Thesis, University of Alberta, Edmonton, AB, Canada, 2023.
18. Fox, J.-P.; Marianti, S. Joint Modeling of Ability and Differential Speed Using Responses and Response Times. *Multivar. Behav. Res.* **2016**, *51*, 540–553. [CrossRef] [PubMed]
19. Goldhammer, F. Measuring Ability, Speed, or Both? Challenges, Psychometric Solutions, and What Can Be Gained From Experimental Control. *Meas. Interdiscip. Res. Perspect.* **2015**, *13*, 133–164. [CrossRef] [PubMed]
20. Wise, S.L. Response Time as an Indicator of Test Taker Speed: Assumptions Meet Reality. *Meas. Interdiscip. Res. Perspect.* **2015**, *13*, 186–188. [CrossRef]
21. Hecht, M.; Weirich, S.; Siegle, T.; Frey, A. Effects of Design Properties on Parameter Estimation in Large-Scale Assessments. *Educ. Psychol. Meas.* **2015**, *75*, 1021–1044. [CrossRef]
22. Fishbein, B.; Foy, P. Scaling the TIMSS 2019 Problem Solving and Inquiry Data. In *Methods and Procedures: TIMSS 2019 Technical Report*; Martin, M.O., von Davier, M., Mullis, I.V.S., Eds.; TIMSS & PIRLS International Study Centre, Boston College: Chestnut Hill, MA, USA, 2021; pp. 17.1–17.51.

23. Kingston, N.M.; Dorans, N.J. Item Location Effects and Their Implications for IRT Equating and Adaptive Testing. *Appl. Psychol. Meas.* **1984**, *8*, 147–154. [CrossRef]
24. Wu, Q.; Debeer, D.; Buchholz, J.; Hartig, J.; Janssen, R. Predictors of Individual Performance Changes Related to Item Positions in PISA Assessments. *Large-Scale Assess. Educ.* **2019**, *7*, 5. [CrossRef]
25. Thurstone, L.L. Ability, Motivation, and Speed. *Psychometrika* **1937**, *2*, 249–254. [CrossRef]
26. van der Linden, W.J. A Hierarchical Framework for Modeling Speed and Accuracy on Test Items. *Psychometrika* **2007**, *72*, 287–308. [CrossRef]
27. Tijmstra, J.; Bolsinova, M. On the Importance of the Speed-Ability Trade-Off When Dealing with Not Reached Items. *Front. Psychol.* **2018**, *9*, 964. [CrossRef] [PubMed]
28. Goldhammer, F.; Naumann, J.; Stelter, A.; Tóth, K.; Rölke, H.; Klieme, E. The Time on Task Effect in Reading and Problem Solving Is Moderated by Task Difficulty and Skill: Insights from a Computer-Based Large-Scale Assessment. *J. Educ. Psychol.* **2014**, *106*, 608–626. [CrossRef]
29. van der Linden, W.J. A Lognormal Model for Response Times on Test Items. *J. Educ. Behav. Stat.* **2006**, *31*, 181–204. [CrossRef]
30. Klein Entink, R.H.; Fox, J.-P.; van der Linden, W.J. A Multivariate Multilevel Approach to the Modeling of Accuracy and Speed of Test Takers. *Psychometrika* **2009**, *74*, 21–48. [CrossRef] [PubMed]
31. Scherer, R.; Greiff, S.; Hautamäki, J. Exploring the Relation between Time on Task and Ability in Complex Problem Solving. *Intelligence* **2015**, *48*, 37–50. [CrossRef]
32. De Boeck, P.; Jeon, M. An Overview of Models for Response Times and Processes in Cognitive Tests. *Front. Psychol.* **2019**, *10*, 102. [CrossRef] [PubMed]
33. Li, F.; Cohen, A.; Shen, L. Investigating the Effect of Item Position in Computer-Based Tests. *J. Educ. Meas.* **2012**, *49*, 362–379. [CrossRef]
34. Yin, L.; Foy, P. TIMSS 2023 Assessment Design. In *TIMSS 2023 Assessment Frameworks*; Mullis, I.V.S., Martin, M.O., von Davier, M., Eds.; TIMSS & PIRLS International Study Centre, Boston College: Chestnut Hill, MA, USA, 2023; pp. 71–85.
35. Vida, L.J.; Brinkhuis, M.J.S.; Bolsinova, M. Speeding up without loss of accuracy: Item position effects on performance in university exams. In Proceedings of the 14th International Conference on Educational Data Mining, Virtual, 29 June–2 July 2021.
36. Martin, M.O.; Mullis, I.V.S.; Foy, P. TIMSS 2019 Assessment Design. In *TIMSS 2019 Assessment Frameworks*; Mullis, I.V.S., Martin, M.O., Eds.; TIMSS & PIRLS International Study Centre, Boston College: Chestnut Hill, MA, USA, 2017; pp. 81–91.
37. Cousineau, D.; Chartier, S. Outliers Detection and Treatment: A Review. *Int. J. Psychol. Res.* **2010**, *3*, 58–67. [CrossRef]
38. Berger, A.; Kiefer, M. Comparison of Different Response Time Outlier Exclusion Methods: A Simulation Study. *Front. Psychol.* **2021**, *12*, 675558. [CrossRef]
39. Muthén, L.K.; Muthén, B.O. *Mplus User's Guide*, 8th ed.; Muthén & Muthén: Los Angeles, CA, USA, 2017.
40. Hu, L.; Bentler, P.M. Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria versus New Alternatives. *Struct. Equ. Model.* **1999**, *6*, 1–55. [CrossRef]
41. Xia, Y.; Yang, Y. RMSEA, CFI, and TLI in Structural Equation Modeling with Ordered Categorical Data: The Story They Tell Depends on the Estimation Methods. *Behav. Res. Methods* **2019**, *51*, 409–428. [CrossRef] [PubMed]
42. Kline, R.B. *Principles and Practice of Structural Equation Modeling*, 5th ed.; The Guildford Press: New York, NY, USA, 2023.
43. Bowerman, B.L.; O'Connell, R.T. *Linear Statistical Models: An Applied Approach*, 2nd ed.; Duxbury: Belmont, CA, USA, 1990.
44. Comrey, A.L.; Lee, H.B. *A First Course in Factor Analysis*, 2nd ed.; Lawrence Erlbaum Associates, Inc.: Hillsdale, NJ, USA, 1992.
45. Hambleton, R.K.; Traub, R.E. The Effects of Item Order on Test Performance and Stress. *J. Exp. Educ.* **1974**, *43*, 40–46. [CrossRef]
46. Ollennu, S.N.N.; Etsey, Y.K.A. The Impact of Item Position in Multiple-Choice Test on Student Performance at the Basic Education Certificate Examination (BECE) Level. *Univers. J. Educ. Res.* **2015**, *3*, 718–723. [CrossRef]
47. Sax, G.; Cromack, T.R. The Effects of Various Forms of Item Arrangements on Test Performance. *J. Educ. Meas.* **1966**, *3*, 309–311. [CrossRef]
48. Davis, J.; Ferdous, A. *Using Item Difficulty and Item Position to Measure Test Fatigue*; American Institutes for Research: Washington, DC, USA, 2005.
49. Yoo, N. Item Position and Motivation Effects in Large-Scale Assessments. Ph.D. Thesis, Columbia University, New York, NY, USA, 2020.
50. Rose, N.; Nagy, G.; Nagengast, B.; Frey, A.; Becker, M. Modeling Multiple Item Context Effects with Generalized Linear Mixed Models. *Front. Psychol.* **2019**, *10*, 248. [CrossRef]
51. Goldhammer, F.; Kroehne, U.; Hahnel, C.; De Boeck, P. Controlling Speed in Component Skills of Reading Improves the Explanation of Reading Comprehension. *J. Educ. Psychol.* **2021**, *113*, 861–878. [CrossRef]
52. OECD. *PISA 2018 Technical Report*; OECD: Paris, France, 2019.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.