International Journal of Assessment Tools in Education
2025, Vol. 12, No. 3, 771–786
https://doi.org/10.21449/ijate.1532862
journal homepage: https://dergipark.org.tr/en/pub/ijate
Research Article

Item block position and format effects in e-TIMSS among the low- and high-achieving countries

Neşe Öztürk Gübeş[1*]

[1]Burdur Mehmet Akif Ersoy University, Faculty of Education, Department of Educational Sciences, Burdur, Türkiye

ARTICLE HISTORY

Received: Aug. 14, 2024

Accepted: June 4, 2025

Keywords:

Item position effect, Explanatory item, Response theory, TIMSS.

Abstract: The Trends in International Mathematics and Science Study (TIMSS) was administered via computer, eTIMSS, for the first time in 2019. The purpose of this study was to investigate item block position and item format effect on eighth grade mathematics item easiness in low- and high-achieving countries of eTIMSS 2019. Item responses from Chile, Qatar, and Malaysia which were low-achieving countries as well as Republic of Korea, Chinese Taipei, and Singapore which were high-achieving countries, were used in the study. The block position and item format effects were investigated within explanatory item response theory framework. The results revealed that there was a negative and statistically significant item block position effect in all low-and high-achieving countries, and it is more prominent in the low-achieving countries. As item block increased, students' probability of giving a correct response to an item decreased. Additionally, the results showed that all high- and low-achieving countries had a negative and significant item format effect in that multiple-choice items appeared easier compared to constructed response items.

1. INTRODUCTION

In international large-scale assessments such as the Trends in International Mathematics and Science Study (TIMSS) or Programme for International Student Assessment (PISA), items are exhibited within a test using a booklet design and in different positions. The advantages of using multiple booklets ensure test security and administering a greater number of field -test items fixed into different booklets (Bulut *et al*., 2017). However, administering the same items in different orders may have undesirable effects on test and item characteristics (Leary & Dorans, 1985). Item position effect is defined as the interaction between item's position in a test booklet and examinees' performance on that item (Qian, 2014). There are two kinds of item-position effects, namely practice or learning effect, and fatigue effect. The practice effect can occur when an item at the beginning of a test is more difficult than the same item implemented at the end of the test. The fatigue effect can occur when an item at the beginning of the test is less difficult than the same item implemented at the end of the test (Hahne, 2008; Hohensinn *et al*.,

*CONTACT: Neşe ÖZTÜRK GÜBEŞ nozturk@mehmetakif.edu.tr Burdur Mehmet Akif Ersoy University, Faculty of Education, Department of Educational Sciences, Burdur, Türkiye

e-ISSN: 2148-7456

2008; Yoo, 2020). As Wu *et al*. (2019) reported, if all examinees would be affected by the item position effect in the same way, not considering position effects would not result in unfair comparisons. However, in previous research studies, it was indicated that item position impacted low-achieving test takers more than high -achieving test takers (Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Klosner & Gellman, 1973; Wise *et al*.,1989). In this vein, the first aim of this study is to investigate item position effect among the low- and high-achieving countries within the scope of computerized version of the Trends in International Mathematics and Science Study (eTIMSS) in 2019, specifically eighth grade mathematics item easiness. Item format also influences success scores (Hastedt & Sibberns, 2005). Thus, another aim of this research is to investigate how item format effects item easiness in the relevant low- and high -achieving countries in eTIMSS 2019.

The large-scale assessment such as PISA, TIMSS, and Progress in International Reading Literacy Study (PIRLS) use multiple – matrix-sample booklet designs (Gonzalez & Rutkowski, 2010). In these designs, several hundred questions are used, and each student is administered a particular combination of test items. These item subtests are usually called booklets which are arranged using rules and techniques such as multiple matrix sampling (Gonzalez & Rutkowski, 2010; Hecht *et al*., 2015). In a matrix sampling design, items are gathered into blocks where each of blocks contains a definite number of items and each booklet comprises of multiple blocks. The booklets are randomly distributed to students. TIMSS, PIRLS, and National Assessment of Educational Progress (NAEP) employ balancing methods to control position effects. The balanced incomplete block (BIB) is one of the designs for balancing item position. In the BIB design, every block holds equally frequency across all booklets and every pair of blocks appears together in booklets with equal frequency (Frey *et al*., 2009). The block positions change across different booklets, but the selections of items and their position within a block are fixed so that in the BIB design item position does not change within block (Weirich *et al*., 2014). It should be noted that in this study the term "item position" refers to the block position in which the item releases.

When assembling blocks of items, the orders of items within or across blocks can affect item parameters. In previous research studies, item block position effect in TIMSS and PISA assessments has been frequently investigated. To start with, Hartig and Buchholz (2012) analyzed item position effects and individual persistence with the data from PISA 2006 science assessment from 10 countries within the multilevel item response model, and found significance and negative item position effect in all countries, which was more noticeable in the low-achieving countries. Deeber and Janssen (2013), on the other hand, investigated the item-position effect in Turkey PISA 2006 reading, math, and science data with explanatory item response theory (EIRT) framework. They revealed that for all their literacies, items were more difficult when placed in later positions. Similarly, Deeber *et al*. (2014) used multilevel item response theory (IRT) framework to analyze item position effect and the variance in examinee effort throughout testing in PISA 2009 paper-and-pencil reading assessment, and found a negative position effect in all countries. Nagy *et al*. (2018) analyzed position effect in science, mathematics and reading tests administered in the German extension to the PISA 2006 study, and reported negative position effects in all domains. Wu *et al*. (2019) also investigated item position effect within multilevel IRT framework via using data of PISA 2006 science, 2009 reading, and 2012 mathematics assessments, and indicated a whole negative item block position effect in all the countries in all PISA domains. Yoo (2020) modeled item position effect within Structural Equation Modelling (SEM) framework via using TIMSS 2015 eighth grade mathematics data from the U.S. sample, and concluded that items tended to be harder to get correct when they arise in later part of the test. Demirkol and Kelecioğlu (2022) investigated item position effect in PISA 2015 Turkey sample within the EIRT framework, and found negative and significant effect in reading and mathematics domains. In a very recent study, Liu

et al. (2024) used a SEM approach to examine position effects on students' ability and test-taking speed, using data from eTIMSS 2019 fourth grade problem -solving and inquiry tasks. They found a negative position effect when item blocks were in the first half of a test session, and reported that students' performance was better. As a result, it can be claimed that fatigue effects have been commonly reported in research studies which focused on item block position effect in PISA and TIMSS assessments.

There are also research studies which focus on item format effect in TIMSS and PISA assessments. Hastedt and Sibberns (2005), for instance, investigated the differences in success between constructed response (CR) and multiple choice (MC) items using data sets from TIMSS 1995 and 1999. They found small differences between the scale scores based on different item formats and concluded that the Eastern European countries did not perform well on the CR items as much as on the MC items. In a similar vein, Liou and Bulut (2020) examined item format effect on eighth grade Taiwanese students' TIMSS 2011 science performance, and their proportion correct statistics showed that the CR items were more difficult compared to the MC items. İlhan *et al*. (2020) examined item format effect on fourth and eighth grade Turkish students' performance in TIMSS 2015 math assessment and stressed that Turkish students were more successful in the MC items compared to the CR items, no matter what the cognitive level of the item was. The studies which focused on students' success in different item formats at PISA showed that MC items had the highest success rates (Demir, 2010; Özer-Özkan & Özaslan, 2018). In a very comprehensive study, Marcq *et al*. (2024) examined students' responses of mathematics assessment from 71 countries PISA 2018. They revealed that the CR items were more difficult than the MC items across all countries, and it was more predominant among the low-achieving countries.

1.1. The Aim and Significance of the Research

In international large-scale assessments such as TIMSS, IRT models are used to transform each student raw score into a single scale score. Therefore, correct modeling of the item responses is very important to compare countries and students validly (Christiansen & Janssen, 2021). Modeling the measurement model without considering item position effects may threat local independence assumption of IRT (Christiansen & Janssen, 2021; Gonzalez & Rutkowski, 2010; Hartig & Buchholz, 2012) because "the probability for giving correct responses usually is assumed to depend only on properties of items and persons which are assumed to be independent from presentation conditions and item context" (Hartig & Buchholz, 2012, p.419). The position effects of blocks in different booklets can be one of major causes of biased item parameter estimates. If an item appeared in a block at the beginning of a booklet, students may give correct answers more often than the items appearing at the end of a booklet because of growing fatigue, reduced motivation or merely a lack of time. The difference in item difficulties may be because of the variation of item's block position and not because of the cognitive demands of the item (Frey *et al*., 2009). Additionally, item position can affect interchangeability between test forms (Meyers *et al.,* 2008; Sideridis *et al.,* 2023). In the relevant literature, it is reported that the presence of item or item block position effect threats item parameter invariance (Meyers *et al*., 2008), validity of test results (Hahne, 2008; Hohensinn *et al*., 2008) as well as conclusions drawn from test results. Besides, large position effects lead to measurement error while estimating item parameters and test scores (Hahne, 2008). Therefore, it is notable to determine and quantify item position effect in TIMSS, which makes it possible for countries to monitor and take decisions for improving their educational system. Investigating item position effect is also important for test practices. "For example, if negative item position effects on performance are known, the maximum test length that can be administered to test-takers without overly impairing the assessed performance can be determined" (Hartig & Buchholz, 2012, p. 419).

In international large-scale assessments like TIMSS, multiple choice and constructed response items are used together. There are some research studies related to TIMSS and PISA and these studies showed that item format sometimes had an influence on students' performance (Demir, 2010; Hastedt & Sibberns, 2005; İlhan *et al*., 2020; Liou & Bulut, 2020; Marcq *et al*., 2024; Özer-Özkan & Özaslan, 2018). In 2019, for the first time, TIMSS was administered via computer. In the literature, the studies which focused on item block position effect and item format effect were conducted with paper-pencil versions of TIMSS assessments. However, do item format and block position effect have an influence on mathematics item easiness of computerized version of TIMSS 2019? Previous studies reported that item position impacted low-achieving test takers more than high -achieving test takers (Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Klosner & Gellman, 1973; Marcq *et al.*, 2024; Wise *et al.*, 1989). If we consider the countries which participated in eTIMSS 2019, the research questions of the present study are: (1) Would the position of item block have affected the lower- and higher- achieving countries in the same way and direction? (2) Would the item format have affected the low- and high- achieving countries in the same way and direction?

The aim of this study is to investigate the effects of item block position and item format effect on eTIMMS 2019 eighth grade mathematics item easiness among the selected low-and high-achieving countries.

2. METHOD

2.1. Data

This study was conducted with 280 eighth grade mathematics items in 14 achievement booklets of eTIMSS 2019. In addition to 14 achievement booklets, eTIMSS 2019 also has two special ebooklets (Booklet 15 & Booklet 16) which were designed to assess Problem Solving and Inquiry (PSI) tasks. All eTIMSS countries (but no paperTIMSS countries) took the PSI items (Mullis *et al*., 2021). The Booklet 15 and Booklet 16 which have PSI task were excluded from the current study.

TIMSS is an international assessment which is designed to gather information about students' achievement in mathematics and science, and has been conducted every four-year period since 1995. TIMSS seventh cycle was conducted in 2019 at the fourth and eighth grades in 64 countries and 8 benchmarking systems (Mullis & Martin, 2017). In 2019, for the first time, TIMSS was administered via computer and a computerized version was introduced as eTIMSS. More than half of the 64 participating countries chose to administer eTIMSS whereas the rest of the countries continued to administer paper-based version (Mullis *et al*., 2020). The TIMSS uses mathematics and science scales. The midpoint of scales are 500 points and the standard deviation of scale is 100 points. The TIMSS assessments have been reported on the same scale metrics since 1995 and therefore participant countries have the chance to track students' growth across every TIMSS's cycles. In this study, eTIMSS 2019 grade eighth grade mathematics test item responses from Chile ($\bar{X}$ = 441), Qatar ($\bar{X}$ = 443) and Malaysia ($\bar{X}$ = 461), which have overall average mathematics scale score below midpoint ($\bar{X}$ = 500); Republic of Korea ($\bar{X}$ = 607), Chinese Taipei ($\bar{X}$ = 612) and Singapore ($\bar{X}$ = 616), which have overall average mathematics scale score above midpoint, were used. The sample sizes for each booklet in the selected countries are presented in Table 1. As seen in Table 1, there were totally 4115 students in Chile, 3884 students in Qatar, 7065 students in Malaysia, 3861 students in Republic of Korea, 4915 students in Chinese Taipei, and 4853 students in Singapore samples.

TIMSS uses a matrix sampling approach that includes gathering all assessment pool of mathematics and science items at each grade level in 14 achievement booklets. TIMSS groups assessment items into item blocks which approximately include 10 to 14 items at the fourth grade, and 12 to 18 items at the eighth grade. TIMSS 2019 totally has 28 blocks at each grade

14 blocks containing mathematics item and 14 blocks containing science items. Each student booklet contains two blocks of mathematics and two blocks of science items. Half of the booklets begin with the two mathematics blocks and continue with the two science blocks. On the other hand, the other half of the booklets begin with two science blocks and continue with the two mathematics blocks (Martin *et al*., 2017). In TIMSS, each item occurs at two different block positions (Block Position 1 and 4 or Block Position 2 and 3) with equal frequency. Item position is fixed in each block. In the current study, item block position is considered as item position and four item block positions, ranging from position one to position four.

Table 1. *Sample sizes of selected countries.*

Countries	Girl		Boy		System Missing		Total
	N	Percent	*N*	Percent	*N*	Percent	
Chile	1973	47.95	2113	51.35	29	.70	4115
Qatar	1897	48.84	1979	51.06	8	.20	3884
Malaysia	3702	52.40	3363	47.60	0	0	7065
Republic of Korea	1920	49.73	1938	50.19	3	.08	3861
Chinese Taipei	2449	49.83	2464	50.13	2	.04	4915
Singapore	2366	48.75	2486	51.23	1	.02	4853

The TIMSS 2019 eighth grade assessment includes a broad variety of selected response and constructed response items. The paper- and eTIMSS 2019 assessment have two general types of selected response items: single selection and multiple selection. Students choose one of the four options in single selection items, while in multiple selection items they choose more than one options. While most of the selected response items were scored as 1-score point, multiple selection items were scored as 2-score points (fully correct or all parts answered correctly gets 2 score points; partially correct or most part answered correctly gets 1 score points; incorrect or no parts answered correctly gets 0 score points). CR items which require writing or typing words or numbers, or dragging and dropping for eTIMSS were scored as 1- or 2-score points. The 1-score point CR items were scored as 1 for correct and 0 for incorrect answers. The 2-score point items were scored as 2 for fully correct answers, 1 for partially correct answers, and 0 for incorrect answers (Cotter *et al.*, 2020).

This study was conducted with the help of explanatory item response models for dichotomously scored items. Although partial credit model (PCM; Masters, 1982) can give more detailed information about the performance of students' score categories of polytomous scored items, it has some limitations. PCM is mathematically more complex to use than the dichotomous Rasch model, and it has some difficulties in explaining item measures (He & Wheadon, 2013). In the current study, we preferred to recode polytomous item responses as dichotomously (as in other studies such as Debeer & Janssen, 2013; Demirkol & Kelecioğlu, 2022; Wu *et al*., 2019; all of which focused on TIMSS and PISA). In this way, 2-scored MC and CR items were recoded as 1 point for fully correct or when all parts were answered correctly, and 0 point for partially correct and incorrect answers. Omitted or invalid responses were scored as incorrect and not reached items were recorded as 9 as in TIMSS scoring process (Debeer & Janssen, 2013; Wu *et al*., 2019). The 12 items which have no valid cases in Chile, Qatar, Malaysia, Republic of Korea, Chinese Taipei, and Singapore samples were not included in this study. The excluded items are ME62342, ME62345BA, ME62345BB, ME62345BC, ME62345BD, ME62345B, ME72038, ME72211B, ME62048A, ME62048B, ME62048C, ME62048.

In this study, information regarding eTIMSS 2019 eighth grade mathematics items was gathered from "eT19_G8_Item Information" excel file from TIMSS international data base. Table 2 shows frequency of item format.

Table 2. *Distribution of common items based on item format.*

Common Items	CR	MC	Total
Booklet 1&2	13	15	28
Booklet 2&3	9	8	17
Booklet 3&4	13	18	31
Booklet 4&5	10	9	19
Booklet 5&6	7	8	15
Booklet 6&7	8	9	17
Booklet 7&8	11	11	22
Booklet 8&9	14	6	20
Booklet 9&10	12	8	20
Booklet 10&11	9	7	16
Booklet 11&12	11	6	17
Booklet 12&13	9	6	15
Booklet 13&14	12	15	27
Booklet 1&14	5	11	16
Total	143	137	280

Note. CR: Constructed Response; MC: Multiple Choice

2.2. Data Analyses

In this study, the item block position and item format effect were investigated within explanatory item response models. The typical item response theory models are descriptive. In these models, each item is modeled with its own set of parameters, and person parameters are also estimated. On the other hand, item response theory models can be used to understand how the item responses can be explained by item and/or person properties. These models are called as "explanatory item response theory models" (Wilson & De Boeck, 2004). Most of explanatory item response theory (EIRT) models are exceptional forms of generalized linear mixed models (GLMMs) or nonlinear mixed models (NLMMs) (Wilson *et al*., 2008). In these models, item responses are taken into consideration as repeated observations which are embedded within each person in a multilevel structure. If a GLMM includes unique parameters for both items and test takers, then it turns into Rasch model for 1-0 scored data (Bulut *et al*., 2021). The Rasch model can be reformulated for GLMM as follows (Wilson *et al*., 2008):

$$\eta_{pi} = \theta_p + \beta_i \quad (1)$$

With $\beta_i = \sum_k \beta_i X_{ik}$ and where θ_p is the ability level of individual p; β_i item-specific coeficients or item parameters represent item easiness (item easiness is the negative of the item difficulty parameter in traditional IRT modeling); η_{pi} is the average of a distribution that has a logit or probit link to the probability denoted in Equation 1. Wilson and De Boeck (2004) explained the Rasch Model as doubly descriptive models which are descriptive for both the individual and item sides of the matrix. It explains person variations via a random person parameter θ and describes item variations via unique or fixed item parameters (Wilson *et al*., 2008). When individual-level covariates are included in the model, the EIRT model is named as a person explanatory model or latent regression Rasch model; when item-level covariates are included but person-level covariates are not in the EIRT model, it is called as item explanatory model or the linear logistic test model (LLTM); when both item and individual covariates are both included in the model, the EIRT model is called doubly explanatory model or latent regressing LLTM model (Wilson *et al*., 2008). In this study, item explanatory model, in other words the LLTM, was employed.

In the LLTM, item aspects are used to define how the probability of replying an item correctly changes due to the characteristics of the items. In the current study, we used "+ -parametrization," in the LLTM equation which leads to an interpretation of β_i as the item easiness (Wilson *et al*., 2008):

$$\eta_{pi} = \theta_p + \sum_{k=0}^{K} \beta_k X_{ik} \quad (2)$$

Where X_{ik} is the value item i on item aspect k ($k=0,\ldots,K$), and β_k is the regression weight of the item aspect k (Wilson *et al*., 2008). With item explanatory models, the effects of several item properties (such as item position, item format) and their interactions on item easiness variation can be investigated. The LLTM assumes that item easiness can be perfectly estimated by the item properties, which is very strong assumption for the model. An error term can be added to the LLTM (Janssen *et al.*, 2004):

$$\eta_{pi} = \theta_p + \sum_{k=0}^{K} \beta_k X_{ik} + \varepsilon_i \quad (3)$$

$$\varepsilon_i \sim 0, \sigma_\varepsilon^2$$

Adding the random item effect $(1|\text{item})$ to the LLTM implies adding a random error to the model, and the LLTM with random error is very useful. A larger random error variance shows that the item covariates have less explanatory power to explain the item easiness (De Boeck *et al.*, 2011; Janssen *et al.*, 2004). In this study, LLTM plus random error was used to model item block position and item format effects.

The *eirm ()* function from the eirm (Bulut, 2021) R package was used in analyzing the model. The *eirm ()* function needs a "long form" for the data to be modeled. The typical test data are wide-format, and there are several observations which were made for several individuals. However, in long-format data, responses are nested within persons, each row represents one item and thus, each person has multiple rows. In this study, the data were restructured to long format with using *melt ()* function in the reshape2 (Wickham, 2007) R package. After that, long-format data were used to describe item-level predictors which are item block position and item format. To make the item block position effect identifiable, a reference position has to be chosen and is fixed to zero (Debeer & Janssen, 2013). For this study, the first block position was chosen as reference position and recoded as "0", hence the block position was coded 0 to 3. The item format of MC was chosen as reference and coded as "0", constructed response items were coded with given "1". To investigate how item block position (BP) and item format (Iformat) account for item easiness, the LLTM was estimated as follows:

$$\text{Model} = \text{"responses} \sim 1 + \text{BP} + \text{Iformat} + (1|\text{student}) + (1|\text{item})\text{"} \quad (4)$$

In Equation 4, responses for the binary response, 1 for adding intercept to the model, BP and Iformat were item covariates (fixed effects), $(1|\text{student})$ defines random effects for students and $(1|\text{item})$ defines random effects for items.

In addition, tidyverse (Wickham *et al*., 2019) R package was used to calculate descriptive statistics of long format data and eirm (Bulut, 2021) R package was used to conduct the LLTM analyses and obtain graphics.

3. RESULTS

The mean of items by block position and item format are presented in Table 3. As seen in Table 3, in all countries, the mean of MC items is larger than that of CR items. Up to Block 3, the item mean increases as the order of blocks increases in all countries. However, the items get minimum average at Block 4 in all countries (except for Chinese Taipei, the average of items at Block 1 and Block 4 are the same in Chinese Taipei). The LLTM was estimated to investigate how block position and item format account for item easiness. Table 4 presents the LLTM results for eTIMSS 2019 for the chosen six countries.

Table 3. *The mean of items in eTIMSS 2019 mathematic assessment by block position and item format.*

Countries		Block Position				Item Format	
		Block 1	Block 2	Block 3	Block 4	MC	CR
Chile	*N*	2057	2047	2058	2051	4115	4115
	Mean	.313	.327	.350	.277	.434	.211
Qatar	*N*	1929	1922	1955	1952	3884	3884
	Mean	.319	.350	.368	.284	.450	.223
Malaysia	*N*	3531	3529	3534	3533	7065	7065
	Mean	.409	.412	.439	.387	.521	.311
Republic of Korea	*N*	1942	1939	1919	1916	3861	3861
	Mean	.647	.686	.695	.638	.745	.598
Chinese Taipei	*N*	2452	2451	2463	2458	4915	4915
	Mean	.651	.705	.717	.651	.746	.628
Singapore	*N*	2424	2423	2429	2427	4853	4853
	Mean	.678	.698	.707	.669	.753	.630

Note. N: Sample Size; CR: Constructed Response; MC: Multiple Choice

In this study, the LLTM analyses were conducted separately for each country and θ-scale was not same across countries. Therefore, to be able to compare estimated effects, the block position and item format effects were standardized by using the standard deviation of the ability level of each country (σ_θ): ($\gamma^* = \gamma/\sigma_\theta$) (Debeer *et al.*, 2014; Hartig & Buchholz, 2012; Wu *et al.*, 2019). Table 5 presents the standardized item block position and item format effects for each country.

If we look at block position effects in Table 4, there is a negative and statistically significant effect in all eTIMSS 2019 among the low-achieving (γ_{Chile}= -.13; z_{Chile}= -12.68, p<.001; γ_{Qatar}= -.13; z_{Qatar}= -11.36, p<.001; $\gamma_{Malaysia}$= -.11, $z_{Malaysia}$= -13.65, p<.001) and high-achieving (γ_{Korea}=-.06, z_{Korea}=-4.71; $\gamma_{Chinese\ Taipei}$= -.04, $z_{Chinese\ Taipei}$=-3.37, p<.001; $\gamma_{Singapore}$= -.05, $z_{Singapore}$= -4.67, p<.001) countries.

According to standardized item block position effect in Table 5, we can say that the strength of effects varies across countries, and it is more prominent in the low-achieving countries. While the country with the highest standardized block position effect is Chile (γ^*= -.13), the country with the lowest standardized block position effect is Chinese Taipei (γ^*= -.02).

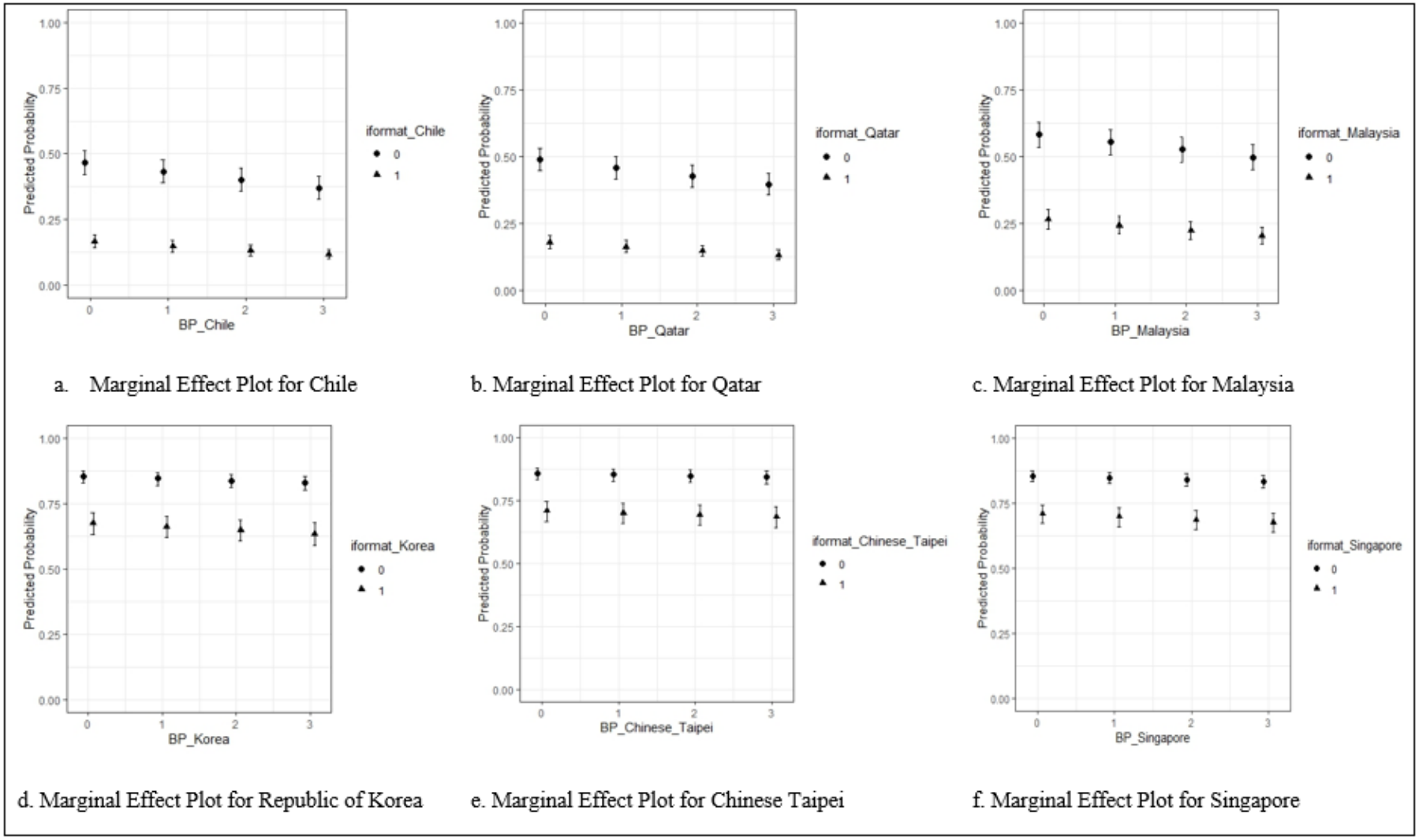
Negative position effect means that asking an item one block further leads to decreasing the probability of giving a correct response to that item. In terms of the probability of giving correct responses, this could result in a decrease of .03 for Chile and .004 in Chinese Taipei for a student with average ability (θ=0) in case of an item with average difficulty (β=0) (Wu *et al.*, 2019). The graphs in Figure 1 show that the predicted probabilities of MC and CR items decrease as block position increases in the low-achieving countries. However, in the high-achieving countries, while the predicted probabilities of CR items decrease as item block position increases, the predicted probabilities of MC items remain roughly the same as block position increases.

Table 4. *The LLTM results.*

	Chile		Qatar		Malaysia		Republic of Korea		Chinese Taipei		Singapore	
	γ (SE)	z	γ (SE)	z	γ (SE)	z	γ (SE)	z	γ (SE)	z	γ (SE)	z
Fixed Effect												
Intercept	-.14(.10)	-1.43	-.05(.09)	-.54	.34(.10)	3.42***	1.76(.10)	18.37***	1.80(.10)	18.02***	1.78(.09)	19.62***
Block Position	-.13(.01)	-12.68***	-.13(.01)	-11.36***	-.11(.01)	-13.65***	-.06(.01)	-4.71***	-.04(.01)	-3.37***	-.05(.01)	-4.67***
Item Format	-1.49(.13)	-11.55***	-1.47(.12)	-12.70***	-1.36 (.13)	-10.11***	-1.03(.13)	-8.17***	-.91(.13)	-6.82***	-.89(.12)	-7.35***
Random Effects												
	Var.	*SD*	*Var.*	*SD*	*Var.*	*SD*	*Var.*	*SD*	*Var.*	*SD*	*Var.*	*SD*
Students	1.01	1.00	1.25	1.12	1.65	1.29	2.54	1.60	2.69	1.64	2.29	1.51
Items	1.15	1.07	.93	.96	1.26	1.12	1.10	1.05	1.24	1.11	1.01	1.00

Note. The estimated parameters in the Table 4 represent easiness; ***p<.001

Figure 1. *Predicted probabilities of correct answers.*



Note. In graphs, BP represents block position, "0" represents multiple choice items, "1" represents constructed response items.

Table 4 shows that the estimated item format effects are negative and statistically significant for all eTIMSS 2019 among the low-achieving (γ_{Chile}= -1.49, z_{Chile}=-11.55, p< .001; γ_{Qatar}= -1.47, z_{Qatar}= -12.70, p< .001; $\gamma_{Malaysia}$= -1.36, $z_{Malaysia}$= -10.11, p< .001) and high-achieving (γ_{Korea}= -1.03, z_{Korea}= -8.17, p< .001; $\gamma_{Chinese\ Taipei}$= -.91, $z_{Chinese\ Taipei}$= -6.82, p< .001; $\gamma_{Singapore}$= -.89, $z_{Singapore}$= -7.35, p< .001) in this study.

According to standardized item format effect results in Table 5, we can say that strength of item format effects varies across countries and it is more prominent in the low-achieving countries. While Chile (γ^*= -1.49) has the highest item format effect, Chinese Taipei (γ^*= -.55) has the lowest item format effect.

Table 5. *Standardized item block position and item format effects.*

Countries	$\gamma^*_{\text{Block Position}}$	$\gamma^*_{\text{Item Format}}$
Chile	-.13***	-1.49***
Qatar	-.12***	-1.33***
Malaysia	-.09***	-1.05***
Republic of Korea	-.04***	-.64***
Chinese Taipei	-.02***	-.55***
Singapore	-.03***	-.59***

Note. $\gamma^*_{\text{Block Position}}$: Standardized Item Block Position effect; $\gamma^*_{\text{Item Format}}$: Standardized Item Format Effect; ***p<.001

The negative item format effects in this study imply that students' probability of giving a correct response to an item decreased when the items were CR. For instance, in terms of the probability of correct responses, this would result in a decrease of .32 in Chile and a decrease of .13 in Chinese Taipei for a student with average ability (θ= 0) in case of an item with average difficulty (β= 0) (Wu *et al.*, 2019).The graphs in Figure 1 configure this result, for all countries and for all block positions, in that the predicted probabilities of the MC items are higher than those of the CR items.

4. DISCUSSION and CONCLUSION

In this study, item block position and item format effect on eTIMSS 2019 eighth grade mathematics item easiness parameters in the low- and high-achieving countries were investigated within the scope of EIRT. The LLTM was used in data analyses. The eighth grade mathematics item responses from eTIMSS 2019 in the low-achieving countries, namely Chile, Qatar, and Malaysia and the high-achieving countries, namely South Korea, Chinese Taipei, and Singapore were used. The LLTM results of this study showed that item block position effects of the low-and high-achieving countries were negative and statistically significant. Our study confirmed the previous research studies (Deeber & Janssen, 2013; Deeber *et al.*, 2014; Demirkol & Kelecioğlu, 2022; Hartig & Buchholz, 2012; Le, 2007; Liu *et al.*, 2024; Nagy *et al.*, 2018; Wu *et al.*, 2019; Yoo, 2020) which found negative item position effect in TIMSS or PISA. We can say that items became more difficult when administered in later positions in eTIMSS 2019 eighth grade mathematics assessment. One of the explanations for negative item position effect is fatigue effect. In TIMSS assessment, eighth grade students have 45 minutes for taking Part 1 (Block 1 and Block 2) and 45 minutes for Part 2 (Block 3 and Block 4) of the assessment. Eighth grade students may feel more tired, less motivated, and rushed at the last part of the assessment (Liu *et al.*, 2024).

In addition, as voiced in previous researches, in current study, the negative item position effect was stronger in the low-achieving countries (Debeer *et al.*, 2014; Hartig & Buchholz, 2012; Wu *et al*., 2019). As Wu *et al*. (2019) stated, lower performance in those countries may be because of their students' being more impacted by item position effects. In a similar vein, Hoheninn *et al*. (2011) pointed out that "if this effect [fatigue effect] is detected in a data set, it has to be ensured that it is not caused simply by a speed effect, meaning that examinees had not enough

testing time to reach the last items." (p. 498). In a very recent study, Zheng *et al*. (2023) examined test-taking behaviors by using data from eTIMSS 2019 eighth grade mathematics of the USA, England, Singapore, and the United Arab Emirates. They found that existence or prevalence of disengaged students may not be directly related to country performance, and there is not a linear relationship between achievement and speediness at the country level. However, their study is limited to only two mathematics blocks (ME01 and ME02) which appeared in the first part of the Booklet 1. They also stated that when blocks are in a different position or together with some other blocks, results may differ. Debeer and Janssen (2013) also emphasized that "item-position effects (especially "fatigue" effects) should not be due to an increasing amount of non-reached items towards the end of the test." (p. 169). In TIMSS 2019, the average percent of not-reached items for Block 2 (which was at the end of the first half of each booklet before break) and Block 4 (which was the last block in the booklets) is as follows: 4.0 for Chile, 2.5 for Qatar, 1.8 for Malaysia, .3 for Republic of Korea, .2 for Chinese Taipei, and .4 for Singapore (Fishbein *et al*., 2020). We can say that the low-achieving countries of current study (Chile, Qatar, and Malaysia) have relatively higher proportion of not reached items than the high-achieving countries (Republic of Korea, Chinese Taipei, and Singapore). Treating not-reached items as not administered may affect item difficulty of items at the end of the test (Wu *et al*., 2019). The strong item position effect of the low-achieving countries may be partially due to relatively more proportion of not-reached items. However, it should be noted that not-reached items can also come from disengaged behaviors of examinees (Pools, 2022). In other words, examinees can omit responses without making an adequate effort. Examinees can also skip difficult items to provide more time and energy to the items that they have a higher probability of answering correctly. Therefore, disengagement may be observed in difficult items and among some lower-ability students more frequently. Disengagement can be problematic in many low-stakes assessments such as PISA and the National Assessment of Educational Progress (NAEP) where scores do not have any consequences on the examinees (Kuang & Sahin, 2023). TIMSS is one of low-stake assessments, and the source of not-reached items of TIMSS should be investigated carefully.

Another result of this study is the negative and statistically significant item format effect in all countries, and it was also found that item format effect is more prominent in the low-achieving countries (Chile, Malaysia, and Qatar). When the items were CR, the predicted probability of items would decrease. In other words, in the current study, the MC items appeared to be easier compared to the CR items in all countries. Although Hastedt and Sibberns (2005) found small differences between the scales based on the MC and CR item formats in TIMSS 1995 and 1999 math data, our finding related to item format is consistent with previous research which is related to TIMSS 2011 (Liou & Bulut, 2020) and TIMSS 2015 (İlhan *et al*., 2020). One explanation for this finding may be that probability of guessing correctly on a MC item is not trivial because MC items require choosing an answer from a set of response options. Additionally, test takers may identify a response elimination strategy, eliminate implausible distractors, and then guess from the rest of options (Martinez, 1991). By contrast, CR items require test taker to develop an answer that illustrates the knowledge required for an acceptable response (Robinson, 1993). Another explanation for this result may be educational systems of countries. CR items require higher-order thinking skills, and TIMSS also makes it possible for countries to monitor and make decisions for improving their educational system. For example, Malaysian students' poor performance in TIMSS cycles was attributed largely to their lack of higher order thinking skills and accordingly, they revised mathematics and science curriculum, and begin to ask students to respond to questions that require thinking at a higher cognitive level (Kelly *et al*., 2020).

To conclude, TIMSS implements a balanced booklet design exhibiting item blocks in four different position within a booklet. Fishbein *et al*. (2020) and Tyack *et al*. (2024) reported that this counterbalancing helps to eliminate impact of item position on the item statistics in TIMSS

assessments. For the first time in 2019, TIMSS used a computer-based "eAssessment system", named eTIMSS. However, the results of this study showed that exhibiting item blocks in different positions within a booklet can still have an impact on mathematics item difficulty in eTIMSS eighth grade assessment, and it is more prominent at eTIMSS 2019 among the low-achieving countries. To make more fair comparisons across countries, item block position effect can be modeled within the GLMM framework (as in Ong and Pastor's (2022) study) to statistically control the effect of item position on ability estimates. Negative item position effect may occur due to fatigue effect, speed effect or motivation loss (Hartig & Buchholz, 2012; Hohensinn *et al*., 2011). More research should be conducted, and after investigating whether or not item position effect is caused by a construct-irrelevant variance, it can be modeled via IRT models to make fair comparisons across countries (Wu *et al*., 2019).

In this study, item position effect was examined together with its association with item difficulty. Item position effect can also affect item discrimination or both item discrimination and item difficulty (Ma & Harris, 2025). In future research, exploring the effects of item position effect on difficulty and discrimination could be more informative and contributive to the practice. This study is limited to only six countries, eighth grade mathematics assessments and computerized version of TIMSS. For generalizability of results, in future studies, item block position effect can be studied via using data from paper based TIMSS and other eTIMSS countries, science achievement, and also fourth grade assessments. Many characteristics of items and students may affect item responses. In the current study, only block position and item format were examined so in further studies, other aspects of items (for example content and cognitive domain) or characteristics of test takers (for example, motivation or anxiety) could be investigated.

The TIMSS uses both dichotomous and polytomous scored items in its assessment. This study is limited to item block position and format effects for dichotomous item response models. In further studies, item position and format effects for TIMSS assessment could be explored for both dichotomous and polytomous IRT models. Besides, whether using dichotomous Rasch model to analyze polytomous items in EIRM may cause loss or not could be investigated.

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 author.

Orcid

Neşe Öztürk Gübeş https://orcid.org/0000-0003-0179-1986

REFERENCES

Bulut O. (2021). eirm: *Explanatory item response modeling for dichotomous and polytomous item responses* [Computer software]. Available from https://github.com/okanbulut/eirm

Bulut, O., Gorgun, G., & Yildirim-Erbasli, S. (2021). Estimating explanatory extensions of dichotomous and polytomous Rasch models: The eirm package in R. *Psych, 3*(3), 308-321. https://doi.org/10.3390/psych3030023

Bulut, O., Quo, Q., & Gierl, M.J. (2017). A structural equation modeling approach for examining position effects. *Large-Scale Assessments in Education, 5*(8), 2-20. https://doi.org/10.1186/s40536-017-0042-x

Christiansen, A., & Janssen, R. (2021). Item position effects in listening but not in reading in the European survey of language competences. *Educational Assessment, Evaluation and Accountability, 33*, 49-69. https://doi.org/10.1007/s11092-020-09335-7

Cotter, K.E., Centurino, V.A.S., & Mullis, I.V.S. (2020). Developing the TIMSS 2019 mathematics and science achievement instruments. In M.O. Martin, M. von Davier, & I.V.S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 1.1 1.36). Boston College. https://timssandpirls.bc.edu/timss2019/methods/chapter-1.html

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement, 50*(2), 164-185. https://www.jstor.org/stable/24018105

Debeer, D., Buchholz, J., & Hartig, J. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39*(6), 502-523. https://doi.org/10.3102/1076998614558485

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from lme4 package in R. *Journal of Statistical Software, 39*(12), 1-28. https://doi.org/10.18637/jss.v039.i12

Demirkol, S., & Kelecioğlu, H. (2022). Investigating the effect of item position on person and item parameters: PISA 2015 Turkey sample. *Journal of Measurement and Evaluation and Psychology, 13*(1), 69-85. https://doi.org/10.21031/epod.958576

Demir, E. (2010). *The students achievement in Turkey, according to the question types used in program for international student assessment (PISA) cognitive domain tests* [Unpublished master's thesis]. Hacettepe University.

Fishbein, B., Foy, P., & Tyack, L. (2020). Reviewing the TIMSS 2019 achievement item statistics. In M. O. Martin, M. von Davier, & I.V.S. Mullis (Eds.), *Methods and Procedures: TIMSS 2019 Technical Report* (pp. 10.1-10.70). Boston College. https://timssandpirls.bc.edu/timss2019/methods/chapter-10.html

Frey, A., Hartig, J., & Rupp, A.A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39-53. https://doi.org/10.1111/j.17453992.2009.00154.x

Gonzalez, E.J., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessment. In von Davier, M. & Hastedt, D. (Eds.) *IERI Monograph Series: Issues and methodologies in Large – Scale Assessments* (Vol. 3, pp. 125-156).

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50*(3), 379-390. https://bit.ly/3aHHyGD

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54*(4), 418-431. https://psycnet.apa.org/record/2013-10658-006

Hastedt, D., & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS surveys. *Studies in Educational Evaluation, 31*(2-3), 145-161. https://doi.org/10.1016/j.stueduc.2005.05.007

He, Q., & Wheadon, C. (2013). Using the dichotomous Rasch model to analyze polytomous items. *Journal of Applied Measurement, 14*(1), 44-56. https://pubmed.ncbi.nlm.nih.gov/23442327/

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Modeling booklet effects for nonequivalent group designs in large-scale assessment. *Educational and Psychological Measurement, 75*(4), 568-584. https://doi.org/10.1177/0013164414554219

Hohensinn, C., Kubinger, K.G., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation, 17*(6), 497-509. https://doi.org/10.1080/13803611.2011.632668

Hohensinn, C., Kubinger, K.D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large -scale assessments using linear logistic model. *Psychology Science Quarterly, 50*(3), 391-402.

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

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. P. De Boeck & W. Wilson (Ed.), *Explanatory item response models: A generalized linear and nonlinear approach* (p. 189-210). New York: Springer.

Kelly, D.L., Centurino, V.A.S., Martin, M.O., & Mullis, I.V.S. (Eds.) (2020). *TIMSS 2019 Encyclopedia: Education Policy and Curriculum in Mathematics and Science*. Boston College, https://timssandpirls.bc.edu/timss2019/encyclopedia/

Klosner, N.C., & Gellman, E.K. (1973). The effect of item arrangement on classroom test performance: Implications for content validity. *Educational and Psychological Measurement, 33(2),* 413-418. https://doi.org/10.1177/001316447303300224

Kuang, H., & Sahin, F. (2023). Comparison of disengagement levels and the impact of disengagement on item parameters between PISA 2015 and PISA 2018 in the United States. *Large-Scale Assessments in Education, 11*(4), 1-31. https://doi.org/10.1186/s40536-023-00152-0

Le, L.T. (2007, July). *Effects of item positions on their difficulty and discrimination: A study in PISA science data across test language and countries* [Conference presentation]. 72nd Annual Meeting of the Psychometric Society, Tokyo, Japan. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1001&context=pisa

Leary, L.F., & Dorans, N.J. (1985). Implications for altering the context in which test item appear: A historical perspective on immediate concern. *Review of Educational Research, 55*(3), 387-413. https://www.jstor.org/stable/1170392

Liou, PY., & Bulut, O. (2020). The effects of item format and cognitive domain on students' science performance in TIMSS 2011. *Research in Science Education, 50*, 99-121. https://doi.org/10.1007/s11165-017-9682-7

Liu, J.X., Bulut, O., & Johnson, M.D. (2024). Examining position effects on students' ability and test-taking speed in the TIMSS 2019 problem-solving and inquiry tasks: A structural equation modeling approach. *Psychology International, 6*(2), 492-508. https://doi.org/10.3390/psycholint6020030

Ma, Y., & Harris, D.J. (2025). Investigating approaches to controlling item position effects in computerized adaptive tests. *Educational Measurement: Issues and Practice, 44*(1), 44 54. https://doi.org/10.1111/emip.12637

Marcq, K., Donayre, E.J.C., & Braeken, J. (2024). The role of item format in the PISA 2018 mathematics literacy assessment: A cross-country study. *Studies in Educational Evaluation, 83*, 1-14. https://doi.org/10.1016/j.stueduc.2024.101401

Martin, O.M., Mullis, I.V.S., & Foy, P. (2017). TIMSS 2019 Assessment Design. In Mullis, I.V.S., & Martin, M.O. (Eds.), *TIMSS 2019 Assessment Frameworks* (pp.80-91). Boston College. http://timssandpirls.bc.edu/timss2019/frameworks/

Martinez, E.M. (1991). A comparison of multiple -choice and constructed figural response items. *Journal of Educational Measurement, 28*(2), 131-145. https://doi.org/10.1111/j.1745-3984.1991.tb00349.x

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174. https://doi.org/10.1007/BF02296272

Meyers, J.L., Miller, G.E., & Way, W.D. (2008). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*(1), 38-60. https://doi.org/10.1080/08957340802558342

Mullis, I.V.S., & Martin, M.O. (Eds.). (2017). *TIMSS 2019 Assessment Frameworks*. Boston College. http://timssandpirls.bc.edu/timss2019/frameworks/

Mullis, I.V.S., Martin, M.O., Fishbein, B., Foy, P., & Moncaleano, S. (2021). *Findings from the TIMSS 2019 Problem Solving and Inquiry Tasks.* Boston College. https://timssandpirls.bc.edu/timss2019/psi/

Mullis, I.V.S., Martin, M.O., Foy, P., Kelly, D.L., & Fishbein, B. (2020). *TIMSS 2019 International results in mathematics and science*. Boston College. https://timssandpirls.bc.edu/timss2019/

Nagy, G., Nagengast, B., Frey, A., Becker, M., & Rose, N. (2018). A multilevel study of position effects in PISA achievement tests: student- and school-level predictors in the German tracked school system. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 422-443. https://doi.org/10.1080/0969594X.2018.1449100

Ong, T.Q., & Pastor, D.A. (2022). Uncovering the complexity of item position effects in a low-stakes testing context. *Applied Psychological Measurement, 46*(7), 571-588. https://doi.org/10.1177/01466216221108134

Özer Özkan, Y., & Özaslan, N. (2018). Student achievement in Turkey, according to question types used in PISA 2003-2012 mathematic literacy tests. *International Journal of Evaluation and Research in Education, 7*(1), 57-64. http://doi.org/10.11591/ijere.v7i1.11045

Pools, E. (2022). Not reached items: An issue of time and of test taking disengagement? The case of PISA 2015 reading data. *Applied Measurement in Education, 35*(3), 197-221. https://doi.org/10.1080/08957347.2022.2103136

Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement, 38*(7), 518-534. https://doi.org/10.1177/01466216145343

Robinson, P. (1993). The politics of multiple-choice versus free-response assessment. In R.E. Bennett & W.C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 313-323). Routledge.

Sideridis, G., Hamed, H., & Jaffari, F. (2023). The item position effects in international examinations: the roles of gender. *Frontiers Psychology*, *14*, 1-10. https://doi.org/10.3389/fpsyg.2023.1220384

Tyack, L., Fishbein, B., Bristol, J., Mao, T., & Gonzalez, G. (2024). Reviewing the TIMSS achievement data. In M. von Davier, B. Fishbein, & A. Kennedy (Eds.), *TIMSS 2023 Technical Report (Methods and Procedures)* (pp. 10.1-10.17). Boston College. https://doi.org/10.6017/lse.tpisc.timss.rs7695

Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement, 38*(7), 535-548. https://doi.org/10.1177/0146621614534955

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. http://www.jstatsoft.org/v21/i12/

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes A., Henry, L., Hester, J., Kuhn, M., Pedersen T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1-6. https://doi.org/10.21105/joss.01686

Wilson, M., & De Boeck, P. (2004). Descriptive and explanatory item response models. In P. De Boeck, & P. Wilson (Eds.), *Explanatory Item Response Models: A generalized linear and nonlinear approach* (pp. 43-74). Newyork: Springer.

Wilson, M., De Boeck, P., & Carstensen, C.H. (2008). Explanatory item response models: A brief introduction. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 91-120). Hogrefe & Huber Publishers.

Wise, L.L., Chia, W., & Park, R. (1989, March). *Item position effects for test of word knowledge and arithmetic reasoning* [Conference presentation]. The Annual Meeting of the American Educational Research Association, San Francisco, CA.

Wu, Q., Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2019). Predictors of individual performance changes related to item positions in PISA assessments. *Large-Scale Assessments in Education, 7*(5). https://doi.org/10.1186/s40536-019-0073-6

Yoo, N. (2020). *Item position and motivation effects in large -scale assessments* [Unpublished doctoral dissertation]. Columbia University.

Zheng, X., Sahin, F., Erberber, E., & Fonseca, F. (2023). Identification and cross-country comparison of students' test-taking behaviors in selected eTIMSS 2019 countries. *Large-scale Assessments in Education, 11*(32). https://doi.org/10.1186/s40536-023-00179-3