ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Araştırma Makalesi (Research Article); 2022; 13(1); 69-85

Investigating the Effect of Item Position on Person and Item Parameters: PISA 2015 Turkey Sample *

Sinem DEMİRKOL ** Hülya KELECİOĞLU ***

Abstract

Different positions of items in booklets affect the probabilities of correct answers. This effect is called the item position effect in the literature, which causes variances in the item and person parameters. The aim of this study is to investigate the item position effect within the framework of explanatory item response theory. The analyses of this research were carried out on the PISA 2015 Turkey sample, and the item position effect was examined in the domains of reading and mathematics. In addition, the effect of the item position in different item formats (open response and multiple choice) was investigated. According to the results, the item position effect decreased the probability of answering the item correctly, and this effect was higher in reading than in mathematics. Furthermore, in the domain of mathematics, open response items were affected more than multiple-choice items by the item position. In the reading domain, open response and multiple choice items were affected similarly. The results of the analysis show that there were undesirable effects of the item position, and these effects should be taken into account.

Keywords: Item position, explanatory item response theory, item format, item easiness, mathematics and reading domain, PISA 2015

Introduction

In large-scale exams, different booklets are generally used. In some of them, such as High-school Entrance Exam (LGS), Advanced Proficiency Test (AYT), and Academic Personnel and Postgraduate Education Entrance Exam (ALES), the items are located in different positions among booklets. The main aim of this practice is to prevent test takers from cheating and to increase test security. In some exams, such as Trends in International Mathematics and Science Study (TIMSS), Programme for International Student Assessment (PISA), booklets consist of both different and common items. The aim here is to ensure that more items are applied and to expand the scope of the exam. With the developing technology, computer-based test applications have started to become very popular in recent years. In computer-based tests, the items can be located in different positions. In addition, items can be applied in different positions according to the test taker's ability level in computerized adaptive testing applications. The scores obtained from all these exams are considered as the maximum performance level of the test takers (Goff & Ackerman, 1992). Therefore, it is very important that the scores serve the intended goal.

It is assumed that the use of booklets has no effect on the response behavior of test takers or this effect is negligible (Albano, 2013; Asseburg & Frey, 2013). In other words, it is accepted that the answers given by the test takers to the items are independent of the booklet selection. Violation of this assumption is a source of undesirable variability that is not indicated by the ability level of test takers and causes context effects. Context effect was defined by Wainer and Kiely (1987), as "any influence or interpretation that an item may acquire purely as a result of its relationship to the other items making up a specific test" (p. 187). As the test and item characteristics change, (e.g., the length of the test

* The present study is a part of PhD Thesis conducted under the supervision of Prof. Dr. Hülya KELECİOĞLU and prepared by Sinem DEMİRKOL.

** PhD. Student, Hacettepe University, Faculty of Education, Ankara-Turkey, dmrklsinem@gmail.com, ORCID ID: 0000-0002-9526-6156
*** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934

To cite this article:
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 in Education and Psychology, 13*(1), 69-85. https://doi.org/10.21031/epod.958576

Received: 28.06.2021
Accepted: 12.12.2021

form, the type of item, the order of domains, the positions of the items, the cognitive levels of items, ordering the items according to their difficulty levels), the context of the item in the test also changes (Leary & Dorans,1985).

MacNicol (1956) used three different test forms in which the items were ordered from easy to hard, hard to easy, and at random order. According to the results, the easy to hard item orders were significantly more easy than the hard to easy, while the easy to hard item orders were not significantly different from the random order. Sax and Carr (1962) changed the difficulty levels of the items preceding a target item and stated that this difference had a significant effect on the difficulty of the target item. Smouse and Munz (1968) investigated the relationship between the arrangement of items according to difficulty levels and the anxiety level of test takers and stated that there were significant interactions between these two variables. Wise et al. (1989) found that item position and other context effects affected low-achieving test takers more than high-achieving test takers. Rose et al. (2019) examined the domain order and item position effects by using the data of a multidimensional computerized test, and they found that the domain order effect had a significant effect. Albano et al. (2020) found that different content order methods affected item and test statistics.

The definition of context effects as systematic effects on the response behavior of test takers indicates that there are different effects depending on the item characteristics among the booklets. A well-documented one of these effects is the item position. The item position refers to the effect of an item in different positions among the booklets on item and person parameters (Brennan, 1992; Wainer & Kiely, 1987). When examining the item position from the perspective of item difficulty, the difficulty of the item may increase or decrease towards the end of the test. Therefore, item position needs to be investigated and, if necessary, incorporated into the measurement models. It may also be useful for testing applications to consider the possible effects of item position. For example, if the effects of item position on ability estimation are known, the maximum test length that will affect the ability estimation can be determined.

In most of the international exams, item and person parameters are analyzed using the item response theory (IRT) models. In order to use IRT models, it must provide the assumption of local independence. Position of the item has an effect on the item parameters and probability of correct answer (Bulut et al., 2017; Debeer & Janssen, 2013; Le, 2007; Nagy et al., 2018; Weirich et al., 2016). Therefore, the item position effect causes bias in item parameters and violates the assumption of local independence (Albano, 2013). Furthermore, in equating and linking studies using common items, there is an assumption of invariance of item parameters among booklets. This assumption is the core of the linking process based on common items (Cook & Petersen, 1987; Kolen & Brennan, 2004). The item position effect violates this assumption and may cause linking bias (Debeer & Janssen, 2013; Meyers et al., 2009).

In general, item position has two opposing effects on item difficulty. These effects are interpreted as practice (learning) and fatigue effects. The decrease in item difficulty towards the end of the test is interpreted as a practice or learning effect. This effect may be due to test takers becoming more acquainted with the test material or the format. The increase in item difficulty towards the end of the test is interpreted as the fatigue effect. This effect may occur due to a decrease in the motivation levels of test takers towards the end of the test, an increase in fatigue, an increase in anxiety level, or distraction (Kingston & Dorans, 1982).

The question of whether item position affects the measured traits has a long history. Mollenkopf (1950) stated that the different positions of the items in the mathematics and verbal subtest caused bias in the item difficulty. Guertin (1954) found that the item position effect in the arithmetic subtest caused a significant change in performance. Hambleton and Traub (1974) investigated the number of correct answers in the two different test forms, in which the item order changed from easy-to-difficult and difficult-to-easy. They found that the number of correct answers was quite high in the test form ordered in increasing difficulty. Whitely and Dawis (1976) examined the effect of item position on item difficulty and found that item position caused variation in item difficulty.

Research on item position effects has focused on three goals. First, to examine the bias in item parameter estimates (Debeer & Janssen, 2013; Meyers et al., 2009), second, to develop and evaluate test designs that are expected to minimize this bias (Frey & Bernhardt, 2012; Frey et al., 2009; Gonzalez & Rutkowski, 2010; Hecht et al., 2015; Weirich et al., 2014), and the third is to develop appropriate models to estimate the item position effect (De Boeck et al., 2011; Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Janssen et al., 2004; Tuerlinckx & De Boeck, 2004).

Debeer and Janssen (2013), in their study using the PISA 2006 data set, found that the item position effect increased item difficulty. In addition, as a result of the simulation study, they stated that the position effects caused bias in the item parameter estimates. Nagy et al. (2018) found that moving towards the end of the test increased the difficulty of the items and decreased the item discrimination. In addition, it was determined that the variation of the item position among test takers was related to the reading speed and motivation levels, and the item position effect was lower in test takers with high reading speed and motivation. Le (2007) found that items were more difficult towards the end of the test. When the item discriminations were examined, small variations occurred in the cluster positions. Schweizer et al. (2009) used the Advanced Progressive Matrices (APM) test, which was developed by Raven (Raven et al., 1997), to investigate the item position effect. They found that the position of the item had positive effects, and the items were easier towards the end of the test. Kingston and Dorans (1984) stated that there were positive item position effects in their analysis using the data of the Graduate Record Exam. Hanne (2008) examined the position effect on reasoning items and found that there was no statistically significant item position effect.

Regarding the second goal, Hecht et al. (2015) and Weirich et al. (2014), in their study on test designs, determined that the bias due to item position could be minimized by applying balanced test designs. They recommended that if test forms (booklets) were created in such a way that each item was included in an equal number in each position, the bias due to item position would affect each item in the same magnitude, and therefore item position would not have a different effect on item parameters.

The third goal is to develop and evaluate models that can be used to estimate item position effects. Meyers et al. (2009), Whitely and Dawis (1976), and Yen (1980) studied the item position effect using a two-stage method. In these studies, as a first step, the difficulty of items in different positions in test forms was estimated separately for each test form, and then the differences between item difficulties were referred to as a function of item position. Bulut et al. (2017) and Nagy et al. (2018) used structural equation models to investigate item position effects in their studies. Christiansen and Janssen (2020), Debeer and Janssen (2013), Weirich et al. (2014), Wu et al. (2019) used generalized linear mixed models to investigate the item position effect in their study. This approach is also known as explanatory item response models with the addition of item and person characteristics (De Boeck & Wilson, 2004).

The position of an item can be considered as a variable that can be used to investigate whether the items' probability of answering correctly depends on the item position. Position effects can be specified as linear or non-linear effects (Trendtel & Robitzsch, 2018). In addition, if this effect is considered as a random effect among items or test takers, it can be investigated whether item positions are homogeneous or heterogeneous (Hartig & Buchholz, 2012). In other words, it can be examined whether the item position effect is fixed for all test takers or whether this effect varies among test takers (Debeer & Janssen, 2013; Nagy et al.,2018; Weirich et al., 2016). In addition, it can be examined whether the effect of item position varies in different item formats or items with different cognitive domains (Le, 2007).

Purpose of the Study

Position effects can cause variation in the probabilities of correct answers. If these effects are due to the characteristics of the test takers (motivation level, anxiety level, etc.) or the properties of the item (cognitive domain, item type, etc.), it can be considered that there is an effect other than ability on test scores. Position effects are possible sources of undesirable variations in test scores. A better understanding of the potential interaction of these effects can help avoid biased parameter estimates

in large-scale assessments. The aim of this study is to investigate the effect of the item position in different domains (reading and mathematics) and in different item formats (multiple choice and open response), and furthermore, to investigate whether this position effect varies among test takers.

Method

In this study, the effect of item position on item difficulty parameter and probability of correct answer was examined with Explanatory IRT models.

Working Group

The working group of the research consists of 2418 (50.5% female, 49.5% male) students who participated in the PISA 2015 Turkey sample and responded to the items in the domain of reading, and 2373 (50.6% female, 49.4% male) students who responded to the items in the mathematics domain.

Data Collection Methods

PISA takes place every three years. In each cycle, a different domain is designated as the main domain. The main domain in PISA 2015 is science. However, for the first time in 2015, in order to diminish the possible potential of systematic measurement errors due to the incomplete measurement of the scope, the number of items in minor domains (reading and mathematics) is increased. These changes strengthened the structural scope of the minor domain cycles in the PISA 2015 and introduced an innovative approach in which each of the science, reading, and mathematical literacy domains can be considered as the main domain (Organisation for Economic Co-operation and Development [OECD], 2017).

A total of 66 main booklets were used in PISA 2015 Turkey application. These booklets were composed with different combinations of clusters in reading (R1-R6), mathematics (M1-M6), science (S1-S12), and collaborative problem solving (C1-C3). There are thirty different test forms that combine two of the four areas, and 88% of students take one of these forms. These 30 forms provided strong pairwise covariance information between science and each of the three other domains. Thirty-six additional forms provided covariance information among the three minor domains. 12% of students received one of these forms (OECD, 2017). In the study, the effect of item position was investigated in the domains of mathematics and reading, so 60 booklets containing the domains of mathematics and reading were used in the analysis of this study (the other six booklets do not contain items from the mathematics and reading domains). The table regarding the allocation of item clusters to the test booklets in the PISA 2015 was given in the appendix. The booklets consist of four different clusters, and the cluster positions change among booklets. The positions of the items in the clusters were fixed. For this reason, the item position variable was considered on a cluster position. The first cluster in the booklet was coded as 0, the second cluster 1, the third cluster 2, and the fourth cluster 3. In this way, the position variable was considered as a variable that takes a value between 0 and 3.

All items in the domain of reading and mathematics were included in the analysis. In the reading domain, there are 88 items, 42 of which are multiple-choice (11 complex-31 simple) and 46 of which are open-response items. In the mathematics domain, there are 69 items, 29 of which are multiple-choice (13 complex- 16 simple) and 40 open-response items. PISA used both dichotomous and partial credit scoring (OECD, 2017). In order to fit the dichotomous IRT models, partial credit items were dichotomized (Debeer & Janssen, 2013; Hartig &Buchholz, 2012; Trendtel &Robitzsch, 2018; Wu et al., 2019). Seven items in the reading domain and five items in the math domain were scored in partial credit (incorrect, partially correct, correct). The items scored in partial credit were dichotomized by scoring the full credit as *correct* (1) and all partial credits as *incorrect* (0). In this study, the same procedure as PISA used for item calibration, missing responses on omitted items (no response) were treated as incorrect and all other missing responses (not reached) were treated as not administered (Christiansen & Janssen, 2020; OECD, 2017; Trendtel & Robitzsch, 2018).

Data Analysis

The item position effect was investigated with the explanatory item response theory models. De Boeck and Wilson (2004), explain explanatory item response theory models as special cases of what are called generalized linear mixed models (GLMM). They described GLMM models as:

> Models that require a transformation in the form of a link function before the dependent variable is related to the linear predictors are called generalized linear models (GLM; McCullagh & Nelder, 1989). They are "generalized" because the freedom of a transformation is allowed before they are linear. If such a model includes a random effect, it is called a generalized linear mixed model (GLMM; Breslow & Clayton, 1993; Fahrmeir & Tutz, 2001; McCulloch & Searle, 2001) (p. 35).

Generalized linear mixed models allow the formulation of MTK models in a multi-level framework, responses to items are treated as repeated measures nested within individuals (De Boeck & Wilson, 2004).

Traditional IRT models are used to investigate the characteristics of a person (latent ability) and to analyze the items used to investigate these characteristics (Lord, 1980). Traditional IRT models are called *descriptive IRT models* because they describe persons' performance in terms of ability and item properties in terms of discrimination and difficulty according to used models (De Boeck & Wilson, 2004). Variations in person and item parameters are not considered in descriptive IRT models (Hartig & Buchholz, 2012). Within the framework of the generalized linear mixed models (GLMM) to explanatory IRT models, person and item properties are included to explain common variability across person or item parameters (De Boeck & Wilson, 2004). The Multilevel Rasch model, the latent regression Rasch model, the linear logistic test model (LLTM), and the latent regression LLTM are widely-used forms of explanatory IRT models (Desjardins & Bulut, 2018).

Analyses were started with a multilevel Rasch model (M0). Equation 1 belongs to the multilevel Rasch model. In formulated models, the probability of answering the item correctly is interpreted as logit, and the difficulty of the item is interpreted as item easiness.

$$logit\left[P\left(Y_{pi}=1\right)\right]=\theta_p+\beta_i \quad (1)$$

Where Y_{pi} is the response of person *p* to item *i*. θ_{p} is the latent ability estimation of person *p*, and β_{i} is the easiness of item *i*. Item position can be specified as a variable to explain common variability of the item difficulty. The original item difficulty parameter in the Rasch model can be decomposed to the difficulty of the item in the reference position and the effect of the item positions. The model of item position fixed effect (M1) can be formulated in Equation 2 (Debeer & Janssen, 2013).

$$logit\left[P\left(Y_{pik}=1\right)\right]=\theta_p+\beta_i+\gamma\left(k_{pi}-1\right) \quad (2)$$

Item position fixed effect was formulated by Debeer and Janssen (2013) as in Equation 2 where Y_{pik} is the response of person *p* to item *i* in position *k*. θ_{p} is the ability estimation of person *p*, β_{i} is the easiness of item *i* at the reference position. k_{pi}, is the position of item *i* that is presented to person *p*, and γ is the fixed effect representing the overall effect of item position among individuals.

In Equation 2, it is assumed that the item position effect is fixed among individuals. The formula for the model (M2) that allows the item position effect to variation among individuals is given in Equation 3.

$$logit\left[P\left(Y_{pi}=1\right)\right]=\theta_p+\beta_i+\left(\gamma+\delta_p\right)\left(k_{pi}-1\right) \quad (3)$$

δ_{p} is the random effect of item position among persons. In other words, it is the deviation of the person *p* from the general position effect $(\delta_{\mathrm{p}}\sim(0,\sigma_{\delta}^{2}))$, referred to as *persistence* (Hartig & Buchholz 2012). Defines the ability of individuals to maintain their performance during the exam. The positive correlation between individuals' ability levels and persistence indicates that students with high ability

levels tend to be less affected by item position, while a negative correlation indicates that students with high ability levels tend to be more affected by item position during the test (Hartig & Buchholz, 2012; Weirich et al.,2016; Wu et al., 2019). The sum of $(\gamma + \delta_p)$ indicates the individual variation on test performance of the item being in different positions. A positive value indicates an increase in performance, while a negative value indicates a decrease.

Model Fit Indices

The model fits of M0, M1, and M2 models were compared. The Model M1 is the more complex version of the Model M0, and the Model M2 is the more complex version of the Model M1. First, the M0 and M1 models, then the models M1 and M2 were compared. In order to compare the model data fit, the chi-square likelihood ratio test was applied, and the Akaike's information criterion (AIC) and Bayesian information criterion (BIC) values were also used. The chi-square likelihood ratio is used to compare the likelihoods between the null model and the alternative model.

$$D = -2\log\left(\frac{likelihood_{null}}{likelihood_{alternative}}\right) \quad (4)$$

$$D = deviance_{null} - deviance_{alternative} \quad (5)$$

Where $deviance = -2\log(likelihood)$. D distributes χ^2 with approximately df_{A-N} degrees of freedom. df_A is the number of parameters estimated with the alternative model, and df_N is the number of parameters estimated with the null model.

$$AIC = 2 * df + deviance \quad (6)$$

$$BIC = \log(sample\ size) * df + deviance \quad (7)$$

Multilevel models have different sample sizes for each level. In the literature, the choice of BIC sample size depends on the specific area and the type of multilevel data being modeled (McCoach & Black, 2008). Also, the selection of the sample varies according to the software packages. The eirm package used in this study uses the level-1 observations. Therefore, in this study, BIC values were obtained using the number of level-1 observations.

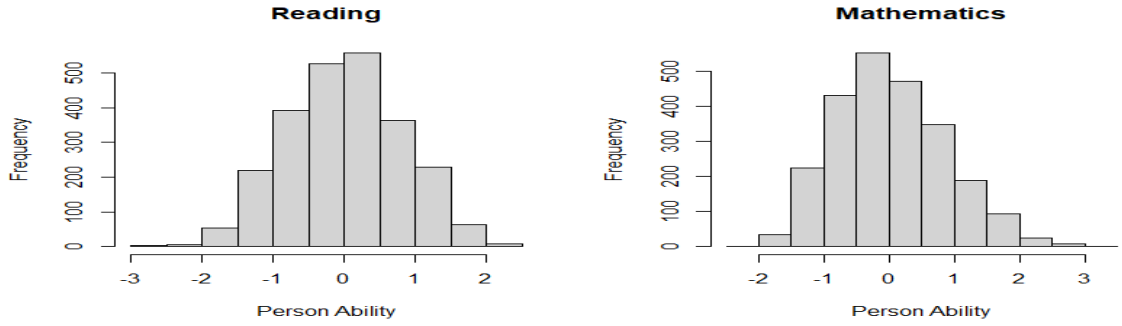
Analyses were carried out within the framework of generalized linear mixed models, with can be fitted in R using of the eirm package (Bulut, 2021), which is suitable for the analysis of explanatory item response theory models. The eirm package is essentially a wrapper for the lme4 package (Bates et al., 2014), which is capable of estimating various GLMMs using a maximum likelihood method.

Results

The first model (M0) was the multilevel Rasch model, which did not include any explanatory variables at the item and individual level. The fixed effects of the items (item easiness) and the random effects of the persons (persons' ability level) were investigated. In the domain of reading, the variance of the random effect was 0.891, and the standard deviation of this variance was 0.944; in the domain of mathematics, the variance of the random effect was 1.019, and the standard deviation of this variance was 1.01. Figure 1 shows the random effect (persons' ability level) estimated by the Rasch model in the domains of reading and mathematics, respectively.

Figure 1

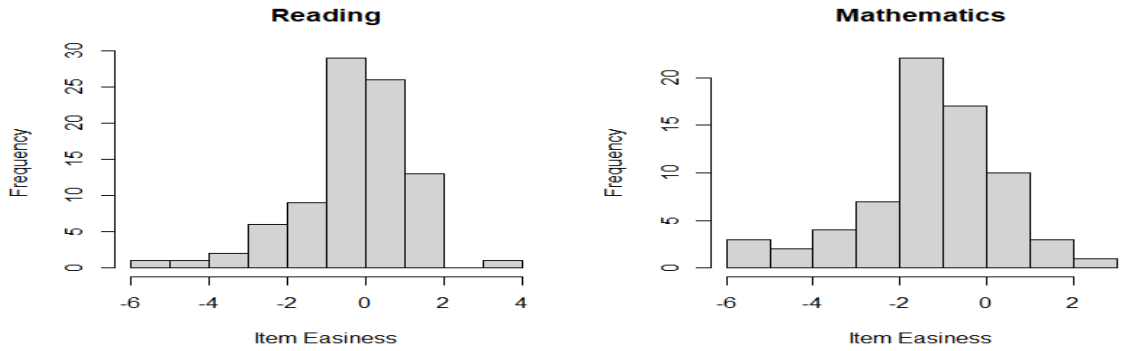
Ability Level of Persons in Reading and Mathematics



When the item fixed effects (item easiness) were examined, the easiness of the reading items ranged from -4.51 to 3.34, and the easiness of the mathematics items ranged from -5.88 to 2.45. Figure 2 shows the distribution of the easiness of items in the reading and mathematics domains, respectively.

Figure 2

Item Easiness in Reading and Mathematics



The first model (M0), was the baseline model. Models M1 and M2 were designed to investigate item position fixed effects and item position random effects, respectively. Table 1 displays the model fit indices of these models.

First, models M0 and M1 were compared. In the reading domain, model M1 indicated significantly better model-data fit than model M0. AIC, BIC, and logLik values were smaller for M1. In the domain of mathematics, AIC and logLik values were smaller for M1, while M1 BIC was slightly larger. The $\chi^2 = 89.96$ and $\chi^2 = 5.292$were statistically significant at $\alpha = .05$ ($p = .000$, $p = .021$) in reading and mathematics, respectively. This result supports the inclusion of the item position fixed effect in the model.

Table 1

Model Fit Indices

Field	Npar.	Model	AIC	BIC	logLik	Deviance	χ2
Reading	89	M0	64640	65439	-32231	64462	
	90	M1	64552	65360	-32186	64372	89.995***
	92	M2	64542	65368	-32179	64358	13.863***
Mathematics	70	M0	44564	45175	-22212	44424	
	71	M1	44561	45181	-22209	44419	5.292*
	73	M2	44538	45175	-22196	44392	27.095***

* $p < .05$; ** $p < .01$; *** $p < .001$

In order to investigate the item position random effect and correlation between latent ability and persistence, the model M2 was designed by adding the random effect of item position (random slope) to the model M1. When the models M1 and M2 were compared, for both domains, model M2 had lower AIC, BIC, and logLik values. Furthermore, the chi-square likelihood test supported that the model (M2) that included the random effect of the item position had significantly better model data fit than model M1 ($\chi^2 = 13.863, p < .000$; $\chi^2 = 27.095, p < .000$).This result was interpreted as the item position effect was not fixed and not all test takers were equally susceptible to the effect of item position. Since the model M2 fits the data better, the item position main effect was examined with the model M2. The table 2 displays the random and fixed effects of item position for the model M2.

Table 2

Fixed and Random Effects of Item Position

	Reading		Mathematics	
Parameter	Estimate	SE	Estimate	SE
Fixed effect				
Position	-0.144***	0.015	-0.065**	0.020
Random effect				
σ_θ^2		0.886		0.850
σ_δ^2		0.057		0.030
$\rho_{\theta\delta}$		-0.250		0.370

* $p < .05$; ** $p < .01$; *** $p < .001$

In the reading domain, the variance of item position was 0.057, and the standard deviation of this variance was 0.239. In the domain of mathematics, the variance of item position was 0.030, and the standard deviation of this variance was 0.174. The fact that the model (M2), which included the random effect of item position among persons, had better model-data fit than the fixed model (M1), this finding supported that the effect of item position seems to vary significantly across the test takers.

In the reading domain, the correlation between latent trait estimates and persistence was found to be -.25. This value indicated that the relationship between the ability levels of persons and the persistence was negative. This shows that the position effect was less for test takers with lower ability levels. In the domain of mathematics, the correlation between latent trait estimates and persistence item position was found to be .37. This value indicated that the relationship between the ability levels of persons and the persistence was positive. This shows that the position effect was less for test takers with higher ability levels.

Table 2 shows that the fixed effect of the item position in the reading domain was -0.14 logit, and this value was statistically significant. It needs to be kept in mind that, in this study, item easiness was estimated, not item difficulties. Moving any item cluster position by one would lead to an increase of 0.14 logits in the difficulty of the items. In the domain of mathematics, the main effect value of item position was estimated as -0.06 logit. This value was statistically significant. In the mathematics domain, moving any item cluster position by one would increase the difficulty of the item by 0.06 logits.

Table 3

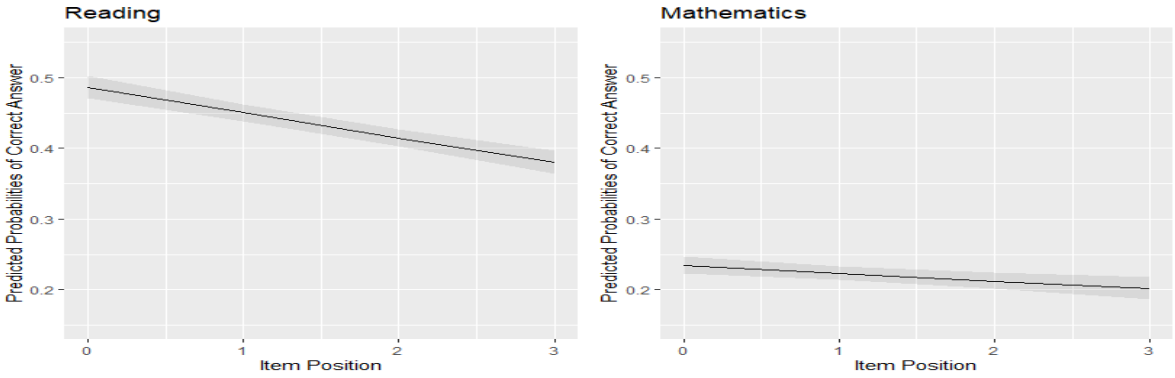
Predicted Probabilities of Correct Answer by Item Position and 95% Confidence Interval

	Reading		Mathematics	
Position	Predicted	%95 CI	Predicted	%95 CI
0	0.49	[0.47, 0.50]	0.23	[0.22, 0.25]
1	0.45	[0.44, 0.46]	0.22	[0.21, 0.23]
2	0.41	[0.40, 0.43]	0.21	[0.20, 0.22]
3	0.38	[0.36, 0.40]	0.20	[0.19, 0.22]

Table 3 shows the predicted probabilities of correct answer according to item position in the reading and mathematics domains, respectively. For example, while the predicted probability of correct answer to an item in the reference position (cluster 1) in the reading domain was .49, locating the same item in a cluster later decreased this probability to .45. Likewise, while the predicted probability of correct answer to an item in the mathematics domain was .23 (cluster 1), locating the same item in a cluster later decreased this probability to .22. Figure 3 shows predicted probabilities of correct answer (with confidence intervals) according to item position in reading and mathematics, respectively.

Figure 3

Predicted Probabilities of Correct Answer According to Item Position in the Reading and Mathematics



In this part of the study, it was investigated whether the effect of item position varied in multiple-choice (MC) and open-response (OR) item types. The items in the reading and mathematics domains consisted of multiple-choice and open-response questions, and the main effect of item position was investigated according to item formats. Table 4 shows the fixed effect of item position according to the item format of the items in the reading and mathematics.

Table 4

Item Position Effect in Different Item Formats

	Reading		Mathematics	
Parameter	Estimate	SE	Estimate	SE
Fixed effect				
MCPosition	-0.137***	0.017	-0.025	0.019
ORPosition	-0.123***	0.020	-0.053*	0.025

Note. MCPosition is the item position effect in multiple-choice items and OR Position is the item position effect in open-response items.
* $p < .05$; ** $p < .01$; *** $p < .001$

In the reading domain, multiple-choice items were slightly more affected by item positions than open-response items. Moving a multiple-choice item to the next cluster increased the difficulty of the item by 0.14 logits, while moving an open-response item to the next cluster increased the difficulty of the item by 0.12 logits.

In the domain of mathematics, it was found that the item position effect was not significant in multiple-choice items. Moving a multiple-choice item to the next cluster increased the difficulty of the item by 0.02, but this increase was not significant at the 0.05 level ($p = .196$). In other words, the difficulty of multiple-choice items in the domain of mathematics was not affected by the position of the item. Moving an open-response item to the next cluster increased the difficulty of the item by 0.05 logits, and this effect was statistically significant.

Table 5

Predicted Probabilities of Correct Answer and 95% Confidence Interval for Item Format

	Reading				Mathematics			
	MC		OR		MC		OR	
position	Prt.	%95CI	Prt.	%95CI	Prt.	%95CI	Prt.	%95CI
0	0.51	[0.49,0.52]	0.46	[0.44, 0.48]	0.38	[0.37, 0.40]	0.13	[0.12, 0.15]
1	0.47	[0.46, 0.49]	0.43	[0.41, 0.44]	0.38	[0.36, 0.39]	0.13	[0.12, 0.14]
2	0.44	[0.43, 0.45]	0.40	[0.38, 0.41]	0.37	[0.36, 0.38]	0.13	[0.12, 0.13]
3	0.41	[0.39, 0.42]	0.37	[0.35, 0.39]	0.36	[0.35, 0.38]	0.12	[0.11, 0.13]

Note. MC is the multiple-choice items, OR is the open-response items, and Prt. is the predicted probabilities value.

Table 5 shows that, in the reading domain, while the predicted probability of correct answers to multiple-choice items in the reference position was approximately .51, this ratio was approximately .46 for open-response items. In the reading domain for both item formats, an item moved from the reference position (1st cluster) to the last position (4th cluster). This decreased the predicted probabilities of correct answers by approximately .10. Figure 4 shows predicted probabilities of correct answers (with confidence intervals) according to item position in multiple-choice and open-response items, respectively.

Figure 4

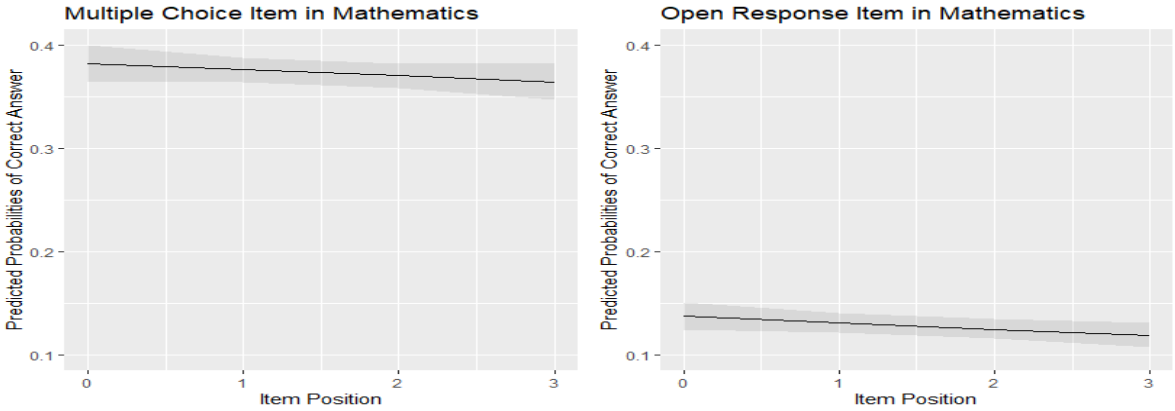
Predicted Probabilities of Correct Answer According to Item Format in the Reading

In the mathematics domain, while predicted probabilities of correct answers for multiple-choice items in the reference position were approximately .38, this ratio was approximately .13 for open-response items. In multiple-choice items, that an item moved from the reference position (1st cluster) to the last position (4th cluster) decreased the predicted probabilities of correct answers by approximately 0.02, while in open-response items, this rate was approximately .01. Figure 5 shows the predicted probabilities of correct answers (with confidence intervals) according to item position in multiple-choice and open-response items in mathematics domain, respectively.

To summarize, the effect of the item position in different item formats varied in the domains of reading and mathematics. In the domain of reading, the item position effect between multiple-choice items and open-response items was similar. In the domain of mathematics, the item position effect in multiple-choice items was not statistically significant, while the item position effect of open-response items was significant.

Figure 5

Predicted Probabilities of Correct Answer According to Item Format in the Mathematics



Discussion and Conclusion

The aim of this study is to investigate the item position effect using explanatory IRT models. Data from PISA 2015 Turkey sample were used. The effect of the item position on the probability of answering correctly and on the item difficulty was investigated in different domains and item formats.

According to the results, the item position effect was negative and significant in both domains. Furthermore, this effect was stronger in reading than in mathematics. Many studies in the literature support this finding. Wu et al. (2019) used different country samples of the 2006, 2009, and 2012 PISA data and found that the item position effect in the reading domain was stronger than in the mathematics domain. Nagy et al. (2018), using PISA 2006 Germany data, found that the effect of item position was higher in reading than mathematics across student and school levels. Hohensinn et al. (2011) stated that the effect of item position in the domain of reading was higher than in the domain of mathematics.

The negative item position effect is interpreted as fatigue effect, as mentioned before. Moving of an item from the reference position to the next position will decrease the probability of answering the item correctly. Although it is assumed that test takers are at the constant motivation level during the exam, this is not usually the case. As time passes in exams that take a long time, test takers become tired, and the items at the end become more difficult with this fatigue (Kingston & Dorans, 1984). Compared to mathematics, an item in the reading test is more likely to be answered correctly (see Table 3). Therefore, it can be thought that the later positions of the items in the reading domain may be more sensitive to the stability of the test takers' effort and attention to solving these items. However, more detailed studies are needed to reach a conclusion on this issue.

The results indicated that not all test takers were equally susceptible to the effect of item position. In the literature, there are studies comparing the fixed and random effects of the item position, as well as studies focused on only one of them (Albano, 2013; Weirich et al., 2016; Wu et al., 2019). Debeer and Janssen (2013) and Nagy et al. (2018) found that models with the random effect of item position had better model-data fit than the fixed model. In this study, the correlation between latent ability and persistence was negative in the reading domain, and it was positive in the mathematics domain. Hence, in the domain of reading, test takers with higher abilities were more affected by position effect, while test takers with higher abilities were less affected by position effect in the domain of mathematics. Debeer and Janssen (2013) found that the correlation between item position and ability estimation of individuals in reading, mathematics, and science was negative in all three domains. Weirich et al. (2016) investigated the random item position effect for classes and persons, they found that the correlation among classes was positive, but this value was negative among persons.

In this study, the effect of item position in different item formats was investigated. In the reading domain, the item position effect in the multiple-choice and open-response items was similar. However, in the domain of mathematics, the item position effect in multiple-choice items was less than in open-

response items, and this effect was not statistically significant. In his study using data from the PISA 2006 science, Le (2007) found that open response items were more affected by the item position than other item formats.

According to the results, item position affects the probabilities of correct answers and difficulty level of items. For this reason, this effect may lead to measurement error, especially in equating, linking, and item calibration studies with the assumption of item parameter invariance. Yen (1980) and Brennan (1992) found that the item position effect caused undesirable effects on the item parameters and equating results. Similarly, Kingston and Dorans (1982) found that item position had a negative effect on the Graduate Record Exam (GRE) equating forms. Kolen and Harris (1990) found that when the ACT math items were included at the end of the test, lack of motivation or fatigue had a negative effect on the equating results. Zwick (1991) attributed the scaling biases in the NAEP reading test between 1984 and 1986 to item position effect.

An item moving from the reference position (1st cluster) to the last position (4th cluster) decreased the probability of answering correctly by approximately 11% and 3% in the reading and mathematics domain, respectively. Controlling the arrangement of the items in the booklets can help ensure that the items function similarly across the booklets (Kolen & Brennan, 2004). Especially in IRT models, there is an assumption of local independence (Lord & Novick, 1968). Local independence requires that dependencies among items and persons are accounted for by parameters in the model (Albano, 2013). For example, with the Rasch model, the probability of answering the item correctly is modeled by the individual's ability and the difficulty level of the item. If a variable other than these parameters has an effect on probabilities of correct answers, it should also be included in the model; otherwise, the assumption of local independence will be violated (Lord & Novick, 1968).

When low-stake assessments are administered, such as PISA and TIMSS, the degree to which test takers give their best effort is often unclear (Wise & Kong, 2005). In national assessments, such as ALES or AYT, where important decisions are made for the future of test takers, test takers' motivation level is higher in these assessments (Wise & DeMars, 2005). The negative effect of item position is also interpreted as the fatigue effect. In other words, examinees do not have a stable cognitive level throughout the test, and their motivation and attention decrease. Therefore, in future studies, the effect of item position can be investigated by using high-stake assessment data. It is assumed that the use of booklets in an exam has no effect on answering behavior (Hahne, 2008). However, the use of booklets can lead to undesirable effects. Different context effects can be investigated in future studies. For example, test mode effects (paper-pencil, computer-based, or adaptive) or domain-order effects. PISA is an exam administered to students in the age group of 15. Whether the age of the test takers is related to the item position effect can be evaluated with a longitudinal study.

Analyses of this study were carried out with Explanatory IRT models. With these models, responses to items are treated as repeated measures nested within individuals in a multi-level framework (De Boeck & Wilson, 2004). In case of missing any covariate in Level 2, the test takers may be removed from the analysis. On the other hand, missing at level 1 only excludes data for person *j* on item *i* (Albano, 2013). In the PISA technical report, no response/omit items are defined as the test takers had an opportunity to answer the question but did not respond (OECD, 2017). Missing data due to omitted items may be related to individuals' ability levels or item characteristics. For example, studies had found that open-response items were more likely to be omitted than multiple-choice items, and difficult items were more likely to be omitted than easy items (Okumura, 2014; Rose et al., 2010). In item position effect studies using PISA data, the method of handling missing data is generally similar (Christiansen & Janssen, 2020; Trendtel & Robitzsch, 2018). In these studies, PISA scoring procedures were generally used, missing responses on omitted items were treated as incorrect, and all other missing responses were treated as not administered (OECD, 2009, 2012, 2014, 2017). Wu et al. (2019) examined different missing data handling methods by selecting countries (Albania, Argentina, Montenegro) with high-level omitted data in the PISA 2012. According to the results of this study, there was a negative position effect in both cases where the omitted data were handled as missing or incorrect. In the method in which the omitted data were handled as incorrect, the item position effect was stronger than the one handled as missing. As can be seen from the explanations, there are different

modeling approaches to handle missing data; however, modeling missing data is beyond the framework of this study. In further research, the effect of item position can be investigated by considering different missing data handling methods.

There are two types of multiple-choice items in PISA, complex multiple-choice and simple multiple-choice (OECD, 2017). In this study, simple and complex multiple-choice types were considered as multiple-choice items without making any distinction between them. In future studies, complex and simple multiple-choice items can be handled separately, and item position effects can be investigated. In this study, also the effect of the item position on the item difficulty parameter was investigated within the framework of generalized linear mixed models. For further researches, the effect of item position on item discrimination parameters can be examined by using generalized nonlinear mixed models.

The item position effect in open-response items was investigated. Open-response items of which 15% is in the reading domain, 12.5% is in the mathematics domain, are partial credit, and the other items are dichotomously scored. A multilevel model requires a consistent scoring format among all items (Hartig & Buchholz, 2012). Therefore, partial credit scored items were dichotomized in this study (Debeer & Janssen, 2013; Trendtel & Robitzsch, 2018; Wu et al., 2019) since the proportion of dichotomous items is larger than the items scored with partial credit. However, this may have affected the depth in scoring the partial credit items. In further research, the effect of item position can be investigated by using models suitable for partial credit scoring.

Declarations

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: Secondary data were used in this study. Therefore, ethical approval is not required.

References

Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement*, *50*(4), 408-426. https://doi.org/10.1111/jedm.12026

Albano, A. D., McConnell, S. R., Lease, E. M., & Cai, L. (2020). Contextual interference effects in early assessment: Evaluating the psychometric benefits of item interleaving. *Frontiers in Education, 5*. https://doi.org/10.3389/feduc.2020.00133

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*(1), 92-104. https://psycnet.apa.org/record/2013-18917-006

Bates, D., Maechler, M., Bokler, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, 5, 225-264. https://doi.org/10.1207/s15324818ame0503_4

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9-25. https://doi.org/10.2307/2290687

Bulut, O. (2021). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses* (R package version 0.3.0) [Computer software]. https://doi.org/10.5281/zenodo.4556285

Bulut, O., Quo, O., & Gierl, M. (2017). A structural equation modeling approach for examining position effects in large-scale assessments. *Large Scale in Assessments in Education*, *5*(8), 1-20. https://doi.org/10.1186/s40536-017-0042-x

Christiansen, A., & Janssen, R. (2020). Item position effects in listening but not in reading in the European Survey of Language Competences. *Educational Assessment, Evaluation and Accountability*, *33*(3), 49-69. https://doi.org/10.1007/s11092-020-09335-7

Cook, L. L., & Petersen, N. S. (1987). Problems related to the use of conventional and item response theory equating methods in less than optimal circumstances. *Applied Psychological Measurement*, *11*(1), 225-244. https://doi.org/10.1177/014662168701100302

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. Springer.

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 the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1-28. https://doi.org/10.18637/jss.v039.i12

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164-185. https://doi.org/10.1111/jedm.12009

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press.

Fahrmeir, L., & Tutz, G. (2001). *Multivariate statistical modeling based on generalized linear models* (2nd ed.). Springer.

Frey, A. & Bernhardt, R. (2012). On the importance of using balanced booklet designs in PISA. *Psychological Test and Assessment Modeling*, *54*(4), 397-417. https://www.psychologie-aktuell.com/fileadmin/download/ptam/4-2012_20121224/05_Frey.pdf

Frey, A., Hartig, J., & Rupp, 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.1745-3992.2009.00154.x

Goff, M., & Ackerman, P. L. (1992). Personality-intelligence relations: Assessment of typical intellectual engagement. *Journal of Educational Psychology*, *84*(4), 537-552. https://doi.org/10.1037/0022-0663.84.4.537

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large scale assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 3, 125-156. https://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_03_Chapter_6.pdf

Guertin, W. H. (1954). The effect of instructions and item order on the arithmetic subtest of the Wechsler-Bellevue. *Journal of Genetic Psychology*, *85*(1), 79-83. https://doi.org/10.1080/00221325.1954.10532863

Hambleton, R. K., & Traub, R. E. (1974). The effects of item order in test performance and stress. *Journal of Experimental Education*, *43*(1), 40-46. http://www.jstor.org/stable/20150989

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly*, *50*(3), 379-390. https://www.psychologie-aktuell.com/fileadmin/download/PschologyScience/3-2008/05_Hahne.pdf

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://www.proquest.com/scholarly-journals/multilevel-item-response-model-positioneffects/docview/1355923397

Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement*, *75*(6), 1021-1044. https://doi.org/10.1177/0013164415573311

Hohensinn, C., Kubinger, K., 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

Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 189-212). Springer. https://doi.org/10.1007/978-1-4757-3990-9_6

Kingston, N. M., & Dorans, N. J. (1982). *The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory* (GRE Board Professional Report GREB No. 79-12bP). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1982.tb01308.x

Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, *8*(2), 147-154. https://doi.org/10.1177/014662168400800202

Kolen, M. J., & Brennan, R. L. (2004). *Testing equating, scaling, and linking: Methods and practice*. Springer.

Kolen, M. J., & Harris, D. (1990). Comparison of item pre-equating and random groups equating using IRT and equipercentile methods. *Journal of Educational Measurement*, *27*(1), 27-29. https://doi.org/10.1111/j.1745-3984.1990.tb00732.x

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*. Paper presented at the 72nd Annual Meeting of the Psychometric Society, Tokyo. https://research.acer.edu.au/pisa/2/

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*(3), 387-413.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Erlbaum.

Lord, F. N., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Addison-Wesley.

MacNicol, K. (1956). *Effects of varying order of item difficulty in an unspeeded verbal test* (Unpublished manuscript). Educational Testing Service.

McCoach, D. B., & Black, A. C. (2008). Evaluation of model fit and adequacy. In A. A. O'Connell & D. B. McCoach (Ed.), *Multilevel modeling of educational data* (pp. 245-272). Information Age Publishing, Inc.

McCullagh, P., & NeIder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models.* Wiley.

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). 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

Mollenkopf, W. G. (1950). An experimental study of the effects on item-analysis data of changing item placement and test time limit. *Psychometrika*, *15*(3), 291-315. https://doi.org/10.1007/BF02289044

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

Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report.* Organisation for Economic Co-operation and Development. https://www.oecd.org/pisa/data/42025182.pdf

Organisation for Economic Co-operation and Development. (2012). *PISA 2009 technical report.* Organisation for Economic Co-operation and Development. http://dx.doi.org/10.1787/9789264167872-en

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 technical report*. Organisation for Economic Co-operation and Development. https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 technical report.* Organisation for Economic Co-operation and Development. https://www.oecd.org/pisa/data/2015-technical-report/

Okumura, T. (2014). Empirical differences in omission tendency and reading ability in PISA: An application of tree-based item response models. *Educational and Psychological Measurement, 74*(4), 611-626. https://doi.org/10.1177/0013164413516976

R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Raven, J. C., Raven, J., & Court, J. H. (1997). *Raven's progressive matrices and vocabulary scales*. J. C. Raven Ltd.

Rose, N., Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (Report No. RR-10-11). Educational Testing Service.

Rose, N., Nagy, G., Nagengast, B., Frey, A., & Becker, M. (2019). Modeling multiple item context effects with generalized linear mixed models. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00248

Sax, G., & Carr, A. (1962). An investigation of response sets on altered parallel forms. *Educational and Psychological Measurement, 22*(2), 371-376. https://doi.org/10.1177/001316446202200210

Schweizer, K., Schreiner, M., & Gold, A. (2009). The confirmatory investigation of APM items with loadings as a function of the position and easiness of items: A two-dimensional model of APM. *Psychology Science Quarterly*, *51*(1), 47-64. https://psycnet.apa.org/record/2009-06359-003

Smouse, A. D., & Munz, D. C. (1968). The effects of anxiety and item difficulty sequence on achievement testing scores. *Journal of Psychology*, *68*(2), 181-184. https://doi.org/10.1080/00223980.1968.10543421

Trendtel, M., Robitzsch, A. (2018). Modeling item position effects with a Bayesian item response model applied to PISA 2009-2015 data. *Psychological Test and Assessment Modeling*, *60*(2), 241-263. https://www.psychologie-aktuell.com/fileadmin/download/ptam/2-2018_20180627/06_PTAM-2-2018_Trendtel_v2.pdf

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 289-316). Springer.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*(3), 185-201. http://www.jstor.org/stable/1434630

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

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2016). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, *41*(2), 115-129. https://doi.org/10.1177/0146621616676791

Whitely, E., & Dawis, R. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement*, *36*(2), 329-337. https://doi.org/10.1177/001316447603600211

Wise, L. L., Chia, W. J., & Park, R. (1989, 27-31 March). *Item position effects for test of word knowledge and arithmetic reasoning*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1-17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

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 Assessment in Education*, *7*(5), 1-20. https://doi.org/10.1186/s40536-019-0073-6

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, *17*(4), 297-311. http://www.jstor.org/stable/1434871

Zwick, R. (1991). Effects of item order and context on estimation of NAEP reading proficiency. *Educational Measurement: Issues and Practice, 10*(3), 10-16. https://doi.org/10.1111/j.1745-3992.1991.tb00198.x

Appendix. Allocation of item clusters to test booklets in PISA 2015 Booklet

Table A

Allocation of Item Clusters to Test Booklets in PISA 2015 Booklet

Percentage of student	Booklet	Cluster 1	Cluster 2	Cluster 3	Cluster 4
	31	S	S	R01	R02
	32	S	S	R02	R03
	33	S	S	R03	R04
	34	S	S	R04	R05
	35	S	S	R05	R06A
33%	36	S	S	R06A	R01
	37	R01	R03	S	S
	38	R02	R04	S	S
	39	R03	R05	S	S
	40	R04	R06A	S	S
	41	R05	R01	S	S
	42	R06A	R02	S	S
	43	S	S	M01	M02
	44	S	S	M02	M03
	45	S	S	M03	M04
	46	S	S	M04	M05
	47	S	S	M05	M06A
33%	48	S	S	M06A	M01
	49	M01	M03	S	S
	50	M02	M04	S	S
	51	M03	M05	S	S
	52	M04	M06A	S	S
	53	M05	M01	S	S
	54	M06A	M02	S	S
	55	S	S	M01	R01
	56	S	S	R02	M02
	57	S	S	M03	R03
	58	S	S	R04	M04
	59	S	S	M05	R05
4%	60	S	S	R06A	M06A
	61	R01	M01	S	S
	62	M02	R02	S	S
	63	R03	M03	S	S
	64	M04	R04	S	S
	65	R05	M05	S	S
	66	M06A	R06A	S	S
	67	S	S	C01	M01
	68	S	S	M02	C02
	69	S	S	C03	M03
	70	S	S	M04	C03
	71	S	S	C02	M05
4%	72	S	S	M06A	C01
	73	M01	C02	S	S
	74	C03	M02	S	S
	75	M03	C01	S	S
	76	C01	M04	S	S
	77	M05	C03	S	S
	78	C02	M06A	S	S
	79	S	S	R01	C01
	80	S	S	C02	R02
	81	S	S	R03	C03
	82	S	S	C03	R04
	83	S	S	R05	C02
4%	84	S	S	C01	R06A
	85	C02	R01	S	S
	86	R02	C03	S	S
	87	C01	R03	S	S
	88	R04	C01	S	S
	89	C03	R05	S	S
	90	R06A	C02	S	S
	91	S	S	C01	C02
	92	S	S	C02	C03
22%	93	S	S	C03	C01
	94	C02	C01	S	S
	95	C03	C02	S	S
	96	C01	C03	S	S