Item Position Effects Are Moderated by Changes in Test-Taking Effort

Applied Psychological Measurement

2017, Vol. 41(2) 115–129

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621616676791

journals.sagepub.com/home/apm

Sebastian Weirich¹, Martin Hecht¹, Christiane Penk²,
Alexander Roppelt¹, and Katrin Böhme¹

Abstract

This article examines the interdependency of two context effects that are known to occur regularly in large-scale assessments: item position effects and effects of test-taking effort on the probability of correctly answering an item. A microlongitudinal design was used to measure test-taking effort over the course of a large-scale assessment of 60 min. Two components of test-taking effort were investigated: initial effort and change in effort. Both components of test-taking effort significantly affected the probability to solve an item. In addition, it was found that participants' current test-taking effort diminished considerably across the course of the test. Furthermore, a substantial linear position effect was found, which indicated that item difficulty increased during the test. This position effect varied considerably across persons. Concerning the interplay of position effects and test-taking effort, it was found that only the change in effort moderates the position effect and that persons differ with respect to this moderation effect. The consequences of these results concerning the reliability and validity of large-scale assessments are discussed.

Keywords

item position effects, generalized linear mixed models, test-taking effort, large-scale assessment

Several methodological studies have recently examined item position effects in an item response theory (IRT) framework (Albano, 2013; Debeer & Janssen, 2013; Hahne, 2008; Hohensinn et al., 2008; Hohensinn, Kubinger, Reif, Schleicher, & Khorramdel, 2011; Meyers, Miller, & Way, 2009; Qian, 2014; Schweizer, Schreiner, & Gold, 2009; Weirich, Hecht, & Böhme, 2014). Item position effects, which belong to the group of context effects (Brennan, 1992; Wainer & Kiely, 1987), imply that parameters of an item (e.g., difficulty or discrimination) in an achievement test vary according to the item's position in the booklet. Considering item position effects on item difficulty, an item administered at the end of a test often is more difficult than the same item administered at the beginning of the test. If an achievement test consists of several test forms

¹Humboldt-Universität zu Berlin, Germany

²Deutsches Institut für Internationale Pädagogische Forschung (DIPF), Berlin, Germany

Corresponding Author:

Sebastian Weirich, Institute for Educational Quality Improvement (IQB), Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany.

Email: sebastian.weirich@iqb.hu-berlin.de

that may include different items in a different order and that are to be linked to a common scale, item position effects might violate the assumption of item parameter invariance across test forms.

As the assumption of item parameter invariance is central to any linking procedure that is based on common items (Cook & Petersen, 1987; Kolen & Brennan, 2004), item position effects might induce a linking bias (Debeer & Janssen, 2013; Meyers et al., 2009). To date, research concerning item position effects has focused on three objectives: first, to examine how severely item parameter estimates may be biased (Debeer & Janssen, 2013; Meyers et al., 2009); second, to develop and evaluate balanced test designs, which are supposed to minimize this bias (Frey & Bernhardt, 2012; Frey, Hartig, & Rupp, 2009; Gonzalez & Rutkowski, 2010; Hecht, Weirich, Siegle, & Frey, 2015; Weirich et al., 2014), and third, to develop models that are suitable for estimating the effects of item position in an IRT framework (Debeer & Janssen, 2013; De Boeck et al., 2011; Hartig & Buchholz, 2012; Janssen, Schepers, & Peres, 2004; Tuerlinckx & De Boeck, 2004).

In the following, the central results regarding these three goals will be briefly reviewed. Focusing on the first goal, Meyers et al. (2009) compared the item parameters of two tests on which the overlapping items differed in their relative positions. They found that about 56% of the variance in changes in Rasch item difficulty could be attributed to changes in item position. Debeer and Janssen (2013) used data from the Programme for International Student Assessment (PISA) 2006 (Organization for Economic Cooperation and Development [OECD], 2006) to show that the difficulty of reading items increased .24 logits on average when administered one cluster position further on the test. Moreover, their simulation showed that position effects may cause a bias in item parameter estimates of the Rasch model of up to .52 logits.

Concerning the second goal, research on test designs has shown that this bias can be minimized by the application of test designs that are balanced with respect to item position (Hecht et al., 2015; Weirich et al., 2014). If several test forms (e.g., booklets) are constructed and distributed to examinees in such a manner that each item occurs at each position an equal number of times, the disturbing effect of item position is expected to influence each item to the same magnitude. Hence, there would then be no differential effect of item position on item difficulty.

The third goal addresses the development and evaluation of models that can be used to estimate item position effects. The framework of generalized linear mixed models (GLMMs; De Boeck & Wilson, 2004) has proven very useful for modeling item position effects when only 1PL estimation is of interest (Debeer & Janssen, 2013; Weirich et al., 2014). The position of an item may be specified as a predictor that can be used to examine whether the probability of success depends on it. Position effects may be specified as linear or nonlinear. Moreover, if items and persons are considered to be random effects, the distribution of items or persons can be specified to depend on position, for example, to investigate whether the effect of the position is homogeneous or heterogeneous across persons or items (Hartig & Buchholz, 2012).

Debeer and Janssen (2013) stated that very little research has addressed *why* item position effects occur. Depending on the direction of the effect, two possible explanations are usually taken into account (Hohensinn et al., 2011; Kingston & Dorans, 1984). An increase in item difficulty during the test might be interpreted as a fatigue effect or an effect of decreasing test-taking effort (e.g., potentially due to decreases in their motivation). In contrast, a decrease in item difficulty may be interpreted as a practice effect (e.g., if examinees become better acquainted with the test material). Although both explanations seem plausible at first glance, Debeer and Janssen (2013) suggested that “the *why* question” should be addressed by including person predictor variables (e.g., test-taking effort) in the model. Qian (2014), for example, analyzed position effects in a 50-min writing assessment and found that the second half of the test was more difficult than the first half. This difference, however, was less pronounced for examinees who

indicated that the test was important for them. However, item position effects conceptually imply a change in item response behavior over the course of the test. Consequently, as item position effects are inevitably related to test time, it is plausible to assume that some covariates that substantively moderate these effects also change over time. According to Debeer and Janssen, it is plausible that examinees' changes in effort during the test can also moderate item position effects, which should result in interaction effects between position and effort (Qian, 2014).

Like position effects, effects of examinees' effort on the test score may be considered another context effect. A relation between self-reported effort and test achievement was found in several studies (for an overview, see Wise & DeMars, 2005). In general, students who report higher motivation invest more effort and achieve higher scores on the test. The topic is frequently discussed as a problem of validity: A lack of test-taking effort does not only lead to a potential underestimation of achievement scores but also to a change in the construct being measured. The test score then would be comprised of two constructs: the ability and the current effort of the examinees (Eklöf, 2010); that is, test scores would not be free from construct-irrelevant variance (Messick, 1984). When interpreting the relation between self-reported motivation and test achievement, it is unclear whether highly motivated students perform better due to their higher motivation or due to their higher actual ability (Wise & DeMars, 2005). Measuring the change in test-taking effort during the course of the test in a *microlongitudinal design* allows for the examination of whether situational motivation affects current response behavior beyond the general correlation between motivation and ability.

To summarize, position effects imply a change in response behavior. If these effects are moderated by a change in test-taking effort, it can be concluded that there is an effect of effort on test performance beyond the examinees' ability. Position effects and effort are both possible sources of construct-irrelevant variance in test scores. A better understanding of the potential interplay of the two effects can help researchers to avoid biased person estimates in large-scale assessments.

To distinguish between test-taking motivation and test-taking effort, the authors adopted Freund, Kuhn, and Holling's (2011) taxonomy. Test-taking motivation is considered a multidimensional construct that also comprises facets of test-related anxiety, challenge, interest, and perceived probability of success. Test-taking effort, as one facet of test-taking motivation, is a unidimensional construct that may change during the course of the test in terms of a variable state (Wise & Smith, 2011).

Research Scope

In the present article, the authors used data collected for the *IQB-Ländervergleich 2012* (Pant et al., 2013), a German nationwide low-stakes large-scale assessment study similar to the U.S. National Assessment of Educational Progress (NAEP). The tasks and items included in this study were developed to evaluate student attainment according to the German National Educational Standards for Math and Science. Test-taking effort was measured 2 times during the test. The change in test-taking effort could only be modeled linearly using a latent growth model that reflected a decline across the course of the test. The intercept and slope of the growth model were used as estimates of initial effort and change in effort. The authors modeled position effects in a generalized linear mixed effects model framework and examined whether position effects were moderated by initial effort and change in effort. Hence, the authors investigated the interaction between position effects and test-taking effort, which have been considered only separately until now.

The following research questions are addressed:

Research Question 1: Prerequisite analysis: To what degree does the test-taking effort of examinees change during a 60-min low-stakes achievement test of scientific literacy?

Research Question 2: Do position effects occur on a 60-min low-stakes achievement test? Are examinees affected heterogeneously by position effects?

Research Question 3: Are position effects moderated by initial effort and change in effort?

The first research question is a replication of results reported by Penk and Richter (2016), who investigated the change in test-taking effort, anxiety, challenge, interest, and perceived probability of success across the course of a mathematical achievement test.

Method

Sample

In the *IQB-Ländervergleich 2012* (Pant et al., 2013), schools were randomly selected within each of the 16 German federal states. Within each school, one or two Grade 9 classes were randomly selected, according to a multistage sampling procedure. As several domains in several school types were tested, the entire design was composed of several subdesigns (Hecht, Roppelt, & Siegle, 2013). To reduce the amount of data and to simplify the analyses, the authors used data from only one domain (scientific literacy) and from several school types (secondary school, comprehensive school) representing academic as well as non-academic school tracks. The sub-sample of $N = 9,410$ ninth graders (51.5% female, average age of $M = 15.7$ years) consisted of examinees who answered the effort scale at both time points. Although the sample cannot be seen as a completely random draw from the population, weights were refrained from using in the analyses. As the sub-sample was chosen to fulfill the requirements of representativeness, no weights were needed.

Measures

As will be explained later in more detail, the test consisted of two parts, with a scheduled processing time of 60 min each and an intermission of 15 min in between. An adequate modeling of initial effort and the change in effort is only possible for the first 60 min of the test. Therefore, modeling of position effects and test-taking effort was limited to the first part of the test.

The test was comprised of 386 dichotomously scored items. Usually, up to three items belong to a common stimulus, that is, a short reading passage. In a preceding multistage development process, items were developed and selected to fit the unidimensional Rasch model. Unsuitable items were discarded so the items used in this study had been shown to fulfill the measurement standards of the Rasch model. Of the 386 items, 227 items had a closed form (i.e., multiple choice or complex multiple choice), and 159 items had an open format. Hence, the examinees were requested to write short answers to open questions, for example, “Why is a chlorine atom not converted to argon by absorption of an electron?” The scientific literacy items were used in a multiple matrix sampling design in which a subset of items (i.e., a booklet) was randomly assigned to each examinee. More specifically, the items were grouped into 31 disjoint blocks (i.e., items were nested within blocks). A single block consists of 11 to 14 items. Hence, a single examinee works on up to 78 items. The time allocated for each block was 20 min. The authors constructed 31 booklets, each consisting of six blocks, according to a Youden square design (Frey et al., 2009), which is balanced with respect to item position. Table A1 is

supplied in the online appendix and gives an overview of the used design. As common in such designs, only the block position was varied across booklets, whereas the item positions within each block remained constant. The modeling of position effects is thereby a means for estimating the magnitude by which item difficulty changes on average if items are placed, for example, at Block Position 2 instead of Block Position 1. The 31 booklets were randomly distributed to the 9,410 students, yielding 341,797 responses overall for the first part of the test.

Test-taking effort was measured with the self-reported test-taking effort scale examined in Penk, Pöhlmann, and Roppelt (2014), who investigated test-taking motivation as a multidimensional construct. Test-taking effort constitutes one facet of this construct and was measured with four items that originally stemmed from Eklöf (2010). All items ranged from 1 (*strongly disagree*) to 4 (*strongly agree*) and referred to the current test situation. All items were adopted (positive formulation only) and translated into German (Penk et al., 2014). Test-taking effort was measured at the beginning of the first part of the test (i.e., before examinees worked on the first part of the test; t1), and at the beginning of the second part of the test (i.e., after 60 min of test processing plus 15 min of intermission; t2). The scale reliabilities of the four effort items yielded satisfactory results. Coefficient alphas (Cronbach, 1951), reliability coefficients, of .82 and .86 at t1 and t2, respectively, were found.

Models

As a first step, the change in test-taking effort was modeled using in Mplus, Version 7.11 (Muthén & Muthén, 1998-2012). The authors estimated the latent difference between the two time points, using a curves-of-factors model (Duncan, Duncan, & Strycker, 2006), which assumes equal intercepts and equal factor loadings across the two time points. This corresponds to the assumption of strong measurement invariance. With only two time points, the change in test-taking effort can only be modeled linearly. Thus, the model contained two latent variables to represent change: an intercept (initial effort at the beginning of the test) and a linear change component (change in effort). To derive point estimates for each person, factor scores were generated. As effort was measured preceding the test (i.e., previous to the first block position) and after the first part of the test (i.e., after the third block position), the authors conceptualized the factor score of the intercept as a measure of initial effort and the factor score of the slope as a measure of change in effort. The variance of the intercept (initial effort) was standardized to be 1 to simplify the interpretation of parameter estimates in the following GLMMs.

As a second step, the R package *lme4* (Bates, Maechler, Bolker, & Walker, 2014; R Core Team, 2014) was deployed to specify two nested GLMMs to model linear item position effects that depended on initial effort and change in effort. Model 1 assumed a linear effect of item position. Model 2 additionally used initial effort and change in effort from the latent linear growth model (LLGM) as predictors as well as two-way interactions of position and effort. Basically, these models are extensions of the multilevel item response model for item position effects (Debeer, Buchholz, Hartig, & Janssen, 2014; Hartig & Buchholz, 2012).

To build the two models, one may start from the one parameter logistic model with a linear item position effect for an assessment with P test takers and I binary test items administered in K positions (Hartig & Buchholz, 2012). As most large-scale assessment studies use a sampling procedure that leads to a hierarchical data structure where students are nested within classes, Debeer et al. (2014; equation 2) suggested to use a multilevel extension of this model. Now one may modify this model in three ways. First, the Level-3 variable is class instead of school, as the sampling procedure used in the study allowed to sample two different classes from the same school. Second, add a fixed-effect parameter for school type to account for possible ability

differences between two school types: academic track schools and non-academic track schools. And third, model random item effects instead of fixed-item effects.

The data are structured in three levels: the responses (Level 1) are nested within persons and items. Persons and items (Level 2) are partially crossed. Persons are nested in classes (Level 3). η_{ipck} is the linear GLMM component for person p in class c solving item i at position k . Hence, $\eta_{ipck} = \text{logit}(Y_{ipck} = 1 | \theta_p, \delta_p)$. At Level 1, this component is predicted by block position k .

Level 1:

$$\eta_{ipck} = \alpha_{ipc} + \gamma_{pc}(k - 1).$$

Here, α_{ipc} is the intercept for person p in class c solving item i . γ_{pc} is the position effect of person p in class c . At Level 2, α_{ipc} and γ_{pc} are further decomposed.

Level 2:

$$\alpha_{ipc} = \alpha_c + \theta_{pc} + \beta_i,$$

$$\gamma_{pc} = \gamma_c + \delta_{pc}.$$

α_c is the class-specific intercept. θ_{pc} is the ability of person p in class c , or more specifically, the deviation of person p from the mean ability of class c . β_i is the easiness of item i . γ_c is the average position effect in class c , and δ_{pc} is the individual deviation of person p from the class-specific position effect γ_c . Similar to Debeer et al. (2014), θ_{pc} and δ_{pc} are random effects and follow a bivariate normal distribution with $\begin{pmatrix} \theta_c \\ \delta_c \end{pmatrix} \sim MVN(\mathbf{0}, \Sigma_{\theta pc})$, where

$\Sigma_{\theta pc} = \begin{pmatrix} \sigma_{\theta pc}^2 & \sigma_{\theta \delta pc}^2 \\ \sigma_{\theta \delta pc}^2 & \sigma_{\delta pc}^2 \end{pmatrix}$. $\sigma_{\theta pc}^2$ is the variance of ability between persons, and $\sigma_{\delta pc}^2$ is the variance of position-specific deviations between persons. β_i is specified as a random effect, following a normal distribution with $\beta_i \sim N(0, \sigma_\beta^2)$. At Level 3, α_c and γ_c are further specified.

Level 3:

$$\alpha_c = \alpha + \theta_c + v_1 \cdot T_c,$$

$$\gamma_c = \gamma + \delta_c.$$

α is the overall intercept, θ_c is the ability of class c , and v_1 is the effect of school type T_c (academic track vs. non-academic track). Weighted effect codes were used with code 1 for academic track schools and $-1 \cdot N_{academic}/N_{non-academic} = -0.60$ for non-academic track schools. γ is the average position effect, and δ_c is the class-specific deviation from the global position effect. θ_c and δ_c follow a bivariate normal distribution with $\begin{pmatrix} \theta_c \\ \delta_c \end{pmatrix} \sim MVN(\mathbf{0}, \Sigma_{\theta c})$, where $\Sigma_{\theta c} = \begin{pmatrix} \sigma_{\theta c}^2 & \sigma_{\theta \delta c}^2 \\ \sigma_{\theta \delta c}^2 & \sigma_{\delta c}^2 \end{pmatrix}$. $\sigma_{\theta c}^2$ is the variance of ability between classes, and $\sigma_{\delta c}^2$ is the variance of class-specific deviations from the global position effect. Model 1 finally reads as,

$$\text{logit}(Y_{ipck} = 1 | \theta_p, \delta_p) = \alpha + \theta_c + v_1 \cdot T_c + \theta_{pc} + \beta_i + (\gamma + \delta_c + \delta_{pc}) \cdot (k - 1). \quad (1)$$

The parameters that correspond to Research Question 2 are γ (position effect) and δ_{pc} (individual deviation from position effect). The authors used the “+parametrization,” which leads to an interpretation of β_i as the easiness of an item i administered at the first position of the test. In *lme4*, Model 1 was specified as

$Y \sim 1 + \text{position} + \text{track} + (1 + \text{position} \mid \text{class}) + (1 + \text{position} \mid \text{person}) + (1 \mid \text{item})$

With the exception of Y and position, all variables were factors, with 9,410 levels for person, 1,218 for class, 386 for item, and 2 levels for school track, respectively. Position was a numeric variable with value 0 for Position 1, value 1 for Position 2, and value 2 for Position 3.

Model 2 used the initial effort I_{pc} and change in effort C_{pc} from the LLGM as fixed-effects predictors at Level 2—both vary between persons but not within persons. In addition, the authors modeled two-way interactions between position and initial effort and between position and change in effort. In the following, Model 2 is described stepwise for each level.

Level 1:

$$\eta_{ipck} = \alpha_{ipc} + \gamma_{pc}(k - 1).$$

At Level 1, Models 1 and 2 do not differ from each other. At Level 2, α_{ipc} and γ_p are further specified.

Level 2:

$$\begin{aligned}\alpha_{ipc} &= \alpha_c + \theta_{pc} + \beta_i + \zeta \cdot I_{pc} + \varphi \cdot (k - 1) \cdot I_{pc} + \kappa \cdot C_{pc} + \lambda \cdot (k - 1) \cdot C_{pc}, \\ \gamma_{pc} &= \gamma_c + \delta_{pc}.\end{aligned}$$

In addition to Model 1, ζ is the fixed effect of initial effort, φ is the interaction of position and initial effort, κ is the mean effect of change in effort, and λ is the interaction of position and change in effort. θ_{pc} and δ_{pc} are random effects and follow a multivariate normal distribution

with $\begin{pmatrix} \theta_{pc} \\ \delta_{pc} \end{pmatrix} \sim MVN(\mathbf{0}, \Sigma_{\theta pc})$, where $\Sigma_{\theta pc} = \begin{pmatrix} \sigma_{\theta pc}^2 & \\ \sigma_{\theta \delta pc} & \sigma_{\delta pc}^2 \end{pmatrix}$.

At Level 3, α_c and γ_c are further specified.

Level 3:

$$\begin{aligned}\alpha_c &= \alpha + \theta_c + v_1 \cdot T_c, \\ \gamma_c &= \gamma + \delta_c.\end{aligned}$$

Model 2 finally reads as

$$\begin{aligned}\text{logit}(Y_{cpik} = 1 \mid \theta_p, \delta_p) &= \alpha + \theta_c + \theta_{pc} + \beta_i + v_1 \cdot T_c + \zeta \cdot I_{pc} + \\ &\kappa \cdot C_{pc} + (\gamma + \delta_c + \delta_{pc} + \varphi \cdot I_{pc} + \lambda \cdot C_{pc}) \cdot (k - 1).\end{aligned}\quad (2)$$

With respect to Research Question 3, the parameters of the two moderation effects are λ for the moderation of the position effect by change in effort and φ for the moderation of the position effect by initial effort. In *lme4*, Model 2 was specified as,

$Y \sim 1 + \text{position} + \text{track} + \text{position} * I + \text{position} * C +$
 $(1 + \text{position} \mid \text{class}) + (1 + \text{position} \mid \text{person}) + (1 \mid \text{item})$

Results

Table 1 and Online Appendix Table A2 list the results for the first research question. Table A2 displays the factor loadings of the LLGM assuming strong invariance, that is, intercept and

Table 1. Results of the Latent Linear Growth Model.

Parameter	Coefficients		
	Estimate	SE	<i>p</i>
Intercept (initial effort)	0.000		
Change	−0.389	0.011	<.001
Var (intercept)	1.000		
Var (change)	0.399	0.017	<.001
Cor (intercept and change)	−0.039	0.016	<.05

factor loading were restricted to be equal across measurement occasions. The model fit was acceptable (comparative fit index [CFI] = 0.992; Tucker–Lewis index [TLI] = 0.989).

Table 1 displays the coefficients of the LLGM. On average, test-taking effort diminished across the course of the test. The variance of the intercept (the initial effort) was standardized to be 1. Hence, a slope coefficient of −0.389 indicates that from *t*₁ to *t*₂ (i.e., from the first block position to the third block position), effort on average decreases by 38.9% of the standard deviation of the initial effort, which may be considered a moderate decline in examinees' effort.

Moreover, the change variance was quite small compared with the intercept variance: Whereas the base level of effort was rather heterogeneous across examinees, the decrease in effort was more homogeneous across examinees. The descriptive results of the factor scores indicated that only 15.1% of the total sample showed an increase in test-taking effort over the course of the test.

The results of the two GLMMs are listed in Tables 2 and 3. For the sake of clarity, the fixed-effect estimates of both models are listed in Table 2, and the random effect estimates are listed in Table 3. In Model 1, the authors specified only a linear item position effect to address the second research question. The effect of school type (academic vs. non-academic) was additionally specified to account for possible ability discrepancies between school types and yielded that the mean ability of students from academic track is $0.739 + 0.60 \times 0.739 = 1.18$ logits higher than the mean ability of students from non-academic school tracks. The regression coefficient γ refers to the linear position effect and represents the mean change in the logit for a correct response when block position increases by 1. The negative coefficient ($\gamma = -0.189$) indicates that the logit for a correct response was on average 0.189 lower when block position increases from 1 to 2, for example.

Considering the random effects (Table 3), the authors had to separate between the between-class part and the between-person part for ability. Considering the between-class part, the variation of the class-specific deviations from the average position effect are small ($SD = 0.038$), whereas the variation of the individual deviations from the average position effect are more substantial ($SD = 0.219$). In other words, not all examinees were influenced by the position effect to the same extent.

Comparing the correlations between ability and the deviations from the average position effect, oppositional results for classes and persons were found. The positive correlation $\sigma_{\theta\delta_{pc}} = .65$ indicates that classes with higher average ability are less affected by the position effect. However, on the between-person level, this correlation is $\sigma_{\theta\delta_c} = -.25$; thus, individuals with higher ability (in relation to the average class ability) seem more affected by the position effect. This finding replicates results reported in Debeer et al. (2014).

Table 2. Fixed Effects and Model Fit for the Two GLMMs.

Parameter	Model 1			Model 2		
	Estimate	SE	<i>p</i>	Estimate	SE	<i>p</i>
Fixed effects						
Intercept (α)	0.146	0.065	<.001	0.212	0.065	<.001
Academic track (v_1)	0.739	0.019	<.001	0.710	0.018	<.001
Position (λ)	−0.189	0.006	<.001	−0.161	0.008	<.001
Initial effort (ζ)				0.241	0.011	<.001
Change in effort (κ)				0.177	0.035	<.001
Initial Effort $\times$ Position (φ)				0.010	0.006	.084
Change in Effort $\times$ Position (λ)				0.084	0.019	<.001
Model fit						
AIC		371,529			370,777	
BIC		371,637			370,927	
Deviance		371,509			370,749	

Note. GLMM = generalized linear mixed models; AIC = Akaike information criterion; BIC = Bayesian information criterion.

To tackle Research Question 3 of whether effort moderates position effect, initial effort and change in effort from the latent difference model were used as additional predictors in Model 2. The main effects of initial effort ($\zeta = 0.241$) and change in effort ($\kappa = 0.177$) were positive. Students who report higher values on the initial effort items also show a higher test performance. Likewise, students with higher values on the change in effort items (i.e., students with lower decrease in effort) show higher test performance.

When the initial effort increases by 1 standard deviation, the logit for correctly answering an item increases by 0.241 on average. This is in line with previous studies (Wise & DeMars, 2005) that indicated a positive relation between motivation and test performance.

Model 2 also parametrizes two-way interactions between position and both initial effort and change in effort. These two interaction effects describe the moderation of the position effect by effort. The interaction of initial effort and position ($\varphi = 0.010$) was not significant ($p = .084$), that is, the position effect does not depend on the self-reported initial effort. The interaction of change in effort and position was significant ($\lambda = 0.084$, $p < .001$), that is, the mean position effect is less pronounced for students with a lower decrease in effort.

Table 2 also includes the model fit for both GLMMs. The likelihood ratio test of Model 2 versus Model 1 indicated that Model 2 provided a better fit to the data: $\chi^2 = 760.34$, $df = 4$, $p < .001$.

Discussion

The occurrence of position effects in large-scale assessments of student achievement is a well-documented phenomenon. Although differences in position effects across school types or countries have already been demonstrated (Debeer et al., 2014), research exploring why these position effects occur has been lacking. The authors began to fill this gap by investigating the interdependence of position effects and test-taking effort. To this end, three consecutive models were estimated. The purpose of the first model was to examine the change in effort across the course of the test. The results of this latent difference model indicated that effort diminished substantially during the first part of the test.

Table 3. Random Effects for the Two GLMMs.

Parameter	Model 1			Model 2		
	Var	SD	Cor	Var	SD	Cor
Between-class part (Level 3)						
$\sigma_{\theta c}^2$	0.146	0.382		0.111	0.334	
$\sigma_{\delta c}^2$	0.001	0.038		0.001	0.030	
$\sigma_{\theta \delta c}$			0.65			0.80
Between-person part (Level 2)						
$\sigma_{\theta pc}^2$	0.513	0.716		0.479	0.691	
$\sigma_{\delta pc}^2$	0.048	0.219		0.048	0.218	
$\sigma_{\theta \delta pc}$			-0.25			-0.27
Random Effects (items)						
σ_{β}^2	1.528	1.236		1.525	1.235	

Note. GLMM = generalized linear mixed model.

In a second step, the authors specified two GLMMs that contained item responses (correct vs. incorrect) as dependent variable. Both GLMMs account for clustered data structure (students nested in classes) and parametrize individual deviations from the average position effect. The first GLMM (Model 1) indicated that persons are differently affected by position effects. Heterogeneity occurred at the person level rather than at the class level. Hence, it would be reasonable to search for person variables that may moderate position effects. In the second GLMM (Model 2), the authors test whether position effects are moderated by initial effort and change in effort. Indeed, both effort variables were identified as significantly contributing to the probability of solving an item. The interactions between (a) position and initial effort and (b) position and change in effort were incorporated into the second GLMM, whereas only the second interaction effect was significant: Position effects were less pronounced for students with higher change in effort (i.e., students with a less pronounced decline in effort). The conclusions that can be made from this study are discussed in the following sections.

Position Effects Are Only Partially due to Effort

The results of the second GLMM (Model 2) indicate that position effects do not vanish even if controlled for initial effort and change in effort. Hence, even in an “ideal” group of persistently highly motivated students, position effects remain. So it is reasonable to assume that position effects have multiple causes: fatigue or exhaustion, increasing listlessness, and decreasing effort.

Examinees Differ Less in (Decreasing) Effort, But Substantially in Position Effects

Results showed that the variance in the change in effort (0.399) is smaller than the variance of the initial effort (1.000): persons are heterogeneous in their initial effort, but more homogeneous in their decreasing effort. In contrast, the individual deviations from the average position effect seen in both GLMMs are substantial. Differences in the decline of effort cannot completely explain why some examinees deviate from the average position effect. The residuals are correlated, which is in line with previous research. For example, Debeer et al. (2014) found a positive correlation between the change in performance that was referred to as persistence and ability at

school level. Similar results were found for the class level in the current study: in classes with a lower mean ability, position effects were much more pronounced.

Practical Implications: Validity

Whether the results of the current study may be considered as a threat to validity depends on whether the variation in test scores due to inter-individual changes in effort or motivation is regarded as construct-relevant or construct-irrelevant variance. The results underline that this theoretical differentiation is relevant to make a statement about validity in commonly used low-stakes large-scale assessment. The test scores virtually do reflect the intended ability construct as well as the persistent commitment of the students and their willingness to solve items in a given test. This connects to a well-known discussion regarding the distinction between competence and performance. Shohamy (1996) stated, “. . . there is a difference between competence and performance, where competence equals ability equals trait, while performance refers to the actual execution of tasks” (p. 148). The authors assume that the actual execution of tasks is influenced by motivation and effort. Imagine two equally competent students taking a test intended to measure one specific cognitive construct. One of the students is highly motivated and invests constantly huge effort, while the other is not. If the willingness to cope successfully with the tasks differs between these two students, the actual test scores and therefore the inferences regarding the levels of ability would vary due to the differences in the invested effort and not because of differing ability. The varying motivation would threaten the validity of the measurement of the students’ achievement if the underlying competence construct excludes volitional components. The interpretation of the test scores of the two students then would not be equally valid.

Nevertheless, the implications of the finding in the context of large-scale assessment need further exploration. It would be particularly interesting to determine how severely such effects practically affect test scores in commonly used IRT models such as Rasch or 2PL/3PL models that ignore motivational context effects.

Practical Implications: Educational Assessment for the Purpose of System Monitoring

It is known from simulation studies that item parameter estimates can be biased by position effects (Meyers et al., 2009). However, position-balanced test designs allow for the estimation of unbiased parameters even if position effects are not parametrized in the measurement model (Hecht et al., 2015). However, the principle of balancing is based on the assumption that context effects are not related to the effects of items or persons. The results of this study question this assumption: A context effect was found to be related to performance and effort.

This may be relevant in the context of educational assessment for the purpose of system monitoring. The results of educational assessments on an aggregated level provide information for schools, states, or countries and serve evaluation purposes. As stated before, position effects were found to be related to performance effort. This may lead to biased estimates of ability differences between different groups of students. For example, the difference in ability between intermediate and academic school track students might be overestimated if position effects are more pronounced in intermediate school tracks due to a stronger decline in effort. Test scores of the low-ability group then are more affected by position effects, and the low-ability group is disadvantaged by more pronounced position effects. So, the specific extent of this potential bias might be investigated in further simulation studies. This study was exploratory and yielded evidence about the extent to which position effects vary between high- versus low-ability groups and highly versus lowly motivated students.

Limitations

Several limitations should be mentioned concerning this study. First, test-taking effort was evaluated via self-report measures. Although Swerdzewski, Harmes, and Finney (2011) showed that these are also valid indicators of student motivation in low-stakes tests, Wise and Smith (2011) pointed out that it is unclear how truthfully examinees indicate their effort. Indeed, a very low test-taking effort may impair the reliability of the scale itself, especially if examinees are requested to complete the effort questionnaire several times during the course of the test. Having in mind related problems, Debeer et al. (2014) instead use effects of item position as an indicator of test-taking effort. The results of the current study, however, suggest that effort is only one facet contributing to position effects. Both cannot be viewed interchangeably. More specifically, what is the best method for measuring change in test-taking effort is unknown yet: self-reported questionnaires administered repeatedly during the test, response time effort (RTE) measures that are based on reaction time (Wise & Kong, 2005), or the change in performance related to item position during testing (Debeer & Janssen, 2013).

A further limitation is that models from the GLMM framework are always linear in the parameters. This feature allows, for instance, the estimation of the Rasch (1PL) model and extended Rasch-based models such as the ones the authors have used. Other IRT models such as the 2PL/3PL are not supported. Hence, the GLMM framework cannot be used to determine whether the item discrimination parameter is affected by the item's position or the person's effort. For this purpose, generalized nonlinear mixed models (GNLMM), which are implemented in the NLMIXED procedure in SAS, for example, need to be chosen (Debeer & Janssen, 2013). Alternatively, GNLMM may be specified within a Bayesian framework (Fox, 2010; Kruschke, 2011; Patz & Junker, 1999), for example, using JAGS (Plummer, 2013) or WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003).

Conclusion

The interdependence of position effects and test-taking effort stresses the need to motivate students repeatedly during the course of a time-consuming large-scale assessment. Examinees' test-taking effort should be kept constant—preferably at a high level. An alternative approach might be to reduce the time duration of the test. Overall, position effects may be diminished if the test length is reduced, for example, to only 80 min instead of 120 min. In the NAEP writing assessment, for example, test length was only 50 min (Qian, 2014). Consequently, the sample size would have to increase to gain the same measurement precision. Still, the fact that position effects do not vanish even in selected sub-samples of persistently high-motivated examinees underlines the importance of position-balanced test designs that guarantees for unbiased item parameters even though position effects are not parametrized in the measurement model (Hecht et al., 2015).

Authors' Note

Data and syntax used for the analyses may be requested from the corresponding author (confidentiality declaration necessary).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Institute for Educational Quality Improvement at Humboldt-Universität zu Berlin, Berlin, Germany.

References

- Albano, A. D. (2013). Multilevel modeling of item position effects. *Journal of Educational Measurement*, 50, 408-426.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4* (Version 1.0-6). Retrieved from <http://CRAN.R-project.org/package=lme4>
- Brennan, R. L. (1992). The context of context effects. *Applied Measurement in Education*, 5, 225-264.
- 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, 225-244. doi:10.1177/014662168701100302
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics*, 39, 502-523.
- Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, 50, 164-185.
- 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, 1-28.
- De Boeck, P., & Wilson, M. (2004). A framework for item response models. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 3-42). New York, NY: Springer.
- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling*. Mahwah, NJ: Lawrence Erlbaum.
- Eklöf, H. (2010, July). *Student motivation and effort in the Swedish TIMSS Advanced Field Study*. Paper presented at the 4th IEA International Research Conference, Gothenburg, Sweden.
- Fox, J.-P. (2010). *Bayesian item response modeling. Theory and applications*. New York, NY: Springer.
- Freund, P. A., Kuhn, J.-T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, 51, 629-634.
- Frey, A., & Bernhardt, R. (2012). On the importance of using balanced booklet designs in PISA. *Psychological Test and Assessment Modeling*, 54, 397-417.
- 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, 39-53.
- Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph*, 3, 125-156.
- Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly*, 50, 379-390.
- Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling*, 54, 418-431.
- Hecht, M., Roppelt, A., & Siegle, T. (2013). Testdesign und Auswertung des Ländervergleichs [Test design and analysis of the IQB national assessment]. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* [The IQB National assessment Study 2012. Competencies in mathematics and the sciences at the end of secondary level] (pp. 391-402). Münster, Germany: Waxmann.

- 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, 1021-1044. doi: 10.1177/0013164415573311
- Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly*, 50, 391-402.
- Hohensinn, C., Kubinger, K. D., 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, 497-509.
- 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). New York, NY: Springer.
- Kingston, N. M., & Dorans, N. J. (1984). Item location effects and their implications for IRT equating and adaptive testing. *Applied Psychological Measurement*, 8, 147-154.
- Kolen, M. J., & Brennan, R. L. (2004). *Testing equating, scaling, and linking: Methods and practice*. New York, NY: Springer.
- Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.
- Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, 21, 215-237.
- 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, 38-60. doi: 10.1080/08957340802558342
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.
- Organization for Economic Cooperation and Development. (2006). *Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Paris, France: Author.
- Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (Eds.). (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* [The IQB National Assessment Study 2012. Competencies in mathematics and the sciences at the end of secondary level]. Münster, Germany: Waxmann.
- Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, 24, 146-178.
- Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-Scale Assessments in Education*, 2, 2-17.
- Penk, C., & Richter, D. (2016). *Change in test-taking motivation and its relationship to test performance in low-stakes assessments*. *Educational Assessment, Evaluation and Accountability*, 1-25. doi:10.1007/s11092-016-9248-7
- Plummer, M. (2013). JAGS: Just another Gibbs sampler (Version 3.4.0) [Computer software]. Retrieved from <https://sourceforge.net/projects/mcmc-jags/>.
- Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement*, 38, 518-534.
- R Core Team. (2014). *R: A language and environment for statistical computing* (Version 3.1.0). Vienna, Austria: R Foundation for Statistical Computing. Available from <http://www.R-project.org/>
- 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, 47-64.
- Shohamy, E. (1996). Competence and performance in language testing. In G. Brown, K. Malmkjaer & J. William (Eds.), *Performance and competence in second language acquisition* (pp. 136-151). Cambridge: Cambridge University Press.
- Spiegelhalter, D. J., Thomas, A., Best, N., & Lunn, D. J. (2003). *WinBUGS user manual* (Version 1.4). Cambridge, UK: MRC Biostatistics Unit.
- Swerdzewski, P. J., Harmes, C. J., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, 24, 162-188.

- Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models* (pp. 289-316). New York, NY: Springer.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for two testlets. *Journal of Educational Measurement*, 24, 185-201.
- Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, 38, 535-548. doi:10.1177/0146621614534955
- Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, 10, 1-17.
- Wise, S. L., & Kong, X. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, 18, 163-183.
- Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139-153). Washington, DC: American Psychological Association.