A Multilevel Study of Position Effects in PISA Achievement Tests:

Student- and School-Level Predictors in the German Tracked School System

Gabriel Nagy
Leibniz Institute for Science and Mathematics Education,
Olshausenstraße 62, 24118 Kiel, Germany
E-Mail: nagy@ipn.uni-kiel.de

Benjamin Nagengast
University of Tübingen,
Europastraße 6, 72072 Tübingen, Germany
Email: benjamin.nagengast@uni-tuebingen.de

Andreas Frey
Friedrich-Schiller-University Jena
Am Planetarium 4, 07743 Jena, Germany
E-Mail: andreas.frey@uni-jena.de
Centre of Educational Measurement at the University of Oslo
Postboks 1161 Blindern, 0318 Oslo, Norway
andreas.frey@cemo.uio.no

Michael Becker
German Institute for International Educational Research
Warschauer Straße 34-38, 10243 Berlin, Germany
E-Mail: becker@dipf.de

Norman Rose
University of Tübingen,
Europastraße 6, 72072 Tübingen, Germany
E-Mail: norman.rose@uni-tuebingen.de

Author Note

Correspondence concerning this article should be addressed to Gabriel Nagy, Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany. E-Mail: nagy@ipn.uni-kiel.de.

The research reported in the paper was supported by a grant from the Federal Ministry of Education and Research to the first author (Ref: 01LSA007).

Abstract

Position effects (PE) cause test material to become more difficult towards the end of a test. We analyzed PEs in science, mathematics, and reading tests administered in the German extension to the PISA 2006 study with respect to their variability at the student- and school-level. PEs were strongest in reading and weakest in mathematics. Variability in PEs was found at both levels of analysis. PEs were stronger for male students, for students with a migration background (science and mathematics), and for students with a less favorable socioeconomic background (reading). At the school level, PEs were stronger in lower school tracks and in schools with a high proportion of students with a migration background. The relationships of the test scores with the covariates partly reflected the covariates' relationships with PEs. Our findings suggest that PEs should be taken seriously in large-scale assessments as they have an undesirable impact on the results.

Keywords: PISA 2006, position effect, predictors of position effects, multilevel modeling, validity

A Multilevel Study of Position Effects in PISA Achievement Tests: Student- and School-Level Predictors in the German Tracked School System

In large-scale assessments of student achievement, item responses are used to derive population and subgroup estimates of proficiency distributions that indicate which sort of tasks the members of different groups are able to solve with a given probability of success (Watermann & Klieme, 2002). Optimally, the probability of solving an item correctly should depend only on the items' characteristics (e.g., their difficulties) and the students' proficiencies. However, most often, this goal is not achieved in practice (Asseburg & Frey, 2013). There is ample evidence that the context in which items are presented within a test affects the rates of correct responses (Brennan, 1992). One well-documented *test context effect* is the gradual decline in the probability of items being solved correctly as they are put closer to the end of a test (Leary & Dorans, 1985). This effect is known as a *position effect* (PE; e.g., Meyers, Miller, & Way, 2009). PEs appear to play a role in virtually all testing situations, occurring in tests of moderate to extensive lengths (Leary & Dorans, 1985). The size of PEs has been found to vary across individuals (Debeer & Janssen, 2013), subject domains (Frey, Bernhardt, & Born, 2017; Nagy, Lüdtke, & Köller, 2016; Nagy, Lüdtke, Köller, & Heine, 2017) schools (Debeer, Buchholz, Hartig, & Jansen, 2014), and even countries (Hartig & Buchholz, 2012). As a consequence, the inferences drawn on the basis of test scores could be affected by PEs (e.g., Nagy, Lüdtke, & Köller, 2016; Nagy, Retelsdorf, Goldhammer, Schiepe-Tiska, & Lüdtke, 2017): Group differences in test scores could partially reflect differences in PEs, and correlations between test scores and student and school characteristics might be sensitive to the variables' relationships with PEs.

The aim of the present article is to extend knowledge about the role that PEs play in large-scale assessments. More specifically, we investigated three interrelated research questions. First, we examined whether the proficiency domains assessed in PISA (i.e., science, mathematics, and reading) are differently impacted by PEs (Nagy et al., 2016; Nagy, Lüdtke, Köller, & Heine, 2017). Second, we examined student-level and school-level correlates of PEs. Here, we focused on the characteristics commonly employed in large-scale assessments that receive high public attention, namely, students' *migration background*, their *parental socioeconomic status* (SES), their *gender*, and the *school type* they attend. Finally, we evaluated the robustness of analyses typically reported in large-scale assessments with respect to PEs. We used the data of the German national extension to PISA 2006 (Prenzel et al., 2008) and we present analyses that are based on a large subsample of 33,480 students nested in 1,030 schools, drawn from the western part of Germany and sharing the same tracked school system.

Position Effects and Test-Taking Behavior

Because research has consistently documented that the size of PEs varies across individuals (Debeer et al., 2014; Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Robitzsch, 2009; Hecht, Weirich, Siegle, & Frey, 2014; Weirich, Hecht, Penk, Roppelt, & Böhme, 2017), some authors have discussed PEs as indicators of individuals' *test-taking persistence* (Debeer et al., 2014; Hartig & Buchholz, 2012). According to this view, PEs can be expected to be related to cognitive and motivational resources that are relevant for maintaining a constantly high level of effort and precision when working on cognitive tasks. The findings gathered so far are in line with this assertion. PEs in abstract aptitude tests were found to be related to basic cognitive capacities, such as general intelligence (Schweizer, Troche, & Rammsayser, 2011), and attention (Ren, Goldhammer, Moosbrugger, & Schweizer, 2012). Lindner, Nagy, Ramos, and Retelsdorf (2017) showed that experimentally depleting students' self-control resources resulted in stronger PEs in a mathematics test. In addition, PEs in a science tests were found to be correlated with decreases in self-reported test-taking effort (Weirich et al., 2017). Similarly, Qian (2014) found motivation to be an important predictor of PEs in the writing task included in the 2007 National Assessment of Educational Progress (NAEP).

Some studies have shown that, beside cognitive and motivational factors, other variables also impact on PEs. Qian (2014) found that school type, as well as other institutional characteristics, were related to PEs. Profound school-type differences in the size of PEs were found in the German PISA 2012 assessment (Nagy et al., 2016; Nagy, Lüdtke et al., 2017) and for the tests of the German educational standards (Nagy, Haag, Lüdtke, & Köller, 2017). Here, PEs were stronger in school types with a lower average achievement that encompass students with less favorable family backgrounds. In addition, the size of PEs in reading was found to vary between schools in PISA 2009 (Debeer at al., 2014), but variability in PEs between learner groups appeared to be small in a science test (Weirich et al., 2017).

School differences in PEs might occur for several reasons. They could be due to school differences in the student composition defined with respect to student-level variables related to the size of PEs. For example, students' motivation might differ between schools, leading to school differences in PEs. In addition, the composition of students within a school could give rise to contextual effects (Raudenbush & Willms, 1995), which means that the student characteristics aggregated on the school level predict PEs over and above the individual characteristics. Specific constellations of student characteristics might constitute a climate which makes students more reluctant to maintain their effort when working on a test.

Consequences of Position Effects for Inferences about Proficiencies

As test scores are affected by students' test-taking persistence, the relationships of PEs with the covariates under study can be considered to be a threat to the validity of the inferences drawn about the correlates of students' proficiencies. This issue is relevant in large-scale assessments, where the relationships of test scores are often generalized beyond the test. Large-scale assessments aim to provide information about the capability of members of different subpopulations (e.g., gender groups) to successfully perform tasks characteristic of real-life situations (Watermann & Klieme, 2002). However, in situations where the size of PEs differs between groups, differences in test scores might lead to biased conclusions because the extent to which PEs can be generalized to real-life settings is unclear. In most real-life settings, students are not required to draw on their proficiencies for such an extended time as that required in large-scale assessments (e.g., 2 hours in PISA), which means that many real-life problems are less affected by the individuals' persistence in focusing on the relevant tasks. Furthermore, many real-life problems have a high-stakes character, making it more likely that individuals invest full effort across longer periods of time, whereas this seems less likely in the low-stakes situations of large-scale assessments (DeMars, 2007).

The potential impact of PEs on the conclusions drawn from test scores was exemplified in the longitudinal extension to the German PISA 2012 assessment. Nagy et al. (2016; Nagy, Lüdtke et al., 2017) found that ignoring PEs and changes therein resulted in negative estimates of proficiency gains in nonacademic tracks for reading and science, and that this effect vanished once PEs were accounted for. Similarly, ignoring PEs resulted in differences in reading gains in favor of girls, and these differences disappeared once PEs were controlled for (Nagy, Retelsdorf et al., 2017).

The Present Study

In the present study, we focused on PEs in the German extension to the PISA 2006 study. This study provides a large sample in which the PISA tests were administered in a single language, thereby avoiding artifacts due to the linguistic peculiarities of countries (Kreiner & Christensen, 2014). The German school system is a tracked system, made up of different school types, with large differences in students' achievement and background characteristics (Baumert, Stanat, & Watermann, 2006; Maaz, Trautwein, Lüdtke, & Baumert, 2008). Drawing on this large database, consisting of more than 33,000 students, which is in many respects prototypical of most large-scale assessments, we investigated research questions that are of theoretical and practical concern.

First, we examined whether the domains assessed in PISA, namely, science, mathematics, and reading, are equally prone to PEs. Here, we analyzed two aspects of PEs:

their size (i.e., the average decline in test scores from the first to the last position), and their pattern (i.e., the pattern of declines across positions). Nagy et al. (2016; Nagy, Lüdtke et al., 2017) found the size and the pattern of PEs to differ between the domains assessed in the German PISA 2012 study. Reading was most strongly impacted by PEs; the average PEs for reading were strongest and the declines corresponded to a linear function. Mathematics was found to be most robust to PEs, as the average size of the PEs was smallest, and only the second half of the test was impacted by PEs. Based on these findings, we expected to find a similar pattern in the German extension to the PISA 2006 study.

Second, we examined the variability in PEs on the student and school level. Whereas individual differences in PEs seem to be the rule, less is known about school differences in PEs. Based on the findings of Debeer and colleagues (2014), who documented between-school variability in PEs for the PISA 2009 reading test, we expected to find similar results in our study as well. However, the question of whether between-school variability in PEs exists in science and mathematics remains open. For example, Weirich et al. (2017) found that PEs in science had a rather negligible variability on the class level.

Third, we investigated whether PEs were related to student- and school-level characteristics commonly employed in large-scale assessments. Here, we concentrated on variables that receive a large amount of public attention, namely, students’ migration background, their SES, their gender, and the school type they attend. In PISA 2006, students with a migration background often had problems with the German language (Prenzel et al., 2008). Working on tests presented in German was thus likely to be more difficult for them, so that we expected them to be more prone to PEs. Students’ SES background is known to be related to their valuation of education (Sirin, 2005) and might therefore also affect how they approach standardized testing situations, including their persistence. In terms of gender differences, girls have been reported to have higher levels of grit and self-control, making them more likely to counteract PEs (Duckworth & Seligman, 2006). The school type that students attend was found to be related to PEs in the German PISA 2012 assessment. Students in the academic track (i.e., the Gymnasium) were least affected by PEs (Nagy et al., 2016), and we expected to find the same pattern of results in the PISA 2006 data. Furthermore, we expected that PEs would be strongest in the lowest track (i.e., Hauptschule) because, in this track, several risk factors are concentrated (i.e., low achievement standards, less favorable family backgrounds, lower school motivation, and higher rates of problem behavior; e.g., Baumert et al., 2006). In addition, we also investigated the role of students' migration background and SES, aggregated on the school level. When considered as contextual characteristics, these variables might exert effects over and above the student-level variables on PEs, not only because shared attitudes that hinder learning (Baumert et al., 2006; Perry & McConney, 2010) but also the assessment of learning outcomes could become concentrated in such schools.

Our last question targeted the robustness of inferences about the correlates of students' proficiencies. Based on the findings obtained from the German PISA 2012 study (Nagy et al., 2016; Nagy, Retelsdorf et al., 2017), we expected that the stronger the relationships of the PEs with the covariates under investigation, the more strongly the relationships of test scores would be affected.

Method

Sample

The sample was taken from the German extension study to PISA 2006, which was drawn together with the sample used for the international comparison. The target population was 9th-grade students as well as 15-year-olds. In order to maximize the sample size, both samples were combined, but students with special educational needs were excluded. We excluded further cases in order to facilitate comparisons between school types that are complicated by state differences in the school structure. The school structures in the western and eastern parts of Germany are very different. Because the national extension to PISA 2006 offers a much larger sample for the western parts of Germany, we focused our analyses on western Germany. In addition, we excluded the federal state Saarland because it does not include the intermediate track (Realschule).

Traditionally, in western Germany, students are tracked in the different schools of the three-tiered secondary school system from as early an age as 10 (i.e., Grade 5). The three main secondary school types (Hauptschule with 9 years of schooling, Realschule with 10 years of schooling, and Gymnasium with 12 to 13 years of schooling) differ in the intensity of the curriculum. Comprehensive schools (Integrierte Gemeinschaftsschule) and combined tracked schools are further school types found in almost all federal states.

We only used cases with complete data on the covariates under consideration. We did not impute these variables due to the complicated structure of the achievement data being considered. Given the large sample size, we decided on a listwise deletion of cases, leading to a 16% reduction in the total sample size, and a total sample of *N* = 33,480 students nested in *J* = 1,030 schools, with an average size of *N* = 32.50 students per school.

Covariates

School Type. We distinguished between 4 school types (lower track, *N* = 6,559; intermediate track, *N* = 10,094; academic track, *N* = 10,493; and combined track *N* = 6,334). The combined track included comprehensive schools, as well as combined tracked schools.

Migration Background. Students' migration background was taken from the official school records for each student. Students with at least one parent born in a country other than Germany were considered as having a migration background (24.3%).

Socioeconomic Status. SES was assessed by the Socio-Economic Index of Occupational Status (ISEI; Ganzeboom, De Graaf, Treiman; 1992). The ISEI was derived from parents' reports about their occupations or from students' reports if parent data were missing. In order to facilitate the analyses, the ISEI was *z*-standardized (*M* = 0, *SD* = 1) in the complete sample.

Gender. Students' gender was taken from the school files. Female students were used as the reference category and were coded as 0, male students received a code of 1 (50.2%).

Scoring of Proficiency Tests

Table 1 presents the matrix design employed in PISA 2006 to assess students' proficiencies. Science was the major domain and was measured by seven item clusters (groups of items S1 to S7 in Table 1). Mathematics and reading were minor domains, measured by four clusters in the first case (M1 to M4), and by only two clusters in the second case (R1 and R2). The matrix design is a so-called Youden square design (Frey, Hartig, & Rupp, 2009) which balances for item clusters across positions, in such a way that each item cluster was presented in each position exactly once. Students were randomly assigned to test booklets, and random assignment took place within schools.

PEs were examined on the basis of continuous achievement scores. To this end, we treated each item cluster (R1, R2, M1 to M4, and S1 to S7 in Table1) as a different test, so that, for each examinee, four test scores were derived, which we term *item cluster scores*. For example, students working on booklet B01 received item cluster scores for S1, S2, S4, and S7 (Table 1). Item cluster scores were estimated by applying a four-dimensional Rasch model separately to each booklet. In order to put the item cluster scores on a common metric, item parameters were fixed to the values provided in PISA 2006 (OECD, 2009). Scores were estimated with the plausible value technique (PV; Mislevy, Beaton, Kaplan, & Sheehan, 1992) using the ConQuest program (Wu, Adams, Wilson, Haldane, 2007). We generated five sets of PVs while accounting for the covariates at both levels of analysis (Wu, 2005). The marginal PV reliabilities (Adams, 2005) were good (Table 1). We also derived *conventional* PVs on the basis of a three-dimensional item response theory (IRT) model (science, mathematics, and reading), in which the same covariates were used. These PVs served as a comparison standard for examining the impact of PEs on the results.

A Multi-Level Model for Student and School Differences in Position Effects

The booklet design employed in PISA 2006 offers a good opportunity for examining PEs. Random assignment implies that there is no reason to expect that the distribution of the students' (true) proficiencies or their relationships with the covariates differ between booklets. Therefore, results showing that the item cluster scores follow a decreasing trend across positions provide indications that PEs are at work. In addition, findings showing that such declines are accompanied by systematic changes in the item cluster scores' relationships with the covariates suggest that the covariates are related to PEs (Nagy et al., 2016).

Let y_{ijcp} be the domain score (science, mathematics, or reading) of individual *i* attending school *j* assessed on the basis of cluster *c* administered in position *p* (*p* = 1, 2, 3, 4). The model builds upon a separation of the scores y_{ijcp} into two components: An initial level, η_0, underlying the scores assessed in the first position, and a position component, η_Δ, reflecting the amount of score changes when administering an item cluster in the last instead of the first position in the test. Both components are separated in two parts, one located at the school level (η_{j0}, and $\eta_{j\Delta}$) and one at the student level (η_{ij0}, and $\eta_{ij\Delta}$). The η-variables are linked to the *y*-variables according to the following model:

$$y_{ijcp} = \tau_{c0} + \lambda_{c0}(\eta_{ij0} + \eta_{j0}) + \lambda_{cp}(\eta_{ij\Delta} + \eta_{j\Delta}) + \varepsilon_{ijcp}\ . \qquad (1)$$

In this model, τ_{c0} is a cluster-specific intercept term that accommodates differences in cluster difficulties, and λ_{c0} is a cluster-specific loading parameter that relates the initial level to the y-scores. λ_{cp} is a cluster- and position-specific loading parameter that captures the impact of PEs on the item cluster scores assessed by cluster *c* in position *p* > 1. For clusters appearing in the first position *p* = 1, the λ_{cp}-parameter is fixed to zero (i.e., $\lambda_{c1} = 0$ for all *c*). Finally, ε_{ijcp} stands for the residual of individual *i*'s item cluster score assessed on the basis of cluster *c*, administered in position p, from her or his score predicted by η_0 and η_Δ. The residuals are assumed to be mutually uncorrelated and to have a constant variance across positions, but variances were allowed to differ between item clusters.

The model can be extended to include covariates. On the student level, the equations take the form

$$\eta_{ij0} = \sum_{k=1}^{K} \beta_{0k}^{W} x_{ijk} + \zeta_{ij0} \quad \text{, and} \quad \eta_{ij\Delta} = \sum_{k=1}^{K} \beta_{\Delta k}^{W} x_{ijk} + \zeta_{ij\Delta} \text{ ,} \qquad (2)$$

whereas the school-level model adheres to

$$\eta_{j0} = \sum_{l=1}^{L} \beta_{0l}^{B} w_{jl} + \zeta_{j0} \text{ ,} \quad \text{, and} \quad \eta_{j\Delta} = \alpha_{\Delta} + \sum_{l=1}^{L} \beta_{\Delta l}^{B} w_{jl} + \zeta_{j\Delta} \text{ .} \qquad (3)$$

Here, β_{0k}^{W} and $\beta_{\Delta k}^{W}$ are student-level regression weights that represent the impact of the student-level covariate x_k on the initial score and the PE, respectively, whereas terms β_{0l}^{B} and $\beta_{\Delta l}^{B}$ are the corresponding regression weights that are applied to the school-level covariate w_l. The ζ-variables are random disturbances on the student (Equation 2) and the school level (Equation 3) with zero means and a freely estimated variance-covariance structure. The α_{Δ}-parameter in Equation 3 stands for the structural intercept of the PE, whereas α_0 does not appear in the school-level model because this parameter is fixed to 0 for reasons of model identification.

The interpretation of the regression weights in Equation 3 depends upon the centering of explanatory variables. Whenever a predictor x_m is not group-mean centered, and w_m represents the student-level predictor aggregated on the school level, the regression weights β_{0m}^{B} and $\beta_{\Delta m}^{B}$ can be interpreted as contextual effects affecting the initial level and the PE over and above the effects of x_m (Enders & Tofighi, 2007).

In order to keep the model identified, two sets of identification restrictions need to be imposed. First, the mean or intercept of the initial level (η_0) cannot be estimated as long as all τ-parameters are freely estimated. This parameter was therefore fixed to zero. Second, the measurement scale of the initial scores, η_0, and of the PEs, η_{Δ}, needs to be defined. To this end, we selected a reference cluster *c* = *h* for which we fixed the loadings, so that $\lambda_{h0} = 1$ and

$\lambda_{h4} = 1$. These restrictions put the η-variables and the corresponding structural parameters (i.e., variances, intercepts/means, and regression weights) on the metric of the reference cluster h. In order to facilitate the interpretation of results, we transformed the structural parameters and their standard errors (derived by the Delta method) to be defined with respect to the average clusters.

Parameter Estimation. Our approach to modeling PEs can be considered to be a multilevel structural equation model (Rabe-Hesketh, Skrondal, & Zheng, 2007). It can be estimated by software that makes it possible to handle missing data that arise as a result of the incomplete pairing of item clusters due to the assessment designs (Table 1). We employed the M*plus* 7.4 program (Muthén & Muthén, 2012), using a maximum likelihood estimator with standard errors approximated by first-order derivatives (MLF) by employing the EM algorithm. We chose the MLF estimator because it is a robust method for estimating the asymptotic variance-covariance matrix of parameter estimates in the presence of a substantial amount of missing data and of highly correlated school-level variables (Asparouhov & Muthén, 2012).

The matrix design of PISA 2006 did not allow us to identify the student-level covariance between the initial level and the PE for reading. Item clusters assessed in the first position were not paired with reading clusters in a later position in any booklet (Table 1). In order to resolve this issue, we fixed the corresponding covariance to zero. This restriction could affect the estimate of the variance of PEs on the student level, but has no consequences for the remaining parameter estimates.

Results

In the following sections, we first provide descriptive results derived on the basis of the item cluster scores. We then proceed to the decomposition of PEs on the student level and school level. In the next step, we provide results pertaining to the correlates of the PEs on both levels. In the last section, we investigate the consequences of ignoring PEs when examining the covariates' relationships with the test scores.

Descriptive Results

Figure 1 presents the means and variances of the item cluster scores by position. In order to facilitate the interpretation of the results, all item cluster scores were standardized according to their means and variances when assessed in the first item cluster position (M = 0, SD = 1 for each item cluster score in p = 1). Therefore, mean trajectories across positions can be interpreted similarly to standardized effects (d).

The results provide evidence for PEs in all domains, as all item clusters showed gradual declines in their means across positions. PEs were stronger in reading, with an average change from the first to the last position of $\bar{d} =$ -0.50 standard deviations. Here, decreases were already observed in the second item cluster position. Mathematics was least affected by PEs. The average change from the first to the last position was $\bar{d} =$ -0.26, and mean scores assessed in the second position were almost identical to the first position. Science occupied an intermediate position. The average change in mean scores was $\bar{d} =$ -0.35, and means assessed in the second position were only weakly impacted by PEs ($\bar{d} =$ -.07). The variances of item cluster scores showed a gradual increase across positions. Changes were strongest in reading (average variance of 1.70 in the last position), and weakest in mathematics (average variance of 1.23). The results for changes in means and variances shown in Figure 1 suggest that the item clusters belonging to one domain were differently impacted by PEs, because the trajectories of the means and variances differed between item clusters. However, we found a close correspondence between the item clusters' means and variances, such that larger changes in means were accompanied by larger changes in variances (science: r = -.78, mathematics: r = -.83, reading r = -.97). This observation supports our assumption that both aspects of the data – means and dispersions – reflect a common process.

Variability of Position Effects

The first models provided estimates of the variances of initial levels and the PEs on the student level and the school level that are given on the logit metric (Table 2). The loadings on the PE-variable averaged across item clusters agreed with the descriptive results. PEs were found to vary on the student level, as well as on the school level. In all domains, the variability of initial levels was found to be almost equal on both levels of analysis, so that the intraclass correlations (ICC) for initial values were close to .50 (50% of the variance was due to differences between schools). In contrast, between-school variability in PEs was smaller than the student-level variance. In mathematics, only 4% of the variability was found to be due to school differences in PEs, whereas for science and reading, a larger proportion of variability was found to be located on the school level (12% and 18%, respectively). Note, however, that in the case of reading, the PEs' variance estimates at the student level, as well as the corresponding ICC, should be interpreted with caution.

Correlates of Position Effects

The baseline models were extended by student-level and school-level covariates. In the first set of models, each covariate was considered separately but, when appropriate, covariates were included in both levels of the analysis. In order to facilitate the interpretation of the results, school track was effect-coded in such a way that the effects of school tracks sum to zero.

The relationships of the covariates are reported in Table 3. On the student level, all covariates were found to be related to the initial level in each domain. Male students had higher scores in science and mathematics, but lower scores in reading. Students with a migration background had lower scores, and SES was positively related to test scores in all domains. In addition, the covariates were found to be related to PEs located on the student level, although the patterns of relationships differed somewhat between domains. In all domains, PEs were stronger in male students (negative coefficients). Migration background was related to PEs in science and mathematics (with stronger declines in students with a migration background), but not in reading. SES was related only to PEs in reading, with students with a more favorable background being less affected by PEs (positive coefficients).

The covariates' school-level relationships with the initial level were similar across domains. The initial level was negatively related to the proportion of students with a migration background and positively related to the school-average SES. In addition, we found large differences between school tracks in the expected order. On the school level, the pattern of relationships with PEs was found to be remarkably similar across domains, although relationships appeared to be stronger for reading. PEs were stronger in schools in which a larger proportion of students had a migration background, and a higher average SES was accompanied by weaker PEs. Finally, school type was clearly related to PEs: PEs were strongest in the lowest track and weakest in the academic track in all domains.

In order to provide a more detailed picture of the relationships between PEs and the covariates, Figure 2 illustrates the position-specific item cluster scores for male and female students. As shown, male students had steeper declines in all domains. As such, gender differences in science and mathematics that were in favor of male students (Table 3) decreased across positions because male students were more prone to PEs, whereas for the same reasons, in reading, gender differences in favor of female students increased across positions. Therefore, gender differences that were averaged across positions were smaller (science and mathematics) or larger (reading) than the gender differences in the initial levels. This issue will be covered in the next section, where we consider the impact of PEs on the relationships of test scores.

We now turn to the covariates' unique contributions toward predicting PEs. To this end, we included all covariates simultaneously into the models (Table 4). On the student level, the relationships with initial levels and PEs were very close to the univariate results (Table 3). The regression weights of the school-level covariates that predicted the initial level component decreased once all covariates were used simultaneously, although the pattern of effects was similar to the univariate results. Similarly, the relationships with PEs on the school level changed once all covariates were considered simultaneously. Only the effects of school types on PEs were close to the previous results (Table 3). The effect of the average SES on PEs was no longer reliable in any domain, and the effect of the proportion of students with a migration background was reduced in the case of science and reading but remained rather unchanged in mathematics. These findings suggest that the relationships of the average SES with PEs were largely due to school-track differences, but that the proportion of students with a migration background was a contextual characteristic that had a unique impact on PEs.

Sensitivity of Relationships of Test Scores to Position Effects

In this section we examine the sensitivity of the relationships between test scores and student- and school-level variables to PEs. We focus on univariate results that are often employed in practice. We approach this question by first comparing the covariates' effects on the initial levels with their effects averaged across all item clusters and positions as implied by our model. Note that our model builds upon the assumption that the initial levels are not affected by PEs. Therefore, the average effects of the covariates can be considered to combine (1) their effects on proficiency measures purified for PEs, and (2) their average effects on PEs across all combinations of item clusters and positions. As a consequence, results showing that a covariate's effect on the initial components is different from its average effect indicate that PEs could have an impact on the conclusions. In addition, we also compared the covariates' effects on conventionally constructed PVs, as used in PISA, with their relationships with the initial level given by our model. IRT-based PVs are constructed upon the assumption that item responses and their relationships with covariates are not affected by their position in the test. Therefore, it is possible that the relationships of conventionally constructed PVs with covariates are more robust to PEs.

In order to quantify the impact of PEs on an interpretable metric, we used the PISA metric (*M* = 500 and *SD* = 100 in the international sample). In addition, we focused on relationships that can be regarded as being substantively meaningful (i.e., larger than 10 points on the PISA metric, roughly |*d*| ≥ 0.1). We considered PEs to have a potentially consequential impact on conclusions if effect sizes adjusted for PEs differed by more than 10% from the average effect sizes [(adjusted effect – average effect)/average effect × 100], or from the effect sizes determined on the basis of conventional PVs. Table 5 provides the average effects, the effects on PVs, and the absolute and relative discrepancies between adjusted and unadjusted effects. The effect sizes refer to (1) differences in PISA scores between the levels of categorical variables (gender and migration background), (2) differences between the 80th and 20th percentiles of the distribution of continuous covariates, and (3) differences between average scores in school types and the overall mean. The results show that the covariates' average effect sizes were very close to their relationships with conventional PVs (MAD = 1.4 PISA points). As a consequence, the discrepancy between the adjusted effect sizes and the average effect sizes was close to the differences between the adjusted effects and the effects determined on the basis of conventionally scored tests. Therefore, we only briefly review the results for the conventionally scored tests.

According to our previously defined criteria, in science and mathematics, the student-level relationships of gender and migration background were considered to be sensitive to PEs (note that the effect of migration background in science missed the 10% criterion by 0.2% on the basis of average effects). The effect sizes for gender increased once PEs were accounted for, by 24% in science and 13% in mathematics, whereas the effect sizes for migration background were reduced by 11% and 14%, respectively. For science and mathematics, the differences in the effect sizes for the remaining covariates were below 10%. As a consequence, between-school differences in PEs had a rather negligible impact on the associations of test scores with school characteristics.

The relationships of reading scores with the covariates were most sensitive to PEs. Only the relationship of migration background assessed at the student-level was not affected by PEs, because this variable was not related to the PEs (Table 3). Here, the relationship with SES at the student level was most sensitive to PEs: the effect size was reduced by 32% once PEs were accounted for. In the case of reading, the relationships with the school-level covariates were also affected by PEs. For example, the proficiency gap between schools at the 20th and 80th percentile on the migration background variable decreased by nearly 10 points (18%) when PEs were accounted for. Similarly, the gap between the lowest and highest school types was reduced by 25 points when PEs were controlled for.

Summary and Discussion

In the present article, we investigated PEs in the German extension to the PISA 2006 study. We found evidence that all proficiency domains were impacted by PEs, although to a different degree (Nagy et al., 2016). Mathematics was least affected by PEs as the average declines in the test scores from the first to the last position in the test were smallest, and only the second half of the test was affected by PEs. In contrast, reading was found to be most susceptible to PEs, as declines in scores were strongest and PEs already impacted the second item cluster position. Furthermore, we found PEs to vary on the level of students and schools. The between-school variability in PEs was small in mathematics and largest in reading. In addition, we found PEs to be related to the student-level and the school-level covariates. Finally, we found an indication that the possibility that PEs threaten the validity of inferences drawn about the correlates of students’ proficiencies cannot be ruled out: The stronger the covariates’ relationships with the PEs were, the more the inferences were affected.

On the one hand, our study indicates that PEs are a general phenomenon that affect proficiency tests in all of the domains tested in PISA. Accordingly, we found that the PEs in all domains had quite similar patterns of relationships with the covariates considered. For example, male students showed stronger PEs, and PEs were weakest in schools with more favorable background characteristics (i.e., academic schools, schools with a low proportion of students with a migration background). However, on the other hand, we also found an indication that PEs function in a domain-specific way because PEs in reading were stronger, more variable, and more closely related to school-level variables.

Compared to science and mathematics, the reading test is much easier, at least when presented in the first item cluster position. Hence, it appears that most students who invest full effort have a high chance of solving the reading items correctly. The cognitive demands imposed by the test appear to consist mainly of repeatedly engaging in the process of text reconstruction (Kintsch, 1998), so that items positioned later in the test become highly sensitive to students' persistence in investing effort and maintaining attention. Therefore, PEs in reading are likely to be most susceptible to the characteristics of learner groups that are related to shared negative attitudes towards testing. Of course, our explanation for the susceptibility of reading items to PEs is speculative. More research is called for, in which these issues should be investigated more thoroughly.

Consequences of Position Effects for Inferences

PEs might be seen as a threat to the validity of the inferences that are drawn about the correlates of students' proficiencies because the extent to which they can be generalized beyond the test is unclear. We found some evidence that the relationships of test scores typically obtained in PISA are, to a certain degree, affected by PEs (see also Nagy et al., 2016; Nagy, Retelsdorf, et al., 2017). For example, gender differences in favor of male students in science and mathematics were smaller when conventional test scores were used, because PEs were stronger in male students. For the same reason, gender differences favoring female students were larger in reading. Similarly, we found that adjusting for PEs reduced the relationships of test scores with students' migration background (in science and mathematics) and SES (in reading). In contrast, we found the relationships of school-level characteristics with test scores in science and mathematics to be relatively robust to PEs, but that these relationships depended on PEs in the case of reading.

The relevance of PEs for the conclusions drawn depends on the research question. In the present case, most of the differences between the adjusted and unadjusted results were below 20%, and the conclusions about the existence and the directions of relationships were not sensitive to PEs. Therefore, it appears that such inferences are not strongly affected by PEs, although this possibility cannot be completely ruled out in other applications. However, PEs appear to play a more important role in other research contexts where exact effect sizes matter. For example, certain PISA results, such as gender differences and SES effects, are routinely compared between countries (e.g., Willms, 2006). Here, changes in gender gaps as small as 4 points on the PISA metric (in science and mathematics), as well as reductions in SES effects of 32% (in reading) can have profound consequences for the ranking of countries.

Of course, PEs are only then a threat to the validity of inferences when they are considered as a source of construct irrelevant variance (Messick, 1995). More specifically, practitioners and researchers need to decide whether students' test-taking persistence, as reflected in PEs, should or should not be regarded as part of the proficiency construct. Regarding test-taking persistence as part of the construct being measured is completely legitimate (e.g., Weinert, 2001). However, such a construct would be relatively unhandy since it will include average effects (e.g., lack of motivation at the end of the test) that occur during a period of two hours (the testing time of PISA). Therefore, given the impact that PEs can have on results, researchers should clearly define the proficiency constructs.

Further Research

Test-taking persistence, as indicated by PEs, plays a role in virtually all cognitive tests assessed in low-stakes conditions (Leary & Dorans, 1985). Therefore, PEs should be taken seriously in large-scale assessments such as PISA. Given their potential impact on the results, more research that focuses on the proximal determinants of PEs is needed (Weirich et al., 2017), as well as on the related indicators of students' test-taking behavior, such as students' propensity to skip items (Holman & Glas, 2005) and to provide rapid responses (Wise & Kong, 2005). We believe that such research would help to provide a better understanding of the limitations but also the potential of large-scale assessments. Clearly, a thorough investigation of such issues requires innovative measurement designs that make it possible to fit complex statistical models that provide a fine-grained picture of the underlying mechanisms.

Of course, large-scale assessments are limited in their possibilities for employing more complex designs using a larger number of booklets. However, assessment designs suitable for large-scale assessments that have a rather limited number of test booklets could be optimized in order to better account for PEs (Weirich, Hecht, & Böhme, 2014). Although balanced with respect to the positions of item clusters, the matrix designs used in PISA were not developed with respect to the assessment of PEs. As such, the designs impose limitations on the identification of certain parameters and might provide less accurate results for domains assessed with fewer item clusters. Therefore, we applaud the recent developments undertaken in PISA, which result in minor domains being assessed by a larger number of item clusters, but we also stress the need to evaluate design options that help to better account for the often unwarranted impact of PEs (Weirich et al, 2014).

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation, 31*, 162-172. doi: 10.1016/j.stueduc.2005.05.008

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*, 92-104.

Asparouhov, T., & Muthén, B. (2012). *Saddle points*. http://www.statmodel.com/download/SaddlePoints2.pdf

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and the development of differential environments for learning and development]. In J.Baumert, P.Stanat, & R.Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* (pp. 95–188). Wiesbaden, Germany: VS für Sozialwissenschaften.

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

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. doi: 10.3102/1076998614558485

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

DeMars, C. E. (2007). Changes in rapid-guessing behavior over a series of assessments. *Educational Assessment*, *12*, 23–45. doi: 10.1080/10627190709336946

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender differences in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology, 98*, 198-208. doi: 0.1037/0022-0663.98.1.198

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods, 12*, 121-138. doi: 10.1037/1082-989X.12.2.121

Frey, A., Bernhardt, R., & Born, S. (2017). Umgang mit Itempositionseffekten bei der Entwicklung computerisierter adaptiver Tests [Accounting for item position effects in the development of computerized adaptive tests]. *Diagnostica*, *63*, 167–178. doi: 10.1026/0012-1924/a000173

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. doi: 10.1111/j.1745-3992.2009.00154.x

Ganzeboom, H. B. G., & Treiman, D. J. (2003). Three internationally standardised measures for comparative research on occupational status. In J. H. P. Hoffmeyer-Zlotnik & C. Wolf (Eds.), *Advances in crossnational comparison. A European working book for demographic and socio-economic variables* (pp. 159–193). New York, NY: Kluwer Academic ⁄ Plenum Publishers.

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., 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

Holman, R., & Glas, C. A. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology, 58*, 1-17. doi: 10.1111/j.2044-8317.2005.tb00312.x

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*, 210-231. doi: 10.1007/s11336-013-9347-z

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*, 387-413. doi: 10.3102/00346543055003387

Lindner, C., Nagy, G., Arhuis, W. A. R., & Retelsdorf, J. (2017). A new perspective on the interplay between self-control and cognitive performance: Modeling progressive depletion patterns. *PloS one, 12*, e0180149. doi: 10.1371/journal.pone.0180149

Maaz, K., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Educational transitions and differential learning environments: How explicit between school tracking contributes to social inequality in educational outcomes. *Child Development Perspectives, 2*, 99-106. doi: 10.1111/j.1750-8606.2008.00048.x

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749. doi: 10.1037/0003-066X.50.9.741

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

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161. doi: 10.1111/j.1745-3984.1992.tb00371.x

Muthén, L.K., & Muthén, B.O. (2012). *Mplus User's Guide*. 7th Edition. Los Angeles, CA: Muthén & Muthén.

OECD (2009). *PISA 2006 technical report*. Paris: OECD.

Nagy, G., Haag, N., Oliver, L., & Köller, O. (2017). Längsschnittskalierung der Tests zur Überprüfung des Erreichens der Bildungsstandards der Sekundarstufe I im PISA-Längsschnitt 2012/2013 [Longitudinal IRT scaling of tests of the educational standards for lower secondary level in the PISA longitudinal assessment 2012/2013]. *Zeitschrift für Erziehungswissenschaft, 20*, 259-286. doi: 10.1007/s11618-017-0755-1

Nagy, G., Lüdtke, O., & Köller, O. (2016). Modeling test context effects in longitudinal achievement data: Examining position effects in the longitudinal German PISA 2012 assessment. *Psychological Test and Assessment Modeling, 58*, 641-670.

Nagy, G., Lüdtke, O., Köller, O., & Heine, J. H. (2017). IRT-Skalierung der Tests im PISA-Längsschnitt 2012/2013: Auswirkungen von Testkontexteffekten auf die Zuwachsschätzung [IRT scaling of the tests in PISA longitudinal assessment 2012/2013: Impact of test context effects on the growth estimate]. *Zeitschrift für Erziehungswissenschaft, 20,* 229-258. doi: 10.1007/s11618-017-0749-z

Nagy, G., Retelsdorf, J., Goldhammer, F., Schiepe-Tiska, A., & Lüdtke, O. (2017). Veränderungen der Lesekompetenz von der 9. zur 10. Klasse: Differenzielle Entwicklungen in Abhängigkeit der Schulform, des Geschlechts und des soziodemografischen Hintergrunds? [Changes in reading skills from 9th to 10th grade: differential trajectories depending on school type, gender and socio-demographic background?] *Zeitschrift für Erziehungswissenschaft, 20*, 177-203. doi: 10.1007/s11618-017-0747-1

Perry, L., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *The Teachers College Record, 112*, 1137-1162.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klime, E., & Pekrun, R. (2008). *PISA 2006 in Deutschland – Die Kompetenzen der Jugendlichen im dritten Ländervergleich* [*PISA 2006 in Germany – The competencies in adolescents in the third state comparison*]. Münster, Germany: Waxmann.

Qian, J. (2014). An investigation of position effects in large-scale writing assessments. *Applied Psychological Measurement, 38*, 518-534. doi: 10.1177/0146621614534312

Rabe-Hesketh, S., Skrondal, A., & Zheng, X. (2007). Multilevel structural equation modeling. In S.-Y. Lee (Ed.), *Handbook of latent variable and related models* (pp. 209–227). Amsterdam, The Netherlands: Elsevier.

Raudenbush, S., & Willms, J. D. (1995). The estimation of school effects. Journal of *Educational and Behavioral Statistics, 20*, 307-335. doi: 10.3102/10769986020004307

Ren, X., Goldhammer, F., Moosbrugger, H., & Schweizer, K. (2012). How does attention relate to the ability-specific and position-specific components of reasoning measured by APM? *Learning and Individual Differences, 22*, 1-7. doi: 10.1016/j.lindif.2011.09.009

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests [Methodological challenges in calibrating achievement tests]. In D. Granzer, O. Köller, A. Bremerich-Vos, M. van den Heuvel-Panhuizen, K. Reiss, & G. Walther (Eds.), *Bildungsstandards in Deutsch und Mathematik* (pp. 42-107). Weinheim, Germany: Beltz.

Schweizer, K., Troche, S. J., & Rammsayer, T. H. (2011). On the special relationship between fluid and general intelligence: New evidence obtained by considering the position effect. *Personality and Individual Differences, 50,* 1249–1254. doi: 10.1016/j.paid.2011.02.019

Sirin, S. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75,* 417–453. doi:10.3102/00346543075003417

Watermann, R., & Klieme, E. (2002). Reporting results of large-scale assessment in psychologically and educationally meaningful terms: Construct validation and proficiency scaling in TIMSS. *European Journal of Psychological Assessment, 18,* 190-203. doi: 10.1027//1015-5759.18.3.190

Weinert, F. E. (2001). Vergleichende Leistungsmessung in Schulen - eine umstrittene Selbstverständlichkeit. In F. E. Weinert (Eds.), *Leistungsmessung in Schulen* (pp. 23-43). Weinheim und Basel: Beltz-Verlag

Weirich, S., Hecht, M., Penk, C., Roppelt, A., & Böhme, K. (2017). Item position effects are moderated by changes in test-taking effort. *Applied Psychological Measurement*, *41*, 115-129. doi: 10.1177/0146621616676791

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

Willms, J. D. (2006). Variation in socioeconomic gradients among cantons in French-and Italian-speaking Switzerland: Findings from the OECD PISA. *Educational Research and Evaluation, 12,* 129-154. doi: 10.1080/13803610600587008

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. doi: 10.1207/s15324818ame1802_2

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31,* 114-128. doi: 10.1016/j.stueduc.2005.05.005

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Melbourne, Victoria, Australia: ACER Press.

Table 1

Rotated Matrix Design used in the PISA 2006 Assessment, and Marginal PV-Reliabilities for Item Cluster Scores in Parentheses.

	Booklet												
	B01	B02	B03	B04	B05	B06	B07	B08	B09	B10	B11	B12	B13
Position 1	S1 (.87)	S2 (.89)	S3 (.87)	S4 (.88)	S5 (.92)	S6 (.89)	S7 (.89)	M1 (.88)	M2 (.87)	M3 (.89)	M4 (.82)	R1 (.85)	R2 (.86)
Position 2	S2 (.90)	S3 (.90)	S4 (.88)	M3 (.88)	S6 (.92)	R2 (.90)	R1 (.88)	M2 (.90)	S1 (.86)	M4 (.84)	S5 (.88)	M1 (.88)	S7 (.89)
Position 3	S4 (.90)	M3 (.90)	M4 (.90)	S5 (.90)	S7 (.93)	R1 (.88)	M2 (.91)	S2 (.89)	S3 (.92)	S6 (.91)	R2 (.88)	S1 (.86)	M1 (.87)
Position 4	S7 (.90)	R1 (.89)	M1 (.88)	M2 (.88)	S3 (.90)	S4 (.88)	M4 (.81)	S6 (.89)	R2 (.91)	S1 (.86)	S2 (.85)	S5 (.89)	M3 (.86)

Note. S = science, M = mathematics, R = reading

Table 2

Parameter Estimates in Models without Student-Level and School-Level Covariates

	Science		Mathematics		Reading	
	Est.	(*SE*)	*Est.*	(*SE*)	*Est.*	(*SE*)
Average Basis Function Coefficients						
$\bar{\lambda}_1$	0.00	(-----)	0.00	(-----)	0.00	(-----)
$\bar{\lambda}_2$	0.15	(0.015)**	0.06	(0.028)*	0.41	(0.023)**
$\bar{\lambda}_3$	0.47	(0.012)**	0.48	(0.033)**	0.67	(0.024)**
$\bar{\lambda}_4$	1.00	(-----)	1.00	(-----)	1.00	(-----)
Average Position Effects						
α_Δ	-0.32	(0.009)**	-0.29	(0.015)**	-0.54	(0.020)**
Random Effects						
Student Level						
ψ^W_{00}	0.49	(0.006)**	0.66	(0.008)**	0.55	(0.011)**
$\psi^W_{\Delta\Delta}$	0.13	(0.006)**	0.33	(0.031)**	0.26	(0.025)**
$\psi^W_{0\Delta}$	-0.01	(0.004)	-0.11	(0.016)**	-------	(-------)
School Level						
ψ^B_{00}	0.46	(0.035)**	0.66	(0.045)**	0.53	(0.038)**
$\psi^B_{\Delta\Delta}$	0.02	(0.002)**	0.01	(0.004)**	0.06	(0.014)**
$\psi^B_{0\Delta}$	0.09	(0.007)**	0.06	(0.011)**	0.18	(0.017)**
ICCs						
Initial Scores	.49		.50		.49	
PEs	.12		.04		.18	

Note. $\bar{\lambda}_1$ to $\bar{\lambda}_4$ = position-specific loadings averaged across clusters, ψ^W_{00} = variance in initial scores (student level), ψ^B_{00} = variance in initial scores (school level), $\psi^W_{\Delta\Delta}$ = variance of PEs (student level), $\psi^B_{\Delta\Delta}$ = variance of PEs (school level), $\psi^W_{0\Delta}$ = covariance between initial scores and PEs (student level), $\psi^B_{0\Delta}$ = covariance between initial scores and PEs (school level), ICC = intraclass correlation

***** $p \leq .05$; ****** $p \leq .01$

Table 3

Regression Weights for Student and School Characteristics Predicting Initial Levels and Position Effects. Results from Separate Models Applied to each Variable and Domain.

	Student Level		School Level	
	Initial Level	Position Effect	Initial Level	Position Effect
	Est.(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)
Science				
Male	0.27 (0.01)**	-0.14 (0.01)**		
Migration	-0.35 (0.12)**	-0.08 (0.01)**	-1.58 (0.12)**	-0.31 (0.03)**
SES	0.11 (0.01)**	0.00 (0.00)	1.07 (0.03)**	0.20 (0.01)**
Low Track			-0.78 (0.02)**	-0.16 (0.01)**
Combined Tr.			-0.21 (0.02)**	-0.03 (0.01)**
Intermediate Tr.			0.09 (0.02)**	0.02 (0.01)**
Academic Tr.			0.91 (0.02)**	0.17 (0.01)**
Mathematics				
Male	0.42 (0.01)**	-0.16 (0.02)**		
Migration	-0.27 (0.01)**	-0.14 (0.02)**	-1.89 (0.14)**	-0.31 (0.05)**
SES	0.11 (0.01)**	-0.02 (0.01)	1.31 (0.03)**	0.12 (0.03)**
Low Track			-0.90 (0.02)**	-0.10 (0.03)**
Combined Tr.			-0.30 (0.02)**	-0.03 (0.02)
Intermediate Tr.			0.07 (0.02)**	0.04 (0.02)*
Academic Tr.			1.12 (0.02)**	0.10 (0.02)**
Reading				
Male	-0.17 (0.02)**	-0.09 (0.04)**		
Migration	-0.37 (0.03)**	-0.01 (0.04)	-1.56 (0.12)**	-0.56 (0.09)**
SES	0.07 (0.02)**	0.06 (0.02)**	1.18 (0.04)**	0.34 (0.04)**
Low Track			-0.91 (0.03)**	-0.25 (0.03)**
Combined Tr.			-0.20 (0.02)**	-0.08 (0.03)**
Intermediate Tr.			0.18 (0.03)**	-0.01 (0.03)
Academic Tr.			0.93 (0.03)**	0.34 (0.03)**

Note. SES = Socioeconomic background; Tr. = Track

* *p* ≤ .05; ** *p* ≤ .01

Table 4

Regression Weights for Student- and School-Level Characteristics Predicting Initial Levels and Position Effects. Results from Multivariate Models Applied to each Domain.

	Student Level		School Level	
	Initial Level	Position Effect	Initial Level	Position Effect
	Est.(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)	*Est.*(*SE*)
Science				
Male	0.26 (0.01)**	-0.14 (0.01)**		
Migration	-0.31 (0.01)**	-0.08 (0.01)**	-0.38 (0.05)**	-0.11 (0.03)**
SES	0.08 (0.01)**	0.00 (0.00)	0.34 (0.03)**	-0.01 (0.01)
Low Track			-0.52 (0.02)**	-0.14 (0.01)**
Combined Tr.			-0.17 (0.01)**	-0.03 (0.01)**
Intermediate Tr.			0.10 (0.01)**	0.02 (0.01)**
Academic Tr.			0.59 (0.02)**	0.15 (0.01)**
Mathematics				
Male	0.41 (0.01)**	-0.16 (0.02)**		
Migration	-0.23 (0.01)**	-0.14 (0.02)**	-0.42 (0.06)**	-0.29 (0.06)**
SES	0.08 (0.01)**	-0.02 (0.01)*	0.42 (0.03)**	-0.07 (0.04)
Low Track			-0.61 (0.02)**	-0.09 (0.03)**
Combined Tr.			-0.24 (0.02)**	-0.04 (0.02)*
Intermediate Tr.			0.09 (0.02)**	0.03 (0.01)*
Academic Tr.			0.76 (0.02)**	0.10 (0.02)**
Reading				
Male	-0.17 (0.02)**	-0.10 (0.04)**		
Migration	-0.36 (0.03)**	0.01 (0.05)	-0.18 (0.07)**	-0.21 (0.10)*
SES	0.04 (0.02)**	0.07 (0.02)**	0.43 (0.05)**	-0.01 (0.06)
Low Track			-0.62 (0.03)**	-0.20 (0.04)**
Combined Tr.			-0.15 (0.02)**	-0.08 (0.03)**
Intermediate Tr.			0.20 (0.02)**	-0.01 (0.03)
Academic Tr.			0.56 (0.03)**	0.29 (0.04)**

Note. SES = Socioeconomic background

* $p \le .05$; ** $p \le .01$

Table 5

Average Effect Sizes (Item Cluster Scores), Effects on Conventionally Scored Tests (Conventional Scoring), and Differences to Effects Sizes Adjusted for Position Effects (Difference) for Student- and School-Level Characteristics.

	Male	Migration Background		Socioeconomic Background		Low Track	Combined Tracks	Intermed. Track	Academic Track
	Student	Student	School	Student	School	School	School	School	School
Science									
Item cluster scores									
Avg. Effect Size	18.9	-36.1	-56.3	21.9	102.4	-79.9	-21.1	8.9	92.1
Difference (%)	**6.3(33.5)**	3.5(-9.8)	5.0(-8.9)	-0.1(-0.4)	-8.7(-8.5)	7.2(-9.1)	1.5(-7.1)	-1.0(-11.2)	-7.7(-8.4)
Conventional Scoring									
Effect Size	20.2	-36.5	-53.4	22.0	99.5	-77.9	-20.7	8.8	89.8
Difference (%)	**4.9(24.4)**	**4.0(-11.0)**	2.1(-3.9)	-0.2(-0.9)	-5.8(-5.9)	5.2(-6.7)	1.1(-5.4)	-0.8(-9.6)	-5.5(-6.1)
Mathematics									
Item cluster scores									
Avg. Effect Size	28.2	-24.9	-54.6	17.3	98.8	-72.6	-24.1	6.7	90.0
Difference (%)	**4.8(17.5)**	**4.1(-16.3)**	3.2(-5.8)	1.1(6.1)	-3.3(-3.3)	2.9(-4.0)	1.0(-4.1)	-1.1(-15.8)	-2.9(-3.2)
Conventional Scoring									
Effect Size	29.1	-24.4	-50.7	18.0	94.1	-71.0	-22.4	6.2	87.3
Difference (%)	**3.8(13.1)**	**3.5(-14.4)**	-0.7(1.4)	0.3(1.8)	1.4(1.5)	2.9(-1.8)	-0.6(2.8)	-0.6(-10.0)	-0.2(-0.2)
Reading									
Item cluster scores									
Avg. Effect Size	-17.2	-30.4	-51.9	16.7	102.2	-83.4	-19.2	14.1	88.5
Difference (%)	**3.8(-21.8)**	0.4(1.2)	**8.2(-15.8)**	**-5.5(-32.9)**	**-13.2(-13.0)**	**10.5(-12.5)**	**3.5(-18.1)**	0.5(3.5)	**-14.4(-16.2)**
Conventional Scoring									
Effect Size	-16.1	-28.9	-53.3	16.4	101.4	-84.5	-19.6	16.0	88.1
Difference (%)	**2.6(-16.4)**	-1.0(3.6)	**9.5(-17.9)**	**-5.2(-31.6)**	**-12.5(-12.3)**	**11.5(-13.7)**	**3.8(-19.7)**	-1.4(-8.5)	**-13.9(-15.8)**

Note. SES = Socioeconomic background, A = Differences between dichotomous categories, B = Difference between the 80th and 20th percentiles of the covariate distribution, C = Differences relative to the average proficiency score in school types.

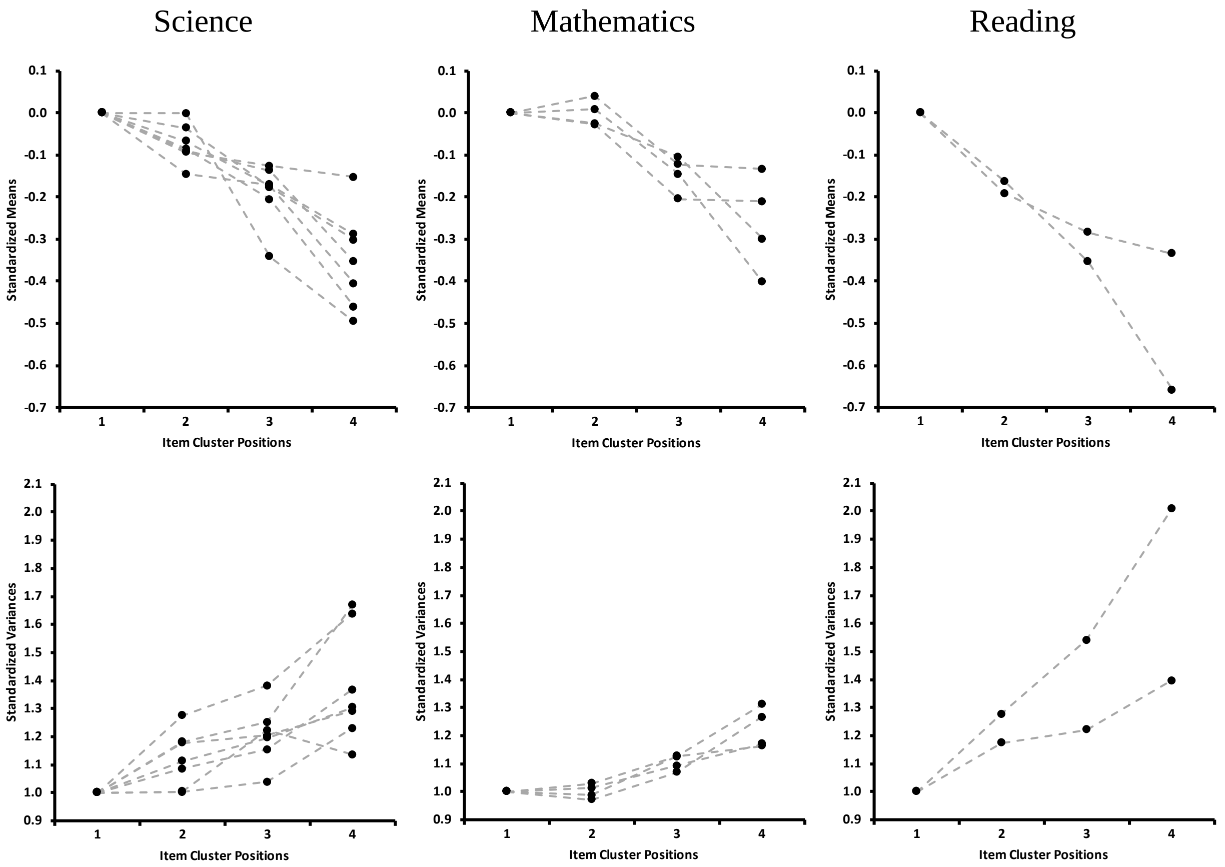

Figure 1. Changes in item cluster means (upper panels), and item cluster dispersions (middle panels) by item cluster positions. Results on the basis of item cluster scores standardized with respect to the first position.

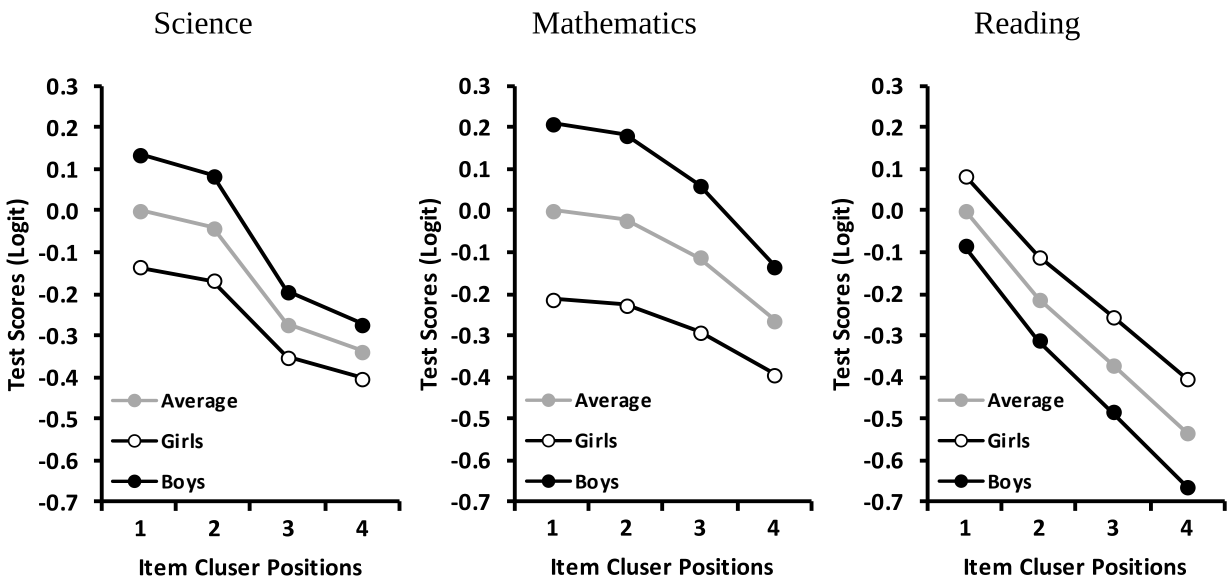

Figure 2. Item cluster scores by item cluster positions for boys and girls.