Psychological Test and Assessment Modeling, Volume 54, 2012 (4), 418-431

A multilevel item response model for item position effects and individual persistence

Johannes Hartig[1] *& Janine Buchholz*[2]

Abstract

The paper presents a multilevel item response model for item position effects. It includes individual differences regarding the position effect to which we refer to as the persistence of the test-takers. The model is applied to published data from the PISA 2006 science assessment. We analyzed responses to 103 science test items from N = 64.251 students from 10 countries selected to cover a wide range of national performance levels. All effects of interest were analyzed separately for each country. A significant negative effect of item position on performance was found in all countries, which is more prominent in countries with a lower national performance level. The individual differences in persistence were relatively small in all countries, but more pronounced in countries with lower performance levels. Students' performance level is practically uncorrelated with persistence in high performing countries, while it is negatively correlated within low performing countries.

Key words: item response theory, item position effects

[1] *Correspondence concerning this article should be addressed to:* Johannes Hartig, PhD, German Institute for International Educational Research (DIPF), Schloßstr. 29, 60486 Frankfurt am Main, Germany; email: hartig@dipf.de

[2] German Institute for International Educational Research (DIPF), Frankfurt am Main, Germany

In standardized assessments, performance of test-takers on test items can be affected by the position of the items within the test. Within long tests, performance may decrease due to fatigue or declining motivation. Performance may also increase due to learning or practice effects during the test. Regardless of the direction, these item position effects violate assumptions made in most item response theory (IRT) measurement models, since the probability for correct responses usually is assumed to depend only on properties of items and persons, which are assumed to be independent from presentation conditions and item context. Therefore, the examination of item position effects and, if necessary, their inclusion in an appropriate measurement model is an advisable undertaking from a psychometric viewpoint.

An examination of item position effects is also interesting from an applied perspective. For example, if negative item position effects on performance are known, the maximum test length that can be administered to test-takers without overly impairing the assessed performance can be determined. More importantly, if test scores are described with reference to construct maps which are based on item difficulties (e.g., Wilson, 2005), effects of the item position should be separated from differences in item difficulties due to item content.

Effects of the item position have been repeatedly investigated with different results. For example, Whitely and Dawis (1976) found different difficulties for items presented at different positions within different test forms. Kingston and Dorans (1982, 1984) found positive position effects for some sections of the Graduate record examinations (GRE) aptitude test which they attribute to a practice effect whose occurrence depends on the item type. Meyers, Miller, and Way (2009) found that changes in item positions between test forms are positively related to changes in item difficulty. Davis and Ferdous (2005) found fatigue effects for reading and math tests in grade 3 and 5. Hohensinn et al. (2008) found a small fatigue effect in a 4th grade math assessment, which was however not replicated in a follow-up study (Hohensinn, Kubinger, Reif, Schleicher, and Khorramdel, 2011). Schweizer, Schreiner, and Gold (2009) and Schweizer, Troche, and Rammsayer (2010) found substantive individual differences in learning effects within the Advanced Progressive Matrices (APM).

Within this paper, we will present an IRT model that can account for different possible item position effects. It is based on the logistic Rasch model for dichotomous responses (with $logit(y) \equiv e^y / (1+e^y)$):

$$P(X_{pi}=1) = logit(\theta_p - \beta_i) \tag{1}$$

Here, $P(X_{pi}=1)$ is the probability for person p answering item i correctly, θ_p is the ability of person p, and β_i is the difficulty of an item i. In the following sections, we will briefly outline general approaches to model item position effects from an item side as well as from an individual differences perspective. Subsequently, we will present the modeling approach used in our study and the research questions linked to the effects estimated within this model.

Effects on the item side

Item position effects can be viewed from the item side, meaning changes of performances related to the item position within a test can be translated into an effect of the item position on the item difficulty β_i. From a measurement model perspective, this means that the item difficulties change depending on the position within a test at which an item is presented, which would violate assumptions in measurement models treating item difficulties as fixed and independent from the presentation context. To account for item position effects on performance, the Rasch model could be extended by a parameter that represents the changes in difficulty across time, as suggested in the dynamic Rasch model proposed by Verhelst and Glas (1993) to capture learning processes.

In tests with fixed item orders, however, the difficulty of the actual item content and effects of the item position within the test cannot be separated since every item is always presented at the same position, e.g. at the beginning or the end of the test. Only if item position and item content are not confounded, the examination of item position effects on item difficulties is possible. This is the case if the same items are presented to different test-takers in different orders, as applied in research designs for item context effects (e.g. Hartig, Hölzel, & Moosbrugger, 2007; Knowles, 1988). A separation of item position and item content is also given in data from large scale studies on student achievement that implement balanced booklet designs, as in the OECD Programme for International Student Assessment (PISA) or the Trends in International Mathematics and Science Study (TIMSS).

Individual differences

The effects of the item position on performance can also be viewed from the perspective of the test-taker, meaning the ability θ_p required to answer the items changes across the test. As long as this effect is the same for all test-takers, effects can be represented either in θ_p or in β_i. However, in addition to the general effect of item position (e.g. a general increase or decrease in performance), test-takers may also *differ individually* in item position effects. The performance of some test-takers may remain stable even across long tests, while the performance of others may decrease or increase faster. This kind of effect would imply systematic dependencies between responses not accounted for by the ability θ_p, and thus would introduce a violation of the assumption of local independence made in most measurement models. To account for individual differences in item position effects, a second latent variable capturing these differences across persons can be included in the model. This approach has been chosen within a structural equation modeling (SEM) framework by Schweizer et al. (2009, 2010) to analyze data from a test presented in a fixed item order; they used a latent variable defined by increasing loadings – similar to a latent growth curve model – to capture individual differences in learning taking place during the course of assessment.

Apart from capturing unaccounted-for local dependencies, analyzing individual differences in item position effects is appealing as these individual differences reflect the

individual tendency to change performance in the course of an assessment. This individual tendency can be interpreted as a test-taker characteristic of its own right, providing additional diagnostic information over and above the individual performance level. Within this paper we are assuming that test performance will probably *decrease* in student assessments and thus refer to the individual effect of the item position as *persistence*: low individual persistence means decreasing performance, high persistence means stable or even increasing performance throughout the assessment. We do so to provide a convenient label to the construct, well knowing that many other processes and more specific substantive constructs (e.g. test motivation or the ability to learn during a test) are also plausible candidates to explain individual differences in item position effects.

Modeling approach

In the present study, effects of the item position are examined on the item as well as on the person side. A model was implemented that can be regarded as a generalization of the Rasch model for dichotomous responses. It takes into account a general effect of the item position on the item difficulty as well as individual differences in this effect (see Equation 1):

$$P\left(X_{pi}=1\right)=logit\left(\theta_p+t_{pi}\left(\gamma+\delta_p\right)-\beta_i\right) \tag{2}$$

Again, $P\left(X_{pi}=1\right)$ is the probability for person p answering item i correctly, θ_p is the ability of person p, and β_i is the difficulty of an item i. Furthermore, t_{pi} is the position within the test at which item i is presented to person p. γ is the linear effect of the item position on item difficulty across the test. A negative estimate for γ indicates a linearly declining, a positive value an increasing performance across item positions. δ_p is a random effect with $\delta_p \sim N\left(0,\sigma_\delta^2\right)$ which captures individual differences in the linear item position effect, similar to the slope factor in a linear growth curve model. Within this paper, we will refer to the individual differences captured in the random effect δ_p as persistence. The sum of the parameters γ and δ_p determine an individual test-taker's trend of performance across item positions – a positive sum indicates increasing performance, a negative sum decreasing performance during the test.

The model can be regarded as an extension of the dynamic Rasch model presented by Verhelst and Glas (1993), including not only a main effect of the item position but also individual differences with respect to this effect. It has to be noted that the model is a restriction of the Verhelst and Glas (1993) model in that the item position effect is constant for all items. Following the adaption of the model given by Verguts and De Boeck (2000), item differences in the position effect are not allowed for. Also, in contrast to the estimation approach chosen by Verguts and De Boeck, the model presented in Equation (2) is implemented as a multilevel IRT model, see the section "Data Analysis" below. A further noteworthy distinction regarding the intended applications is that the dynamic Rasch model is primarily intended to measure learning and models the effect of the number of items answered correctly, while we are interested in general effects of the item position and thus model the effect of the total number of items presented to the test-

taker, regardless of the success in answering them. If the individual position effect would be interpreted as a learning curve, this would correspond to a "noncontingent learning model", in which success in each item is affected by the "number of items previously attempted" (Verguts & De Boeck, 2000, p. 152).

It also has to be noted that a linear trend for the effect of the item position is of course a very restrictive approach and the actual effects of item position can have other and more complex forms. Alternatively to a linear trend modeled in Equation (2), other curves (e.g. including a quadratic trend) could be considered. It would also be possible to introduce dummy variables for each position greater than 0, thus completely freely estimating the changes in performance across the test – this approach was chosen by Alexandrowicz and Matschinger (2008). The linear model has the obvious advantage that both the fixed effect γ as well as the person-level random effect δ_p are more straightforward to interpret than in a model including more fixed or even more random effects. Nevertheless, the possibility of non-linear trends in performance should be considered and the adequacy of the linearity assumption should be checked empirically.

The interindividual variance σ^2_δ in the position effect (i.e. in persistence) is a model parameter of utmost substantive interest for our analyses. If this variance is different from zero, this indicates that from a technical viewpoint, local dependencies exist which are not accounted for by a unidimensional measurement model, and that these dependencies can be represented by individual differences in a linear trend across item positions. From a diagnostic viewpoint, it indicates that there are systematic individual differences in persistence, which might provide diagnostic information in addition to the ability measured by the test.

Another parameter of substantive interest which can be derived from the model applied in this study is the correlation $r_{\theta,\delta}$ between the individual ability θ_p and the individual persistence δ_p. If this correlation is positive, test-takers with a high ability (i.e. overall performance level) tend to be less affected by negative effects of the test length or may even increase their performance across the test, while test-takers with a low ability show a declining performance. A negative correlation $r_{\theta,\delta}$ would indicate that high ability test-takers have a stronger decline (or weaker increase) of performance across the test.

Research questions

The present study examines general item position effects and individual differences in persistence using public data from the PISA 2006 science assessment. Given the heterogeneous findings on item position effects, the PISA database provides an attractive possibility to provide evidence regarding these effects within student assessments on a sound empirical basis. The domain of science was chosen because at the start of our study the dataset from 2006 was the latest dataset publicly available and science was the main domain in the 2006 cycle of the PISA studies. Therefore, the most test items presented to students were science items, and all of the participating students answered science items (while not all students answered reading or mathematics items). Thus, the science assessment data provide the best empirical basis to study position effects within

the PISA 2006 test. We assume the effects to be observed in the science assessment to be of a general nature and have no reason to believe that they should differ from effects to be found within other content domains.

The first aim of our study is to assess the size of a general item position effect in the PISA science assessment. Given the heterogeneous item formats and the lack of feedback within the test, fatigue or declining motivation are more likely to occur than learning, thus we assume a generally negative effect of the item position on performance:

> H1: Item difficulty increases with the position of items within the PISA science assessment.

Second, the variance in students' persistence will be estimated. Given item-position related individual differences found in other studies (e.g. Schweizer et al., 2009, 2010) we assume that there are significant variance components in the position effect.

> H2: The variance of the individual differences in the item position effect is significantly different from zero.

Finally, the correlation between students' performance level and persistence can be estimated. A positive correlation implies that the performance of low performing students decreases faster in the course of the assessment, whereas a negative correlation indicates that higher performing students tend to decrease their performance. Given the findings of Schweizer et al. (2009, 2010) (although in a different content domain), we expect rather a positive than a negative correlation:

> H3: The correlation between students' performance level and the individual differences in the item position effect is greater than zero.

Since the PISA data set offers the opportunity to examine effects of the countries, the consistency of item position effects across different countries with different national performance levels can be examined. To our knowledge, there have been no empirical studies of item position effects in different countries, thus there is no basis to state hypotheses regarding specific differences in item position effects across countries or relations of the relative strength of these effects with country level variables. However, since there is no basis to assume differences between countries either, we do assume that the expected effects can be found in all countries. As an exploratory analysis, the country-level correlations of the item position effects with the national performance level in science (as reported by the OCED) will be inspected. The information which kind of effects can be observed within high and low performing countries might provide insights about the nature of these effects.

Method

Data

The analyses make use of the published science assessment data from the PISA 2006 study (OECD, 2009) consisting of data from N = 397.920 students from 57 countries.

For the analyses, ten countries were selected that cover the range of national performance levels as reported for the PISA 2006 science assessment (OECD, 2007).

Table 1 shows the selected countries, their national science score in PISA 2006, and the number of students from these countries included in the analyses.

The PISA science test items are comprised of a variety of open-ended and closed response formats; all items were selected to fit the unidimensional Rasch model (OECD, 2009). Responses to 103 science test items were used in the analysis. 97 of these items had a dichotomous response format; six items had a polytomous response format with three possible outcomes (incorrect, partial credit, and full credit). To apply the multilevel model which requires a consistent response format across all items, the polytomous items were dichotomized scoring the full credit as correct (1) and all other scores as incorrect (0).

In PISA 2006, items were presented in 13 *item clusters* that were arranged in a fully-linked, balanced incomplete block design with each booklet consisting of four item clusters (OECD, 2009). Each item cluster was presented equally often at all of the four possible cluster positions within the booklet, i.e. cluster position and cluster content were varied independently. In our analyses, position of the cluster in which the item was presented is used as the position t_{pi} in Equation (2). This means we are not using the exact item position within the test but an approximation thereof provided by the cluster position, which is coded from 0 to 3 for the analyses.

Table 1:
Countries selected for the analyses, their national science score in PISA 2006 (OECD, 2007), and the number of students (N) from these countries included in the analyses.

Country	International Science Score	N
1. Brazil	390	9273
2. Finland	563	4712
3. Japan	531	5940
4. Poland	498	5547
5. South Korea	522	5174
6. Sweden	503	4437
7. Thailand	421	6189
8. Tunisia	386	4940
9. Turkey	424	4940
10. United Kingdom	515	13099

Data analysis

The model presented in Equation (2) was implemented as a logistic multilevel model with item responses as level one variable nested within students (e. g. Alexandrowicz & Matschinger, 2008; Kamata, Bauer & Miyazaki 2008; Kamata & Cheong, 2007). The advantage of this approach is that item content and item position can vary independently as response level (level one) variables, and from a multilevel data perspective there are no missing values since there is only one response variable, and only items actually attempted by each student are included. There is no need to construct virtual items for each combination of item and item position as in the procedures described for a "wide" data structure with responses for each individual in one row (e.g. Hohensinn et al., 2011; Verguts & De Boeck, 2000; Verhelst & Glas, 1993). For the PISA science assessment data, the latter procedure would lead to a very large number of virtual items (103 items × 4 cluster positions) and, due to the booklet design, result in a very sparse data matrix. (In the original PISA booklet design, 4 out of 13 clusters are answered by each student, meaning ca. 70% of the possible responses are missing by design. Constructing four virtual items for each item in the test would lead to approximately 93% missing responses.)

To estimate item difficulties β_i, the items were dummy coded on level one, using the first item as reference category. With only the fixed effects for item difficulties, the multilevel model is equivalent to the dichotomous Rasch model. The main effect γ for the item position was estimated by introducing the item cluster position within the booklet (coded starting with 0 for the first position) as an item level predictor. Additionally, a random effect associated with the position effect is introduced to capture individual differences δ_p in persistence.

The analyses were conducted separately for each of the countries, using the student weight provided with the PISA data set as a level 2 weighting variable. All analyses were conducted with the multilevel software HLM 6 (Raudenbush, Bryk, & Congdon, 2004). For each country, the analyses provide the following results of primary interest: (1) the general effect γ of the item position, (2) the interindividual variance σ_δ^2 in persistence, and (3) the correlation $r_{\theta,\delta}$ between the individual ability θ_p and the persistence δ_p.

Results

To obtain an overall impression of the position effect across the cluster position, preliminary to the model-based analyses the percentage of correct responses (based on the dichotomization described above) was calculated for each cluster position within each country; the results are shown in Table 2. It can be seen that there is a substantial decline in performance across cluster position, the difference between the first and last cluster ranges from 3% correct responses in Finland to 8% in Thailand. Additionally, Table 2 shows the correlations of the percentage of correct responses and the cluster position across the four positions. It can be seen that the trend in the performance decline is almost perfectly linear in all countries. The minimum of variance explained by a linear

trend is 87% for Poland. These numbers strongly support the assumption of a linear trend in performance for the data analyzed for this study. Given the minimal deviations from linearity in the raw data, we abstained from applying models with more complex trends.

Table 2:
Percentage of correct responses for each country depending on the item cluster position, and correlations between the percentage of correct answers and the cluster position.

Country	Position 1	Position 2	Position 3	Position 4	Difference 4-1	$r_{(\%\ corr.,\ position)}$
Brazil	35%	33%	31%	29%	7%	-1.000
Finland	67%	66%	65%	64%	3%	-1.000
Japan	63%	60%	58%	55%	7%	-.997
Poland	56%	55%	54%	50%	7%	-.933
South Korea	59%	57%	56%	54%	5%	-.992
Sweden	57%	55%	54%	51%	6%	-.981
Thailand	42%	39%	38%	34%	8%	-.977
Tunisia	35%	33%	31%	28%	7%	-.994
Turkey	42%	39%	38%	34%	7%	-.977
United Kingdom	59%	56%	55%	52%	6%	-.984

Note: The descriptive results were generated without using student weights. $r_{(\%\ corr.,\ position)}$ = Correlation between percentage of correct answers and the cluster position.

Model based results

Table 3 displays the results from the multilevel IRT analyses for each country: fixed effects for the item cluster position, variance components for students' ability level and students' persistence, and the correlation between ability level and persistence. Not surprisingly, given the huge sample sizes on both levels, all fixed effects and variance components, even when numerically small, are significantly different from zero on a 1% alpha level. We therefore refrain from reporting standard errors for the fixed effects. The effects γ within each country are not directly comparable since item difficulties were estimated freely within each country and the θ-scale is not the same across countries. To facilitate the comparison of the cluster position effect between countries, a coefficient γ^* which is standardized using the standard deviation of the ability level within each country is reported in addition to the original parameter γ from Equation (2). The standardized coefficient can be read as the effect of one cluster position on performance in standard deviations in θ within each country.

Table 3:
Fixed effects for the item cluster position, variance components for students' ability level and students' persistence, and correlation between ability level and persistence within each country estimated in the multilevel IRT analyses.

Country	Internat. Science Score	γ	γ^*	σ^2_θ	σ^2_δ	$r_{\theta,\delta}$
Brazil	390	-0.115	-0.129	0.801	0.023	-0.234
Finland	563	-0.060	-0.065	0.854	0.012	-0.032
Japan	531	-0.127	-0.123	1.063	0.027	0.001
Poland	498	-0.095	-0.098	0.945	0.016	-0.045
South Korea	522	-0.072	-0.074	0.938	0.013	-0.019
Sweden	503	-0.100	-0.101	0.988	0.013	-0.017
Thailand	421	-0.130	-0.157	0.683	0.016	-0.306
Tunisia	386	-0.136	-0.163	0.696	0.031	-0.388
Turkey	424	-0.137	-0.157	0.763	0.025	-0.198
United Kingdom	515	-0.102	-0.093	1.216	0.018	-0.099

Note: γ: fixed effect of cluster position; γ^*: fixed effect of cluster position standardized based on the *SD* of the ability level within each country; σ^2_θ: variance of students' ability level; σ^2_δ: variance of students' persistence; $r_{\theta,\delta}$: correlation between ability level and persistence. All parameters are significantly different from zero on a 1% alpha level.

Hypothesis 1 is supported by a significant negative effect of the cluster position on the probability of correct responses – indicating increasing item difficulties – found consistently across all countries. On the "national scales" constructed in the IRT analyses for the ability level θ_p within each country, the effects range from -.065 SD in Finland to -.163 SD in Tunisia. Since the effect γ^* indicates the effect for one cluster position, under the assumption of linearity of the item position effect the total effect occurring across the assessment can be obtained by multiplying γ^* with three. This means the difference between the first and last item cluster corresponds to the difference in performance that would be expected between individuals about half a standard deviation (-0.489) apart in Tunisia and about one fifth standard deviation (-0.194) apart in Finland. These numbers illustrate that the effects are not only statistically significant but also indicate substantial changes in performance, roughly corresponding small to medium effect sizes in Cohen's (1988) classification, using the classification of the effect size *d* for mean differences between independent samples as an orientation.

Hypothesis 2 is supported by significant, although numerically small variances of the individual differences in the cluster position effect (i.e. persistence) in all countries. This indicates that some local dependencies between the science test items exist that can be represented by individual differences in a linear trend across cluster positions.

Hypothesis 3, however, is clearly not supported. The correlations between students' performance level and persistence are either very close to zero (e.g. in Japan, Sweden, and South Korea) or slightly negative (e.g. in Tunisia, Thailand, and Brazil). Combined with the generally increasing item difficulties, this shows that ability and persistence are either unrelated or that students with a high overall performance level have a stronger decline of performance across the test.

Exploratory analyses

To gain further insight into the nature of the position effects observed in the different countries, country-level correlations were calculated between the national performance level in science (as reported by OCED, 2007), the general cluster position effect, the variance of students' persistence, and the correlation between ability level and persistence. The resulting pattern of relations is shown in Table 4.

Surprisingly, the effects estimated for each country are strongly related to the national science score. In countries with higher overall performance in science, the negative cluster position effect is less pronounced, the interindividual variance in persistence is lower, and the correlation between ability level and persistence is higher (i.e. closer to zero, see Table 3). It can also be seen that the effects themselves are correlated on country level, pronounced position effects are associated with higher variance in persistence and a higher correlation between ability and persistence.

Table 4:
Country level correlations between the international science score, the standardized cluster position effect, the variance of students' persistence, and the correlation between ability level and persistence across the ten analyzed countries.

	(1)	**(2)**	**(3)**
(1) International Science Score			
(2) γ^*	0.86		
(3) σ^2_δ	-0.62	-0.76	
(4) $r_{\theta,\delta}$	0.91	0.82	-0.58

Note: γ^* : fixed effect of cluster position standardized based on the *SD* of the ability level within each country; σ^2_δ : variance of students' persistence; $r_{\theta,\delta}$: correlation between ability level and persistence.

Discussion

Within this paper, we presented a multilevel IRT model that can account for item position effects as well as individual differences regarding this effect. Since the model can be specified as a generalized linear mixed model (e.g. Skrondal & Rabe-Hesketh, 2004), the model parameters can be estimated with standard software. The analyses presented demonstrate that the model is applicable even to data from large samples. In the next sections, we will discuss possible substantive implications of our findings as well as technical limitations, including potential further developments.

Substantive implications

The general effect of the cluster position within the PISA 2006 science assessment confirms previous findings on negative effects of test length on performance in achievement tests (e.g. Hohensinn et al., 2008; Meyers et al., 2009). The relatively small size of these effects does not indicate that the two-hour testing session for PISA 2006 was generally too long. However, given the stronger position effects in low performing countries, the possibility that the differences between countries would be less pronounced if the test had been shorter cannot be ruled out. Particularly the negative correlations between performance level and persistence in the low performing countries may indicate that one (although small) contribution to these countries' test results may be that higher performing students aren't able to keep up their performance level across the whole assessment.

While small but significant variances in persistence were found in all countries, these individual differences were more pronounced in countries with low performance levels (e.g. Brazil and Turkey) than in high performing countries (e.g. Finland and Korea). For future studies, the nature of these individual differences could be investigated by examining the correlations with student variables (e.g. gender, test motivation, or cultural resources at home). Additionally, the assumption that the effects found for the domain of science are similar in other content domains should be investigated empirically.

Technical limitations

With the estimation of the model with the HLM software, absolute model or item fit statistics are not available. We assume that since the PISA tests sufficiently fit the Rasch model (OECD, 2009), a more complex model should fit the data at least equally well. Nevertheless, it appears desirable to develop specific statistics for model and item fit for the model used in the analyses. Generally, the model could also easily be specified and estimated using a virtual item approach as used by Hohensinn et al. (2011) or Verguts and De Boeck (2000) and specifying individual differences in the item position effect as a second dimension in a multidimensional Rasch model. This approach would allow the use of different estimation algorithms (particularly marginal maximum likelihood estimation) and would probably allow to generate more information on model and item fit. However, given the sparse data matrix resulting from the construction of virtual items

(93% missing by design), it seems questionable if this approach would yield reliable results for the PISA data set.

Another issue that should be addressed in future analyses of item position effects in student assessments is the sample structure. Since students in the PISA sample are nested within schools, standard errors estimated without taking this into account are probably too small (e.g. Hox, 2010). One way to deal with this problem would be using resampling techniques (e.g. Fay, 1989) to correct the standard errors, a method which is not available in HLM. Another possibility is to extend the model in Equation 2 to explicitly model the school level as another hierarchical level, i.e. decomposing the ability and persistence in variation within and between schools. For the present study, we refrained from that extension because we had no substantive assumptions regarding the school level. A further extension of the model could treat the country level as another hierarchical level of the data structure instead of running analyses country by country. Although attractive to obtain direct estimates for between-country variances in the parameters of interest, the resulting total sample size would pose demands on computing power that would by far exceed the resources of even powerful personal computers.

Finally, we modeled a linear effect of the cluster position, which is in the case of the PISA data at least supported by the linear decline in performance shown in Table 2. For other applications we recommend always to closely inspect the exact form of the position effect before applying and interpreting the linear model used in this study.

References

Alexandrowicz R. & Matschinger H. (2008). Estimating Item Location Effects by Means of a Generalized Logistic Regression Model. *Psychology Science Quarterly, 50,* 64–74.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd edition)*. Hillsdale, NJ: Lawrence Erlbaum.

Davis, J., & Ferdous, A. (2005). *Using item difficulty and item position to measure test fatigue.* Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec.

Fay, R. E. (1989). Theoretical application of weighting for variance calculation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* 212–217.

Hartig, J., Hölzel, B., & Moosbrugger, H. (2007). A confirmatory analysis of item reliability trends (CAIRT): Differentiating true score and error variance in the analysis of item context effects. *Multivariate Behavioral Research Methods, 42,* 157–183.

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 booklet design within large-scale assessment. *Educational Research and Evaluation, 17,* 497–509.

Hox, J. (2010). *Multilevel analysis. Techniques and applications (2nd edition)*. New York: Routledge.

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement modeling. In A. A. O'Connell & D. B. McCoach (Eds.). *Multilevel modeling of educational data* (pp. 345–388). Charlotte: Information Age Publishing.

Kamata, A., & Cheong, F. (2007). Multilevel Rasch model. In M. von Davier & C. H. Carstensen (Eds). *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 217–232). Springer.

Kingston, N. M., & Dorans, N. J. (1982). The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory. *GRE Board Professional Report 79-12bP*. Princeton NJ: Educational Testing Service.

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.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology, 55,* 312–320.

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.

OECD (2007). PISA 2006. *Science competencies for tomorrow's world. Volume 1: Analysis.* Paris: OECD.

OECD (2009). *PISA 2006 Technical Report.* Paris: OECD.

Raudenbush, S.W., Bryk, A.S, & Congdon, R. (2004). HLM 6 for Windows [Computer software]. Skokie, IL: Scientific Software International, Inc.

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.

Schweizer, K., Troche, S. J., & Rammsayer, T. H. (2010). On the special relationship between fluid and general intelligence: New evidence obtained by considering the position effect. *Personality and Individual Differences, 50,* 1249–1254.

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling. Multilevel, longitudinal and structural equation models.* Boca Raton: Chapman & Hall

Verguts, T., & De Boeck, P. (2000). A Rasch model for detecting learning while solving an intelligence test. *Applied Psychological Measurement, 24,* 151–162.

Verhelst, N. D., & Glas, C. A. W. (1993). A dynamic generalization of the Rasch model. *Psychometrika, 58,* 395–415.

Wilson, M. (2005). *Constructing Measures: An Item-response Modeling Approach.* Mahwah, NJ: Lawrence Erlbaum.

Whitely, S. E. & Dawis, R. V. (1976). The influence of test context on item difficulty. *Educational and Psychological Measurement, 36,* 329–337.