JEM JOURNAL of EDUCATIONAL MEASUREMENT
NCME national council on measurement in education

Journal of Educational Measurement
Summer 2013, Vol. 50, No. 2, pp. 164–185

Modeling Item-Position Effects Within an IRT Framework

Dries Debeer and Rianne Janssen
University of Leuven

Changing the order of items between alternate test forms to prevent copying and to enhance test security is a common practice in achievement testing. However, these changes in item order may affect item and test characteristics. Several procedures have been proposed for studying these item-order effects. The present study explores the use of descriptive and explanatory models from item response theory for detecting and modeling these effects in a one-step procedure. The framework also allows for consideration of the impact of individual differences in position effect on item difficulty. A simulation was conducted to investigate the impact of a position effect on parameter recovery in a Rasch model. As an illustration, the framework was applied to a listening comprehension test for French as a foreign language and to data from the PISA 2006 assessment.

In achievement testing, administering the same set of items in different orders is a common strategy to prevent copying and to enhance test security. These item-order manipulations across alternate test forms, however, may not be without consequence. After the early work of Mollenkopf (1950), it repeatedly has been shown that changes in the placement of items may have unintended effects on test and item characteristics (Leary & Dorans, 1985). Traditionally, two kinds of item-position effects have been discerned (Kingston & Dorans, 1984): a practice or a learning effect occurs when the items become easier in later positions, and a fatigue effect occurs when items become more difficult if placed towards the end of the test. Recent empirical studies on the effect of item position include Hohensinn et al. (2008), Meyers, Miller, and Way (2009), Moses, Yang and Wilson (2007), Pommerich and Harris (2003), and Schweizer, Schreiner and Gold (2009).

In the present article, item-position effects will be studied within De Boeck and Wilson's (2004) framework of descriptive and explanatory item response models. It will be argued that modeling item-position effects across alternate test forms can be considered as a special case of differential item functioning (DIF). Apart from the DIF approach, the linear logistic test model of Fischer (1973) and its random-weights extension (Rijmen & De Boeck, 2002) will be used to investigate the effect of item position on individual item parameters and to model the trend of item-position effects across items. A new feature of the approach is that individual differences in the effects of item position on difficulty can be taken into account.

In the following pages we first will present a brief overview of current approaches to studying the impact of item position on test scores and item characteristics. We then present the proposed item response theory (IRT) framework used for modeling item-position effects. After demonstrating the impact of a position effect on parameter recovery with simulated data, the framework is applied to a listening

 Copyright © 2013 by the National Council on Measurement in Education

comprehension test for French as a foreign language and to data from the Program for International Student Assessment (PISA).

Studying the Impact of Item Order on Test Scores

Although interrelated, item-order effects can be distinguished from item-position effects. Item order is a test form property; hence, item-order effects refer to effects observed at the test form level (e.g., the overall sum of correct responses). Item position, on the other hand, is a property of the item. Hence, item-position effects refer to the impact of the position of an item within a test on item characteristics. As will be shown later, item-position effects allow for deriving the implied effects of item order on the test score.

A common approach to studying the effect of item order is to look at the impact of item order on the test scores of alternate test forms which differ only in the order of items and which are administered to randomly equivalent groups. Several procedures have been developed to detect item-order effects in a way that indicates whether equating between the test forms is needed. Hanson (1996) evaluated the differences in test score distributions using loglinear models. Dorans and Lawrence (1990) examined the equivalence between two alternate test forms by comparing a linear equating function of the raw scores for one test form to the raw scores for the other test form with an identity equating function. More recently, Moses et al. (2007) integrated both procedures into the kernel method for observed-score test equating.

In sum, the main purpose of the above procedures is to check the score equivalence of test forms with different item orders that have been administered to random samples of a common population. As a general approach for detecting and modeling item-order and item-position effects, these procedures have certain limitations. First, the effects of item order are only investigated for a particular set of items, making it difficult to generalize the findings to new test forms. Second, the study of item order is limited to a random-groups design with exactly the same items in each alternate test form. Finally, these models only look at the effect of item order on the overall test score. Consequently, item-position effects may remain undetected when the effects of item position cancel out across test forms (as will be shown in the illustration concerning the listening comprehension test). Moreover, focusing on the effect of item position on the overall test score does not allow for an interpretation of the processes (at the item level) underlying the item-order effect.

Studying the Impact of Item Position on Item Characteristics

An alternative approach to modeling the impact of item order is to directly model the effect of item position at the item level using IRT. We first discuss the current use of IRT models to detect item-position effects in a two-step procedure. Afterwards, the framework of descriptive and explanatory IRT models (De Boeck and Wilson, 2004) is used as a flexible tool for modeling different types of item-position effects.

Two-Step Procedures

Within the Rasch model (Rasch, 1960), it repeatedly has been shown that items may differ in difficulty depending on their position within a test form (e.g., Meyers,

et al., 2009; Whitely & Dawis, 1976; Yen, 1980). Common among these studies is the fact that item-position effects are detected in a two-step procedure. First, the item difficulties are estimated in each test form; second, the differences in item difficulty between test forms are considered to be a function of item position. In a recent example of this approach, Meyers et al. (2009) studied the change in Rasch item difficulties between the field form and the operational form of a large-scale assessment. The differences in item difficulty were a function of the change in item position between the two test forms. The model assuming a linear, quadratic and cubic effect provided the best fit, explaining about 56% of the variance of the differences for the math items and 73% of the variance for the reading items.

Modeling Position Effects on Individual Items

The studies using the two-step IRT approach showed that item difficulty may differ between two test forms, the only difference between which is the position of the items in the test forms. These findings may be considered as an instance of differential item functioning (DIF), where group membership is defined by the test form a test taker responded to. Hence, instead of first analyzing test responses for each group and then comparing the item parameter estimates across groups, a one-step procedure seems feasible in which the effect of item position can be distinguished from the general effects of person and item characteristics. Formally, this approach implies that in each test form the probability of a correct answer for person p ($p = 1, 2 \ldots P$) to item i ($i = 1, 2 \ldots I$) in position k ($k = 1, 2 \ldots K$) is a function of the latent trait θ_p and the difficulty β_{ik} for item i at position k. In logit form, this model reads as:

$$logit[Y_{pik} = 1] = \theta_p - \beta_{ik}. \tag{1}$$

When item i is presented at the same position in both test forms, the item has the same difficulty. If not, its difficulty may change across positions.

Using the DIF parameterization of Meulders and Xie (2004), we can decompose (β_{ik}) in (1) into two components:

$$logit[Y_{pik} = 1] = \theta_p - \left(\beta_i + \delta_{ik}^{\beta}\right), \tag{2}$$

where β_i is the difficulty of item i in the reference position (e.g., the position of the item in the first test form) and δ_{ik}^{β} is the DIF parameter or position parameter that models the difference in item difficulty between the reference position and position k in the alternate test form.

The DIF parameterization allows extending the modeling of item-position effects to both the item discrimination α_i and the item difficulty β_i in the two-parameter logistic (2PL) model (Birnbaum, 1968):

$$logit[Y_{pik} = 1] = \left(\alpha_i + \delta_{ik}^{\alpha}\right)\left[\theta_p - \left(\beta_i + \delta_{ik}^{\beta}\right)\right], \tag{3}$$

where δ_{ik}^{α} measures the change in item discrimination depending on the position. This parameter indicates that an item may become more (or less) strongly related to the latent trait if the item appears in a different position in the test. In fact, item-position effects on the discrimination parameter have been studied in the field of personality testing (Steinberg, 1994). More specifically, item responses have been

found to become more reliable (or more discriminating) if they occur towards the end of the test (Hamilton & Shuminsky, 1990; Knowles, 1988; Steinberg, 1994). Up until now, item-position effects on item discrimination have not been found in the field of educational measurement.

Modeling Item-Position Effects Across Items

In (2) and (3), the item-position effects are modeled as an interaction between the item content and the item position. A more restrictive model assumes that the position parameters δ_{ik}^{α} and δ_{ik}^{β} are not item dependent but instead are only position dependent. For example, in (2) one can assume that the item difficulty β_{ik} in (1) can be decomposed into the difficulty of item i (β_i) and the effect of presenting the item in position k (δ_k^{β}):

$$logit[Y_{pik} = 1] = \theta_p - \left(\beta_i + \delta_k^{\beta}\right). \tag{4}$$

For the Rasch model, Kubinger (2008, 2009) derived this model within the LLTM framework.

The model in (4) does not impose any structure on the effects of the different positions. A further restriction is to model the size of the position effects as a function of item position as such by introducing item position into the response function as an explanatory item property (De Boeck & Wilson, 2004). For example, within the Rasch model, one can assume a linear position effect on difficulty:

$$logit[Y_{pik} = 1] = \theta_p - [\beta_i + \gamma(\mathrm{k} - 1)], \tag{5}$$

where γ is the linear weight of the position and β_i is the item difficulty when the item is administered in the first position (when $\mathrm{k} = 1$, the position effect is equal to zero). Depending on the value of γ, a learning effect ($\gamma < 0$) or a fatigue effect ($\gamma < 0$) can be discerned. This model also was proposed by Kubinger (2008, 2009) and by Fischer (1995) for modeling practice effects in the Rasch model. Of course, apart from a linear function, nonlinear functions (quadratic, cubic, exponential, etc.) also are possible.

Modeling Individual Differences in Position Effects

As a final extension of the proposed framework for modeling item-position effects, individual differences in the effect of position can be examined. For example, in (5), γ can be changed into a person-specific weight γ_p. This corresponds to the random weight linear logistic test model as formulated by Rijmen and De Boeck (2002). In a 2PL model, the formulation is analogous:

$$logit[Y_{pik} = 1] = \alpha_i[\theta_p - (\beta_i + \gamma_p(k - 1))]. \tag{6}$$

In (6), γ_p is a normally distributed random effect. In general, γ_p can be considered as a change parameter (Embretson, 1991), indicating the extent to which a person's ability is changing throughout the test. Hence, the model in (6) is two-dimensional and the correlation between γ_p and θ_p also can be estimated.

The use of an additional person dimension to model effects of item position on test responses was proposed by Schweizer et al. (2009) within the structural

equation modeling (SEM) framework. The additional dimension is estimated in a test administration design with a single test form by using a fixed-links confirmatory factor model. More specifically, the factor loadings on the extra dimension were constrained to be a linear or a quadratic function of the position of the item.

A General Framework for Modeling Item-Position and Item-Order Effects

The present framework for modeling item-position effects allows for disentangling the effect of item position from other item characteristics in designs with different test forms. Within the framework, different models are possible. The less restrictive model allows for differences in item parameter estimates across test forms for every item that is included in more than one position across test forms. A more restrictive model is to reduce the observed differences in item parameters across test forms to be a function of item position, changing the model with an item by position interaction into a model with a main effect of item position which is assumed to be a constant effect across test forms. Furthermore, these main effects of item position across test form can be summarized by a trend. This functional form can help practitioners to estimate the size of the item-position effect in new test forms. Finally, individual differences in the trend on item difficulty can be included.

Applicability

The proposed IRT framework for modeling item-position effects can be applied broadly in the field of educational measurement. Because item position is embedded in the measurement model as an item property, the proposed model can deal with different fixed item orders (e.g., reversed item orders across test forms) as well as with random item ordering for every individual test taker separately. Moreover, test forms do not need to consist of the same set of items. As long as there are overlapping (i.e., anchor) items between the different test forms, the impact of item position can be assessed independently of the properties of the item itself.

Although the present framework is focused at the item level, the effect of item position at the test score level also can be captured. The effects on the test score can be seen as aggregates of the position effects on the individual item scores. In an illustration below it will be shown how the test characteristic curve can summarize the effect of item-position effects on the expected test score and how these scores are influenced by individual differences in the size of the linear item-position effect.

Comparison With Other Approaches

As was indicated above, the proposed framework allows for modeling item-position effects in a one-step procedure; this has several advantages in comparison with the current two-step IRT procedures (e.g., having the different test forms on a common scale and testing the significance of additional item-position parameters). The proposed framework also overcomes the above-mentioned limitations of the current approaches for studying the impact of item order on the test scores. First, the item-based approach in principle allows for generalizing found trends in item-position effects to new test forms measuring the same trait in similar conditions. Of

course, the predictions should be checked, as the current knowledge of the occurrence of item-position effects is still limited.

Second, the present framework is applicable in more complex designs than the equivalent-group design with test forms consisting of the same set of items in different orders. Given that the student's ability is taken into account in the proposed IRT framework, the effect of item position also can be investigated in nonequivalent-group designs.

Finally, modeling the effect of item order at the item level can be helpful in looking for an explanation for the found effects. The size and direction of the item-position effects can help in finding an explanation for the effect (see below). Moreover, in the case where individual differences are found in the position effect, explanatory person models (De Boeck & Wilson, 2004) can be used to look for person covariates (e.g., gender, test motivation) that can explain this additional person dimension.

Interpretation of Item-Position Effects on Difficulty

In (4) and (5), a main effect of position on item difficulty is estimated which corresponds to a fixed effect of item position for every test taker. In line with Kingston and Dorans (1984), this effect can be called a practice or learning effect if the items become easier and a fatigue effect if the items become more difficult towards the end of the test. In (6), the effect of item position on difficulty is modeled as a random effect over persons. Again, this parameter may refer to individual differences in learning (if γ_p is positive) or in fatigue (if γ_p is negative).

Although these interpretations are frequently used and also seem self-contained, they can hardly be considered as explanations for the found effects. Instead, explaining a negative γ, for example, by referring to a fatigue effect can be considered as tautological as it is a relabeling of the phenomenon rather than giving a true cause. In fact, explaining item-position effects seems to be similar to explaining DIF across different groups of test takers: one knows that these effects imply some kind of multidimensionality in the data, but as Stout (2002) observed in the case of DIF, it may be hard to indicate on which dimension the different groups of test takers differ. Likewise, when item-position effects are found, this indicates that there is a systematic pattern in the item responses which causes the local item dependence assumption to be violated when these item-position effects are not taken into account in the item response model. However, it may not be clear from the data as such what the cause is of the found effects.

Note that the modeling and interpretation of item-position effects should be distinguished clearly from effects resulting from test speededness. When students are under time pressure, they may start to omit seemingly difficult items (Holman & Glas, 2005) or they may switch to a guessing strategy (e.g., Goegebeur, De Boeck, & Molenberghs, 2010). The present proposed framework, on the other hand, assumes that there is no change in the response process and that the same item response model holds throughout the test (albeit with different position parameters). It also is evident that found item-position effects (especially "fatigue" effects) should not be due to an increasing amount of non-reached items towards the end of the test. Again, item

non-response due to drop out should be modeled with other item response models (e.g., Glas & Pimentel, 2008).

Model Estimation

The proposed models for item-position effects are generalized linear mixed models for the models belonging to the Rasch family or non-linear mixed models for the models belonging to the 2PL family. Consequently, the proposed models can be estimated using general statistical packages (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003; De Boeck & Wilson, 2004). For example, the lmer function from the lme4 package (Bates, Maechler, & Bolker, 2011) of R (R Development Core Team, 2011) provides a very flexible tool for analyzing generalized linear mixed models (De Boeck et al., 2011). Hence, it is well suited for investigating position effects on difficulty in one-parameter logistic models. The NLMIXED procedure in SAS (SAS Institute Inc., 2008) models non-linear mixed effects and therefore can be used to model position effects on difficulty and discrimination in 2PL models (cf. De Boeck & Wilson, 2004). Research indicates that goodness of recovery for the NLMIXED procedure is satisfactory to good (Chen & Wang, 2007; Smits, De Boeck, & Verhelst, 2003; Wang & Jin, 2010; Wang & Liu, 2007). Apart from the lmer and NLMIXED programs, other statistical packages which may rely on other estimation techniques can be used (see De Boeck & Wilson, 2004 for an overview).

Model Identification

For the item-position effects in (2) to (6) to be identifiable, a reference position has to be chosen for which the item-position effect is fixed to zero. For (2) and (3), a reference position has to be defined for every single item. A logical choice is to choose the item positions in one test form. Then, δ_{ik}^{β} expresses the difference in difficulty for an individual item i at position k in comparison with the difficulty of the item in the reference test form. In addition to this dummy coding scheme, contrast coding also can be used when, for example, two test forms have reversed item orders. In this case, the middle position of the test form is considered to be the reference position.

In (4) to (6), the reference position is the same for all items across test forms. For example, in (4), one may choose the first position as the reference position using dummy coding. In this case, δ_{ik}^{β} is the difference in difficulty at position k compared to the first position. In (5) and (6), the first position was chosen as the reference position (γ is multiplied with $(k - 1)$), but any other position can be used.

Model Selection

Most of the models in the presented framework are hierarchically related. Nested models can be compared using a likelihood ratio test. When dealing with additional random effects, as in (6) compared to (5), mixtures of chi-square distributions can be used to tackle the boundary problems (Verbeke & Molenberghs, 2000, pp. 64–76). For non-nested models, the fit can be compared on the basis of a goodness-of-fit measure, such as Akaike's information criterion (AIC; Akaike, 1977) or the Bayesian information criterion (BIC; Schwarz, 1978). Because the models within the

proposed framework are generalized or non-linear mixed models, the significance of the parameters within a model (e.g., the δ_{ik}^{β} in (3) and (4) or the γ in (5)) can be tested using Wald tests.

Simulation and Applications

In the present section, a simulation study first will be described for the case of a linear position effect and random item ordering across test forms. Afterwards, two empirical illustrations will be given. The first deals with a test consisting of test forms with opposite item orders. The second illustration pertains to the rotated block design used in PISA 2009.

Simulation Study

Several studies already have indicated that the goodness of recovery for generalized and non-linear mixed models with standard statistical packages is satisfactory to good (Chen & Wang, 2007; Smits, De Boeck, & Verhelst, 2003; Wang & Jin, 2010; Wang & Liu, 2007). Hence, the purpose of the present simulation study is to illustrate the goodness of recovery for one particular model—namely a model with a linear position effect on item difficulty—in the case of random item ordering across respondents. Moreover, the impact on the parameter estimates when neglecting the effect of item position is illustrated.

Method

Design. Item responses were sampled according to the model in (5). Two factors were manipulated: the size of the linear position effect γ on difficulty and the number of respondents. As a first factor, γ was taken to be equal to three different values (.010, .015, and .020) which were chosen in line with the results in the empirical applications (see below). Such a position effect could be labeled as a fatigue effect. Three different sample sizes were used: small ($n = 500$); intermediate ($n = 1{,}000$); and large ($n = 5{,}000$). The combination of both factors resulted in a 3×3 design. For each cell in the design, one data set was constructed.

For each data set, 75 item difficulties were sampled from a uniform distribution ranging from -1 to 1.5. The person abilities were drawn from a standard normal distribution. Every person responded to 50 items that were drawn randomly from the pool of 75 items. This corresponds to a test administration design with individual random item order and partly overlapping items.

Model estimation. Each simulated data set was analyzed using two models: a plain Rasch model and a model with a linear position effect on item difficulty, as presented in (5). To compare the recovery of both models, the root mean square errors (RMSE) and the bias were computed for both the item and the person parameters.

Results

Table 1 presents the results of the analyses. The likelihood ratio tests indicate that, compared to the model without an item-position effect, the fit of the true model was better in all simulation conditions. For every condition, the estimates of the

Table 1
Simulation Results: Comparison between the Rasch Model and the 1PL Model with Position Effect for the Simulated Data Sets

Simulation conditions		Goodness-of-fit LRT		Estimated position effect		RMSE item difficulties		BIAS item difficulties	
Sample size	Position effect (γ)	$\chi^2(1)$[a]	p	γ	p	Rasch model	Position model	Rasch model	Position model
500	.010	115	<.0001	.011	<0001	.311	.135	.279	−.003
	.015	249	<.0001	.016	<.0001	.533	.148	.519	−.082
	.020	394	<.0001	.021	<.0001	.523	.135	.506	−.030
1000	.010	164	<.0001	.009	<.0001	.275	.096	.260	−.031
	.015	410	<.0001	.014	<.0001	.429	.111	.417	−.048
	.020	805	<.0001	.021	<.0001	.490	.105	.481	−.045
5000	.010	1076	<.0001	.011	<.0001	.263	.047	.259	−.017
	.015	2087	<.0001	.015	<.0001	.380	.041	.377	−.003
	.020	3676	<.0001	.020	<.0001	.501	.051	.506	−.006

[a]When comparing the fit of the position model with the Rasch model.

position effect γ are close to the simulated values, which indicates that the goodness of recovery of the position effect on item difficulty is good, even when sample size is small and item order is random across persons.

The results for the goodness of recovery for the item difficulty parameters show that the model with a linear effect of item position has lower RMSE and bias values in comparison to the Rasch model. The size of the RMSE and bias decreases with increasing sample size for the true model, while this is not the case for the Rasch model. The bias values for the true model are close to zero, while the bias for the Rasch model is close to the RMSE. This implies that the item difficulties are overestimated when the position effect is not taken into account. This overestimation increases with the size of the simulated position effect. In fact, the bias (and RMSE) is about equal to the average impact of the position effect ($25.5 \times \gamma$) in the Rasch model. No differences concerning the RMSE and bias of the person parameters were found between the two models in any of the conditions.

Discussion

The simulation study illustrates the satisfactory goodness of recovery for the parameters in the Rasch model with a linear effect of item position, even with limited sample sizes, randomized item orders and partly overlapping items across test forms. Moreover, it was shown that when the position effect is not taken into account, the resulting item parameters are biased.

The simulation did not show any differences in the recovery of the person parameters between the Rasch model and the true model. This rather unexpected finding presumably is due to the fact that a random item ordering was used across respondents.

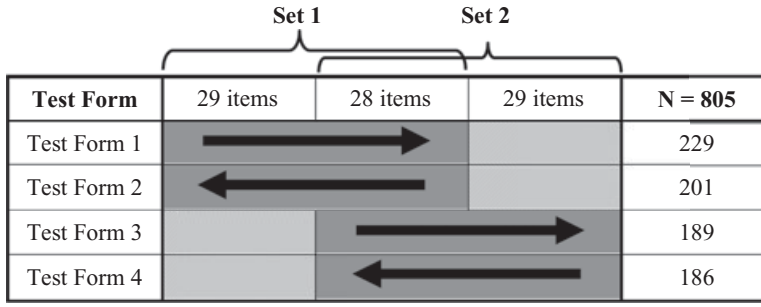

Figure 1. A graphical representation of the test administration design in Illustration I.

Illustration I: Listening Comprehension

As a first empirical example, data from a listening comprehension test in French as a foreign language were used (Janssen & Kebede, 2008). The test was designed in the context of a national assessment of educational progress in Flanders (Belgium), and it measured listening comprehension at the elementary level (the so-called "A2 level" of the Common European Framework of Reference for Languages). There were two overlapping item sets. Each item set was presented in two orders, with one order being the reverse of the other.

Method

Participants. A sample of 1039 students was drawn from the population of eighth-grade students in the Dutch-speaking region of Belgium according to a three-step stratified sampling design. Each student was randomly assigned to one of four test forms.

Materials. The computer-based test consisted of 53 audio clips pertaining to a variety of listening situations (e.g., instructions, functional messages, conversations). Each audio clip was accompanied by one to three questions, and for one clip there were five questions. Students were allowed to repeat the audio clips as many times as they wanted to. In total, 53 audio clips were accompanied by 86 items that were split into two sets of 57 items with 28 items in common. Within each item set, the audio clips were presented in two orders, one being the reverse of the other. This resulted in two alternate test forms for each item set (see Figure 1): Test Form 1 and Test Form 2 for Item Set 1, and Test Form 3 and Test Form 4 for Item Set 2.

Procedure. The computer-based test was accessed via the internet. However, due to server problems, 128 students were not able to take the test. Of the remaining 911 students, 805 students completed their test form: 229, 201, 189 and 186 students for Test Forms 1, 2, 3 and 4, respectively. The number of students dropping out before they reached the end of the test was not increasing towards the end of the test.

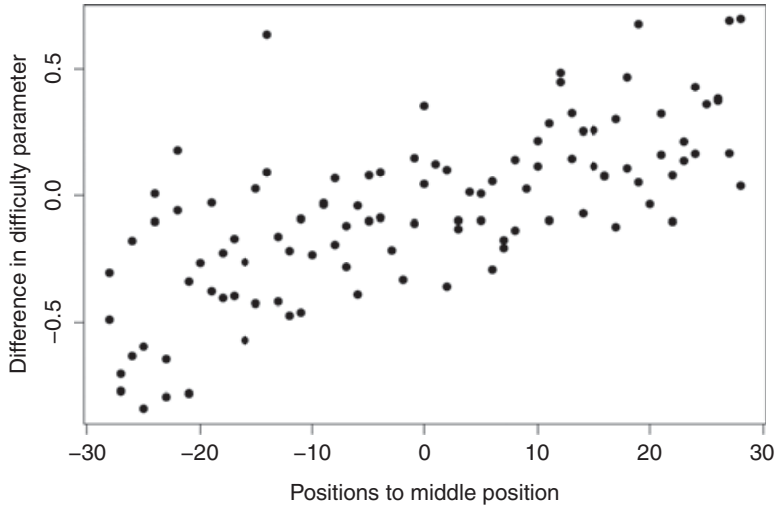

Figure 2. DIF parameters on difficulty within the whole test, according to the distance to the middle position.

Model estimation. The models were identified by constraining the mean and variance of the latent trait to 0 and 1, respectively. To model the position difference between two test forms, contrast coding was used.

Results

Descriptive statistics. No significant differences were found at the level of the total score of each test form. For both Test Form 1 and Test Form 2, the average proportion of correct responses was .76; for both Test Form 3 and Test Form 4, the average was .70. The average performance on the anchor items was identical in the four test forms with an average proportion of correct responses of .74.

Preliminary analyses. Before analyzing the position effects, we compared Rasch and 2PL models for all test forms separately. Likelihood ratio tests indicate that the 2PL model had a significantly better fit for all test forms ($\chi^2(57) = 186$, $p < .0001$, $\chi^2(57) = 159$, $p < .0001$, $\chi^2(56) = 238$, $p < .0001$, and $\chi^2(56) = 190$, $p < .0001$, for Test Forms 1 to 4, respectively). The 2PL analyses revealed that a few items had a very low discrimination parameter which resulted in unstable and extreme difficulty parameter estimates for those items. After dropping these items from further analyses, Item Sets 1 and 2 consisted of 55 and 54 items, respectively. No significant differences in mean and variance were found for students completing the different test forms. Hence, in the following analyses, all students, regardless of which booklet they were assigned, were assumed to come from the same population.

Modeling position effects on individual items. Different models were used to investigate the position effect in a combined analysis of the four test forms. The first model was a contrast-coded 2PL version of the model in (3). The goodness-of-fit measures for this model are presented in the first line of Table 3. Figure 2 shows the

Table 2
Goodness-of-Fit Statistics for the Estimated Models in Item Sets 1 and 2 Combined

Model	*N* parameters	−2logL	AIC	BIC
2PL	162	38649	38973	39733
2PL + position effect per item (DIF)	268	38164	38700	39957
2PL + linear position effect	163	38369	38695	39460
2PL + quadratic position effect	164	38369	38697	39466
2PL + cubic position effect	165	38368	38698	39699
2PL + random linear position effect	165	38307	38637	39411

differences in item difficulties between different positions according to the distance between the positions in the test forms. The plot suggests a linear trend in the effect of item position on item difficulty. The correlation between the differences in difficulty and the item positions was positive, $r = .71$, $p < .0001$.

Modeling item-position effects across items. Further, linear, quadratic and cubic trends were introduced into the measurement model, as in (5). The results of the goodness-of-fit statistics of different models are presented in Table 2. As could be expected from the plot, the model assuming only a linear position effect on difficulty provided the best fit (lowest AIC and BIC; when the model with a linear trend was compared with the 2PL model, the Likelihood Ratio Test was: $\chi^2(1) = 280$, $p < .0001$, compared with the quadratic and cubic models, the Likelihood Ratio Tests were $\chi^2(1) = 0$, $p = 1$ and $\chi^2(2) = 1$, $p = .607$, respectively). The estimated linear position parameter γ equalled .014, $t(804) = 14.81$, $p < .0001$. This indicates that an item became more difficult at later positions.

Modeling individual differences in position effects. A model with random weights for the position effect was estimated, as in (6). As can be seen in Table 2, adding the random weight to the model significantly increases the fit of the model, according to a likelihood ratio test with a mixture of χ^2 distributions ($\chi^2(1{:}2) = 62$, $p < .0001$). The estimated covariance between the position dimension and the latent trait differed significantly from zero ($t(803) = -2.54$, $p = .011$ and $\chi^2(1) = 7$, $p = .008$), which corresponds to a small negative correlation ($r = -.21$). This indicates that the position effect was smaller for students with higher listening comprehension.

Implications of the found position effect. The estimated mean for the random position effect is .013. Its estimated standard deviation was .014. Table 3 presents the effect size of the random position effect in terms of the change in the odds and the probability of a correct response of .50 for three values of γ_p, both when the item is placed one position further and when it is placed 30 positions further in the test. When γ_p is equal to the mean or one standard deviation above the mean, the position effect is positive and the success probability decreases. However, at one standard deviation below the mean the position effect γ_p is just below zero, which suggests that items become easier towards the end of the test. Although this effect is very small for k equal to 1, it accumulates to a considerable effect for k equal to 30.

Table 3
Size of the Random Linear Position Effect for Item Sets 1 and 2 Combined

Position effect		Change in ODDS (Y = 1)		P(Y = 1)[a]	
z(γ)	γ	+ 1 position	+ 30 positions	+ 1 position	+ 30 positions
−1	−.002	1.002	1.049	.500	.512
0	.013	.987	.679	.497	.405
1	.027	.973	.440	.493	.305

[a]When the item has a discrimination equal to 1 and the probability of a correct response in the reference position is .50.

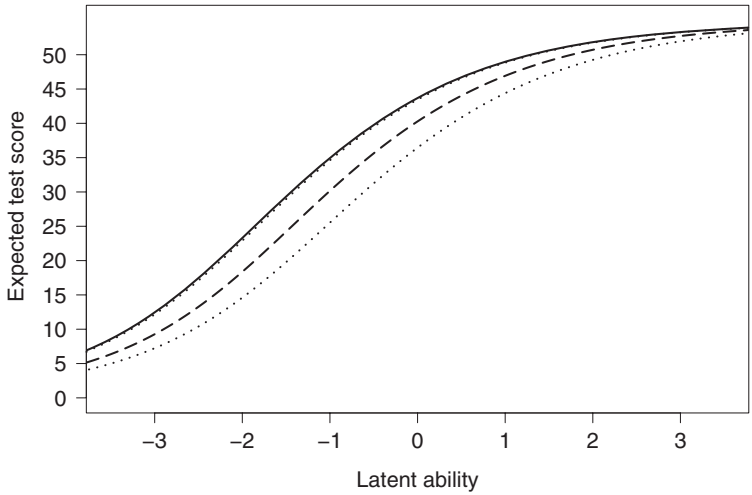

Figure 3. Test characteristic curves (TCCs) for the expected test scores for four different models, based on the parameter estimates of the listening ability data. The solid line represents the TCC of the model without a position effect. The dashed line represents the TCC of the model with an average linear position effect. The two dotted lines represent the TCC of the model with a position effect one standard deviation below the mean, and one standard deviation above the mean, respectively. (One of the dotted lines coincides with the solid line.)

Note that for about 17% of the population, the position effect was negative, so items became easier in later positions.

In order to explore the impact of the position effect on the total test score, the test characteristic curve was calculated for different cases (see Figure 3). The expected test scores under a 2PL model without a position effect are higher than the expected test scores under a 2PL model for persons with an average position effect. When the position effect is one standard deviation above the mean, the impact becomes larger. On the other hand, when the position effect is one standard deviation below the mean, the TCC is almost equal to the TCC of the model without a position effect.

Discussion

The individual differences in the found position effect indicate that not all test takers were susceptible to the effect of item position. Furthermore, although items tended to become more difficult if placed later in the test, the reverse effect was observed for a considerable proportion of test takers (for whom items became easier). The position effect therefore could be interpreted as a person-specific trait (a change parameter that indicates how a person is affected by the sequencing of items in a specific test) rather than a generalized "fatigue effect." It was shown that for some test takers the position effect seriously affects the success probability on items further along in the test. The cumulative effects of these differences in success probabilities were shown in the TCC. Both findings suggest that the position effect is not to be neglected in the present listening comprehension test, although it is not clear what the reason is for the found construct-irrelevant variance.

Illustration II: PISA 2006 Turkey

As another illustration of detecting item-position effects in low-stakes assessments, the data of one country from the PISA 2006 assessment were analyzed. PISA is a system of international assessments that focus on the reading, mathematics, and science literacy competencies of 15-year-olds (OECD, 2006). Almost 70 countries participated in 2006. In each country, students were drawn through a two-tiered stratified sampling procedure: systematic sampling of individual schools from which 35 students were randomly selected.

Method

Design. The total of 264 items in the PISA assessment (192 science, 46 math, and 26 reading items) was grouped in thirteen clusters: seven science-item clusters (S1–S7), four math-item clusters (M1–M4), and two reading-item clusters (R1, R2). A rotated block design was used for test administration (see Table 4). Each student was randomly assigned to one of thirteen test forms in which each item cluster (S1–S7, M1–M4, R1 and R2) occurred in each cluster position once. Within each cluster, there was a fixed item order. Hence, there were only differences in the position of the clusters (i.e., cluster position; ranging from position one to position four). More information on the design, the measures, and the procedure can be found in the PISA 2006 Technical Report (OECD, 2009).

Data set. The data for reading, math, and science literacy were analyzed for Turkey. The Turkish data set for PISA 2006 consisted of a representative sample of 4,942 students (2,290 girls) in 160 schools. For the current analysis we adopted the PISA scoring, where "omitted items" and "not-reached items" are scored as missing responses. Hence, these responses were not included in the analyses. Further, polytomous items were dichotomized; only a full credit was scored as a correct answer.

Model estimation. As PISA traditionally uses 1PL models to analyze the data, item discriminations were not included in the present analyses. For each literacy, four models were estimated: (a) a simple Rasch model; (b) a model assuming a main effect of cluster position as in (4), a model using dummy coding; (c) a model with

Table 4
Rotated Cluster Design Used to Form Test Booklets for the PISA 2006 Study

Test form	Cluster 1	Cluster 2	Cluster 3	Cluster 4
1	S1	S2	S4	S7
2	S2	S3	M3	R1
3	S3	S4	M4	M1
4	S4	M3	S5	M2
5	S5	S6	S7	S3
6	S6	R2	R1	S4
7	S7	R1	M2	M4
8	M1	M2	S2	S6
9	M2	S1	S3	R2
10	M3	M4	S6	S1
11	M4	S5	R2	S2
12	R1	M1	S1	S5
13	R2	S7	M1	M3

a fixed linear effect of cluster position; and (d) a model with a random linear effect of cluster position. For each model, all students were assumed to be members of the same population. The models were identified by constraining the mean of the latent trait to 0. The data were analyzed in R, using the lmer function.

Results

Modeling item-positon effects across items. The Goodness-of-fit-statistics of all four estimated models are presented in Table 5. The likelihood ratio tests indicate that the model with a dummy-coded effect of cluster position produced better fit than the Rasch model ($\chi^2(3) = 78$, $p < .0001$ for math, $\chi^2(3) = 137$, $p < .0001$ for reading, and $\chi^2(3) = 332$, $p < .0001$ for science). For all three literacies, the parameter estimates for the cluster position effect seem to increase across the four clusters (Table 6). This shows that items are more difficult when placed in later positions.

To test whether a linear trend summarizes these effects, the model with cluster position as a main effect was compared with a model with a linear cluster effect. As can be seen in Table 5, the AIC and BIC of both models are comparable, indicating comparable fit for the three literacies. The parameter estimate for the linear cluster effect is positive and significantly differs from zero for each of the three literacies (Table 6). The effect seems to be strongest for the reading items: on average, the difficulty of a reading item increases .240 when it is administered one cluster position further in the test.

Modeling individual differences in position effects. For the three literacies, the likelihood ratio test with a mixture of chi-square distributions indicates that the model with a cluster position dimension provides the best fit ($\chi^2(1{:}2) = 7$, $p = .019$ for math, $\chi^2(1{:}2) = 6$, $p = .032$ for reading, and $\chi^2(1{:}2) = 201$, $p < .0001$ for science). For example, for science the estimated covariance between the position

Table 5
Goodness-of-Fit Statistics for the Estimated Models for Math, Reading, and Science Literacy

	Math				Reading				Science			
Model	N parameters	−2logL	AIC	BIC	N parameters	−2logL	AIC	BIC	N parameters	−2logL	AIC	BIC
Simple Rasch	47	65372	65466	65895	27	41797	41851	42082	193	322687	323073	325110
+ Main effect	50	65294	65394	65851	30	41660	41720	41977	196	322355	322747	324816
+ Fixed linear effect	48	65302	65398	65837	28	41661	41717	41956	194	322358	322746	324795
+ Random linear effect	50	65295	65395	65852	30	41655	41715	41972	196	322157	322549	324618

Table 6
Estimates of the Effect of Cluster Position on Item Difficulty in the PISA 2006 Data for Turkey

	Main effect of cluster position						Fixed linear cluster effect		Random linear cluster effect			
Literacy	Cluster 2	*p*-value	Cluster 3	*p*-value	Cluster 4	*p*-value	weight	*p*-value	weight	*p*-value	*SD*	*r*
Math	.038	.2983	.189	.0002	.356	<.0001	.129	<.0001	.132	<.0001	.213	− .357
Reading	.204	.0014	.468	<.0001	.706	<.0001	.240	<.0001	.241	<.0001	.285	− .531
Science	.100	<.0001	.176	<.0001	.298	<.0001	.099	<.0001	.106	<.0001	.158	−.257

[a]The first cluster position was the reference level.

dimension and the latent trait corresponded to a small negative correlation ($r = -.257$). This suggests that, if values on the latent trait increase, the position effect decreases. For the other literacies, the found effects are similar (Table 6).

Discussion

The effects in the PISA 2006 illustration are comparable with the effects found in the first illustration. The size of the standard deviations for the position effects indicates that there are considerable individual differences in the proneness to the position effect. Again, this indicates that not all test takers were equally susceptible to the effect of item position. Similar to the findings for the listening comprehension test, the correlation between the position dimension and the latent ability was negative for all three literacies. Hence, students with a higher ability tend to have a smaller position effect.

The current analyses took into account only the items that were answered by the students. Omissions and "not reached" items were excluded from the analyses, although they were present in the original data set. In general, non-response is taken as an indicator of low test motivation (e.g., Wise & DeMars, 2005). Consequently, our findings of the general decrease in performance towards the end of the test for those students who still responded to the items also may refer to a decrease in test motivation and to individual differences in the amount of effort they expended on earlier versus later items in the test.

General Discussion

The purpose of the present article was to propose a general framework for detecting and modeling item-position effects of various types using explanatory and descriptive IRT models (De Boeck & Wilson, 2004). The framework was shown to overcome the limitations of current approaches for modeling item-order effects, which either are focused on effects at the test score level or which make use of a two-step estimation procedure. The practical relevance of the proposed models was illustrated with a simulation study and two empirical applications. The simulation study showed that the framework is applicable even with random item orders across examinees. The empirical studies illustrated that item-position effects may be present in large-scale, low-stakes assessments.

Further Model Extensions

The current framework only considers item-position effects for dichotomous item responses. It also would be interesting to model item-order effects in polytomous IRT models. Moreover, the effects of item position may appear not only in response accuracy but they may even have a stronger impact in the time taken to respond to an item (Wise & Kong, 2005). Hence, an extension to models taking response accuracy and response time jointly into account (van der Linden, Entink, & Fox, 2010) seems to be an important step in further understanding these effects.

Limitations

The present framework investigates the effect of item position in explaining lack of item parameter invariance across different test forms. Of course, item position is only one type of context effect that may be responsible for the lack of item parameter invariance. The present model also does not look at effects caused by one item being preceded by another item (e.g., the effect of a difficult item preceding an easy item). Such sequencing effects are a function of item position as well, but these effects refer to the position of subsets of items (e.g., pairs of items), whereas the present framework focuses only on the position of single items within test forms.

The proposed models are limited to position effects that occur independently of the person's response to an item. However, in the case of a practice effect, one can assume that solving an item generally may produce a larger practice effect than trying an item unsuccessfully. Specific IRT models exist that model such response-contingent effects of item position. Examples of these so-called dynamic IRT models are Verguts and De Boeck (2000) and Verhelst and Glas (1993).

As was already explained in the introduction, the present framework focuses on detecting and modeling item-position effects but is not apt for giving explanations for the effects found. Like in DIF research (Zumbo, 2007), building frameworks for empirically investigating item-position effects probably precedes a next generation of research answering "the why question" of the found effects. Further person explanatory models (De Boeck & Wilson, 2004), which try to capture the individual differences in the position effect, could be helpful in finding an explanation. For example, it has been shown that in low-stakes assessments test takers may differ in test motivation, and hence it may be interesting to include self-report measures of test motivation (e.g., Wise & DeMars, 2005) or response time (Wise & Kong, 2005) as an additional person predictor in the IRT model.

As a final limitation, the present framework does not allow for detection of item-position effects in a single test administration, except when the test items belong to an item bank with known item properties. In that case, the effect of a change in item position can be compared to the reference position of the item in the item bank. If an item-position effect is expected within a single test design, it seems advisable to randomly order harder and easier items to avoid bias. Surely, if items are ordered from hard to easy, a positive linear position effect on difficulty would disadvantage lower ability persons and benefit higher ability persons (e.g., Meyers et al., 2009).

Acknowledgments

The present study was supported by several grants from the Flemish Ministry of Education. For the data analysis we used the infrastructure of the VSC—Flemish Supercomputer Center, funded by the Hercules foundation and the Flemish Government—Department EWI.

References

Akaike, H. (1977). On entropy maximization principle. In P. R. Krishnaiah (Ed.), *Applications of statistics* (pp. 27–41). Amsterdam, The Netherlands: North-Holland.

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed effects models using S4 classes*. ⟨http://cran.r-project.org/web/packages/lme4⟩.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Reading, MA: Addison Wesley.

Chen, C., & Wang, W. (2007). Effects of ignoring item interaction on item parameter estimation and detection of interacting items. *Applied Psychological Measurement*, *31*, 388–411.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1–28.

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

Dorans, N. J., & Lawrence, I. M. (1990). Checking the statistical equivalence of nearly identical test editions. *Applied Measurement in Education*, *3*, 245–254.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *65*, 495–515.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374.

Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 131–155). New York, NY: Springer.

Glas, C. A. W., & Pimentel, J. L. (2008). Modeling nonignorable missing data in speeded tests. *Educational and Psychological Measurement*, *68*, 907–922.

Goegebeur, Y., De Boeck, P., & Molenberghs, G. (2010). Person fit for test speededness: Normal curvatures, likelihood ratio tests and empirical Bayes estimates. *Methodology–European Journal of Research Methods for the Behavioral and Social Sciences*, *6*, 3–16.

Hamilton, J. C., & Shuminsky, T. R. (1990). Self-awareness mediates the relationship between serial position and item reliability. *Journal of Personality and Social Psychology*, *59*, 1301–1307.

Hanson, B. A. (1996). Testing for differences in test score distributions using loglinear models. *Applied Measurement in Education*, *9*, 305–321.

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.

Holman, R., & Glas, C. A. W. (2005). Modelling non-ignorable missing-data mechanisms with item response theory models. *British Journal of Mathematical and Statistical Psychology*, *58*, 1–17.

Janssen, R., & Kebede, M. (2008, April). *Modeling item-order effects within a DIF framework*. Paper presented at the meeting of the National Council on Measurement in Education, New York, NY.

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.

Kubinger, K. D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly*, *50*, 311–327.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement*, *69*, 232–244.

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.

Meulders, M., & Xie, Y. (2004). Person by item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 213–240). New York, NY: Springer.

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.

Mollenkopf, W. G. (1950). An experimental study of the effects on item analysis data of changing item placement and test-time limit. *Psychometrika*, *15*, 291–315.

Moses, I., Yang, W., & Wilson, C. (2007). Using kernel equating to assess item order effects on test scores. *Journal of Educational Measurement*, *44*, 157–178.

Organization for Economic Co-operation and Development (OECD). (2006*). Assessing scientific, reading and mathematical literacy: A framework for PISA 2006*. Paris, France: OECD.

Organization for Economic Co-operation and Development (OECD). (2009). *PISA 2006. Technical Report*. Paris, France: OECD.

Pommerich, M., & Harris, D. J. (2003, April). *Context effects in pretesting: Impact on item statistics and examinee scores*. Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

R Development Core Team. (2011). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ⟨http://www.r-project.org/⟩

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Rijmen, F., & De Boeck, P. (2002). The random weights linear logistic test model. *Applied Psychological Measurement*, *26*, 271–285.

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205.

SAS Institute Inc. (2008). *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

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.

Smits, D. J. M., De Boeck, P., & Verhelst, N. (2003). Estimation of the MIRID: A program and a SAS-based approach. *Behavior Research Methods, Instruments, & Computers*, *35*, 537–549.

Steinberg, L. (1994). Context and serial-order effects in personality measurement: Limits on the generality of measuring changes the measure. *Journal of Personality and Social Psychology*, *66*, 341–349.

Stout, W. (2002). Psychometrics, from practice to theory and back. *Psychometrika*, *67*, 485–518.

van der Linden, W. J., Entink, R. H. K., & Fox, J. P. (2010). IRT parameter estimation with response times as collateral information. *Applied Psychological Measurement*, *34*, 327–347.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.

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.

Wang, W., & Jin, K. (2010). A generalized model with internal restrictions on item difficulty for polytomous items. *Education and Psychological Measurement*, *70*, 181–198.

Wang, W., & Liu, C. (2007). Formulation and application of the generalized multilevel facets model. *Educational and Psychological Measurement*, *67*, 583–605.

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

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

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

Yen, W. M. (1980). The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement*, *17*, 297–311.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.

Authors

DRIES DEBEER is Researcher at the Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, 3000 Leuven (PB 3713), Belgium; dries.debeer@ppw.kuleuven.be. His current research interests include psychometric methods, item response models, and educational measurement.

RIANNE JANSSEN is Associate Professor at the Faculty of Psychology and Educational Sciences, KU Leuven, Dekenstraat 2 (PB 3773), 3000 Leuven, Belgium; rianne.janssen@ppw.kuleuven.be. Her current research interests include psychometrics and educational measurement.