Psychological Test and Assessment Modeling, Volume 60, 2018 (2), 165-187

Item position effects in a reading comprehension test: An IRT study of individual differences and individual correlates

Gabriel Nagy[1], Benjamin Nagengast[2], Michael Becker[3], Norman Rose[2] & Andreas Frey[4]

Abstract

Item position (IP) effects typically indicate that items become more difficult towards the end of a test. Such effects are thought to reflect the persistence with which test takers invest effort and work precisely on the test. As such, IP effects may be related to cognitive and motivational variables that are relevant for maintaining a high level of effort and precision. In this article, we analyzed IP effects in a reading comprehension test. We propose an IRT model that includes random IP effects affecting item difficulties and fixed IP effects affecting item discriminations. We found evidence for gradually increasing item difficulties and decreasing discriminations. Variation in IP effects on the item difficulties was systematically related to students' decoding speed and reading enjoyment. The results demonstrate that the relationship between the overall scores and other variables is affected by respondents' test-taking behavior, which is reflected in the random IP effect.

Keywords: reading comprehension, item position effects, correlates of item position effects, item response theory

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

[2] University of Tübingen, Germany

[3] German Institute for International Educational Research, Frankfurt am Main, Germany

[4] Friedrich-Schiller-University Jena, Germany

Scores derived from achievement tests are commonly interpreted to indicate individuals' maximal performance (e.g., Goff & Ackerman, 1992); thereby, it is assumed that test takers maintain their effort throughout a test. However, most often, this is not the case, especially in low-stakes situations (Asseburg & Frey, 2013). Typically, as they get closer to the end of a test, the probability of individuals solving the test items declines (Leary & Dorans, 1985). These effects are subsumed under the term *item position effects* (IP) in the literature. IP effects play a role in virtually all testing situations, recurring on tests of moderate to extensive lengths (Leary & Dorans, 1985). The phenomenon of items becoming more difficult towards the end of the test is commonly referred to as a fatigue effect (Kingston & Dorans, 1984), reflecting declines in test takers' motivation to apply, and their capacity to maintain, a constant level of effort over the course of a test (Ackerman & Kanfer, 2009). Therefore, IP effects can be expected to vary across individuals (Debeer & Janssen, 2013) and to be related to individuals' cognitive and motivational resources that are relevant for maintaining a constantly high level of effort and precision.

This article focuses on IP effects in a reading comprehension test administered to fifth-grade students. Three research questions were addressed. First, we examined the test for the existence of individually varying IP effects. For this purpose, we proposed and tested different versions of the two parameter logistic (2PL) item response theory (IRT) model, including individual differences in IP effects. Second, we extended IRT models with random IP effects to include variables hypothesized to be related to students' motivational (reading enjoyment) and cognitive (decoding speed) resources that are relevant for maintaining a constant level of effortful and precise processing. Finally, we investigated whether disregarding IP effects influences the estimated relationship between test scores and covariates.

Item position effects

Mollenkopf (1950) noted that changing the positions of items in a test affects item characteristics. IP effects of the first kind (IP1) make items appear to be harder or easier when presented in later positions. Alternatively, IP1 effects make individuals appear to be more or less able. Leary and Dorans (1985) referred to IP1 effects that lead to decreasing item difficulties, so that individuals appear to be more able, as *practice effects*, whereas IP1 effects in the opposite direction were referred to as *fatigue effects*. Practice and fatigue effects as defined by Leary and Dorans (1985) refer only to the direction of IP1 effects. The authors did not theorize about the psychological processes underlying IP1 effects; thus, the terms should be understood as loose circumscriptions rather than well-defined constructs.

IP effects of the second kind (IP2) lead to changes in an item's potential to discriminate between different levels of ability. They make items appear to be more or less reliable when located toward the end of a test. Such effects have been intensively investigated in the areas of personality and attitude assessment (e.g., Hartig, Hölzel, & Moosbrugger, 2007), but have rarely been examined in the area of cognitive testing.

Item position effects affecting item difficulties (IP1 effects)

Many studies have provided evidence for negative IP1 effects in achievement tests (Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Hohensinn, Kubinger, Reif, Holocher-Ertl, Khorramdel, & Frebort, 2008; Kingston & Dorans, 1984; Le, 2007; Meyers, Miller & Way, 2009; Schweizer, Schreiner, & Gold, 2009). Most studies treated IP1 effects as fixed effects (FIP1 effects) that do not vary across individuals. A more realistic point of view is that IP1 effects reflect individuals' reactions to a test-taking situation, which means that they should be conceptualized as random effects (i.e., RIP1 effects) that vary across individuals.

Debeer and Janssen (2013) provided evidence for RIP1 effects in various achievement domains tested in the PISA 2006 study and in a test measuring listening comprehension. Similar findings were reported by Robitzsch (2009), using mathematics achievement tests assessed in lower secondary schools, and by Hartig and Buchholz (2012), who analyzed RIP1 effects in the PISA 2006 science achievement test. In addition, in the PISA 2009 reading test, Debeer, Buchholz, Hartig, and Jansen (2014) found RIP1 effects that also varied between schools, but between-school differences in RIP1 effects were negligible in other assessments (Weirich, Hecht, Penk, Roppelt, & Böhme, 2017). In all studies, the sign of the mean RIP1 effect was negative. Some studies investigated RIP1 effects in aptitude tests. Schweizer, Schreiner, and Gold (2009; see also Ren, Goldhammer, Moosbrugger, & Schweizer, 2012) found evidence for an RIP1 effect in the Advanced Progressive Matrices test, and Schweizer, Troche, and Rammsayser (2011) reported an RIP1 effect in a numerical reasoning test.

The correlates of RIP1 effects have been investigated in several recent studies. RIP1 effects in an aptitude test were found to be positively related to general intelligence (Schweizer et al., 2011), and to executive attention (Ren, Goldhammer, Moosbrugger, & Schweizer, 2012). In addition, the RIP1 effect in a science test was found to be correlated with decreases in self-reported test-taking effort (Weirich et al., 2017), so that students with steeper declines in test-taking effort had a more negative IP1 effect. Similarly, Qian (2014) found motivation to be an important predictor of the IP1 effect in the writing task included in the 2007 National Assessment of Educational Progress, although IP1 effects were also impacted by institutional characteristics. Profound school-type differences in the size of the IP1 effect were found in the German PISA 2012 assessment, both for the PISA tests (Nagy, Lüdtke, & Köller, 2016; Nagy, Lüdtke, Köller, & Heine, 2017) and for the tests of the German educational standards (Nagy, Haag, Lüdtke, & Köller, 2017). Most recently, Lindner, Nagy, Ramos, and Retelsdorf (2017) showed that experimentally depleting students' self-control resources resulted in a stronger IP1 effect in a mathematics test.

These results suggest that the correlations of the ability variables with the covariates are affected by RIP1 effects. When RIP1 effects are not separated from the ability variables, the latter are confounded with RIP1 effects. The potential impact that RIP1 effects can have on the conclusions drawn has been exemplified in the longitudinal extension to the German PISA 2012 assessment. Nagy et al. (2016) found that ignoring IP1 effects and changes therein resulted in negative estimates of proficiency gains for reading and science in nonacademic tracks, and that this effect vanished once IP1 effects were accounted for (see also Nagy, Lüdtke et al., 2017). Similarly, ignoring the IP1 effect resulted in differences in reading gains in favor of girls, but these differences disappeared once the IP1 effect was controlled for (Nagy, Retelsdorf, Goldhammer, Schiepe-Tiska, & Lüdtke, 2017).

Item position effects affecting item discriminations (IP2 effects)

IP2 effects on item discriminations have rarely been considered in the area of cognitive testing. Some studies indicate that item discriminations in cognitive tests tend to increase when presented in later positions (e.g., Le, 2007; Mollenkopf, 1950), whereas other findings suggest that the pattern depends on the achievement domain (Kingston & Dorans, 1982). Note, however, that all findings on IP2 effects have been derived on the basis of unidimensional measurement models. Situations in which item responses are affected by RIP1 effects call for multidimensional measurement models (Debeer & Janssen, 2013). The failure to separate the RIP1 effect from the ability variable means that the item discriminations reflect the items' connection with a composite of ability and RIP1 effects rather than the items' connection with the "purified" ability variable. In this article, we propose a multidimensional IRT (MIRT) model in which the RIP1 effect is separated from the ability variable and the IP2 effect is defined solely with respect to the ability variable.

IRT models assessing item position effects

In this section, we describe IRT models that assess IP effects. Their application requires that individuals work on items presented in different positions. This requirement is fulfilled in large-scale assessments that build upon matrix designs in which at least some items are presented in different positions (e.g., Frey, Hartig, & Rupp, 2009).

Formulation of IRT models

All models envisaged are extensions of the traditional 2PL model, which is given as

$$logit\left[P\left(y_{ijp}=1\right)\right]=\alpha_j\left(\theta_i-\beta_j\right), \qquad (1)$$

where y_{ijp} stands for the individual's $i=1,2,\dots,N$ response to item $j=1,2,\dots,J$, presented in position $p=0,1,\dots,P$, α_j and β_j stand for the item discrimination and item difficulty, respectively, and θ_i reflects the value of the ability variable for individual i. Note that we indexed the item position starting from $p=0$ instead of from $p=1$.

The 2PL model is extended to include FIP1, RIP1, and IP2 effects. Some models, which will subsequently be presented, have already been introduced in previous studies. We present these models in a slightly different notation and parameterization and extend them to accommodate a more fine-grained assessment of IP effects, including RIP1 and IP2 effects.

Hohensinn and colleagues (2008) reformulated the linear logistic test model (Fischer, 1973) to assess FIP1 effects. A 2PL version of their model can be written as

$$logit[P(y_{ijp} = 1)] = \alpha_j(\theta_{i0} + \lambda p\delta - \beta_{j0}). \qquad (2)$$

Here, the FIP1 effect is represented by δ, which affects item responses according to a linear function of their position p. As a consequence, β_{j0} now refers to the difficulty of item j presented in the reference position (i.e., $p = 0$), and θ_{i0} stands for the ability of individual i, defined with respect to $p = 0$. λ is a parameter to be fixed by the researcher. It serves the purpose of putting the FIP1 effect δ onto an interpretable scale. In Equation 2, δ reflects the change in all individuals' θ values per unit change in λp, that is, we could define the value of the ability variable assessed in different reference positions as: $\theta_{ip} = \theta_{i0} + \lambda p\delta$ (e.g., Robitzsch, 2009). When λ is fixed to 1, δ reflects the change in θ when an item is moved one position towards the end of the test. By setting λ to $\lambda = 1/m$, δ captures the change in θ when an item is moved *m* positions towards the end of a test. Negative values of δ indicate that individuals appear to be less capable when θ is defined with respect to a later position.

One limitation of the model presented in Equation 2 is the assumption that δ has the same value for each person. This assumption can be relaxed such that

$$logit[P(y_{ijp} = 1)] = \alpha_j(\theta_{i0} + \lambda p\delta_i - \beta_{j0}). \qquad (3)$$

The difference between Equations 2 and 3 is that the subscript *i* is attached to δ, which means that each person might have a different value of δ. This results in an RIP1 effect. Because δ is assumed to differ between individuals, the ability variables defined in different reference positions ($\theta_{ip} = \theta_{i0} + \lambda p\delta_i$) might now exhibit different rank orderings of individuals. The model in Equation 3 is equivalent to the model proposed by Debeer and Jansen (2013). Besides the mean (κ_0) and the variance of θ_0 (ϕ_{00}), the model estimates the mean (κ_1) and variance of δ (ϕ_{11}), and the covariance between θ_0 and δ (ϕ_{01}).

The model in Equation 3 assumes that θ_0 has a constant impact on an item regardless of its position. This assumption might be questioned because items become increasingly influenced by δ. An alternative is to assume that the impact of θ_0 changes across positions:

$$logit[P(y_{ijp} = 1)] = \alpha_{j0}[\gamma^{\lambda p}(\theta_{i0} - \beta_{j0}) + \lambda p\delta_i], \qquad (4)$$

where γ ($\gamma > 0$) is a parameter introduced to capture the change of the impact of the ability variable θ_0 on item responses observed in later positions.

The value of γ stands for the change in an item's discrimination for measuring θ_0 when presented in position $p = m$ relative to the initial position ($p = 0$), and α_{j0} now stands for the item's j discrimination for measuring θ_0 when presented in the initial position $p = 0$. Hence, γ can be interpreted as an indicator of the IP2 effect. A value of $\gamma = 1$ indicates that the impact of θ_0 on an item is not altered by the item's position (i.e., absence of IP2 effects), values smaller than 1 indicate decreasing item discriminations, whereas values greater than 1 indicate increasing item discriminations the later the items are presented in a test.

The model given in Equation 4 can be rewritten as

$$logit[P(y_{ijp} = 1)] = \alpha_{j0}\gamma^{\lambda p}[\theta_{i0} + \gamma^{-\lambda p}\lambda p\delta_i - \beta_{j0}], \quad (5)$$

showing that the position-specific ability variable θ_p is given by $\theta_{ip} = \theta_{i0} + \gamma^{-\lambda p}\lambda p\delta_i$. The RIP1 effect is assumed to contribute in a nonlinear fashion to the value of the position-specific ability variable. The impact is positively accelerated whenever $\gamma < 1$, it is negatively accelerated when $\gamma > 1$, and it is linear when $\gamma = 1$.

Examination of the consequences of item position effects

In the IRT models given in Equations 2 to 4, θ_0 is defined as the ability underlying the responses to items administered in the first position, which is, by definition, not affected by IP effects. Therefore, these IRT models provide estimates of the means and variances of ability distributions that are fully adjusted for IP effects. When the models are extended by covariates, they also provide estimates of the covariances of θ_0 and δ with a covariate k, ϕ_{0k} and ϕ_{1k}, which can be interpreted as the covariance that is fully adjusted for IP effects, and the covariance with the RIP1 effect, respectively. In addition, the IRT models make it possible to study changes in the means (κ_p), variances (ϕ_{pp}), and covariances (ϕ_{pk}) of the ability variables θ_p, defined with respect to different reference positions. If the loading of the IP1 effect is denoted by ω_p (i.e., $\omega_p = \lambda p$ in the case of Equations 2 and 3, and $\omega_p = \gamma^{-\lambda p}\lambda p$ in the case of Equations 4 and 5), the mean of θ_p is given by

$$\kappa_p = \kappa_0 + \omega_p\kappa_1. \quad (6)$$

The variance of θ_p is given by

$$\phi_{pp} = \phi_{00} + \omega_p^2\phi_{11} + 2\omega_p\phi_{01}. \quad (7)$$

The covariances of covariate k with the position-specific ability variables adhere to

$$\phi_{pk} = \phi_{0k} + \omega_p \phi_{1k}. \tag{8}$$

The covariances can be standardized to derive the correlations ρ_{pk} ($\rho_{pk} = \frac{\phi_{pk}}{\sqrt{\phi_{kk}\phi_{pp}}}$).

Issues of model estimation

All models presented can be estimated in a conventional IRT framework by means of marginal maximum likelihood techniques, employing the expectation maximization algorithm. The models that include RIP1 effects (Equations 3 and 4) call for a multidimensional specification, in which the IP effect δ is represented by a distinct dimension. We propose to treat each item × position combination as a separate item (e.g., Nagy et al., 2016). The creation of "virtual" items does not affect the data likelihood as long as the necessary parameter constraints are imposed. In this research, the program M*plus* 7.4 (Muthén & Muthén, 2012) was used. The models can be estimated by any other software suitable for MIRT analyses that enables nonlinear parameter constraints. A description of the model setup is given in the appendix to this article.

The present study

Our study focused on IP effects in a reading comprehension test administered to fifth-grade students. The analyses attempted to shed light on their size and meaning, as well as the consequences of IP effects for assessing students' reading comprehension abilities. The data came from a study that used large student samples, which are representative of the lower secondary school types in two federal states in Germany.

Three research questions of different scopes were examined. The first question addressed the nature of IP effects. Here, we explored (a) whether the reading comprehension test was impacted by a negative IP1 effect, (b) whether the IP1 effect could be conceptualized as varying across individuals (RIP1 effect), and (c) whether IP2 effects occurred. The second research question focused on the correlates of the RIP1 effect. We considered two variables: *decoding speed* and *reading enjoyment*. We chose these variables because they are prototypical exemplars of the cognitive capacities and motivational resources that are related to reading comprehension (Artelt, Schiefele, & Schneider, 2001; Kintsch, 1998). In addition, based on existing theories, expectations about the variables' connection to the RIP1 effect can be deduced, although almost all investigations assume that the relationship of reading comprehension with decoding speed and reading enjoyment reflects ability relations. The last research question addressed the consequences of ignoring IP effects when examining the correlates of students' reading comprehension. Whenever RIP1 effects are not separated from the ability variable, the overall test score confounds two sources of systematic variance. Consequently, the external correlations of the overall test score confound the covariates' relationships with the ability variable and the RIP1 effect (e.g., Ren et al., 2012).

Item position effects in tests of reading comprehension

We expected the reading items to be affected by a negative IP1 effect. Furthermore, we expected IP1 effects to vary across individuals (RIP1 effects). This pattern of results would be in line with findings from related research projects (Debeer & Janssen, 2013; Debeer et al., 2014; Hartig & Buchholz, 2012; Robitzsch, 2009; Weirich et al., 2017). Given that IP2 effects were not systematically investigated in conjunction with RIP1 effects, we were not able to formulate specific expectations about their occurrence. On a general level, we expected item discriminations to decrease rather than increase across positions. Such a pattern would indicate that a second process (i.e., an RIP1 effect) gradually takes over. One consequence of such a phenomenon would be that items presented towards the end of a test provide less information about measures of reading comprehension ability.

Correlates of item position effects in tests of reading comprehension

We expected both variables, decoding speed and reading enjoyment, to be related to reading comprehension as well as to the RIP1 effect. Individuals characterized by a high level of decoding speed read fluently and at a faster pace. A high level of decoding speed frees up resources for higher-level processing (e.g., Artelt, Schiefele, & Schneider, 2001; Kintsch, 1998). As reading is a less exhausting activity for good decoders, they are also more likely to show a higher level of persistence in a test-taking situation.

The impact of motivational variables on reading comprehension is commonly assumed to operate on a long-term basis. Individuals who find reading to be an enjoyable activity are assumed to read more frequently, thereby raising their ability level (Retelsdorf, Köller, & Möller, 2011). One might speculate that individuals with a higher interest in reading might also be more motivated to sustain their effort while working on a test, as indicated in the results of Weirich et al. (2017) and Qian (2014).

Method

Sample

This study drew on data from the Tradition and Innovation in School Systems Study (TRAIN) that included N = 2,830 fifth-grade students from 86 secondary schools in two German federal states (Saxony and Baden-Württemberg). For the analyses, we selected only those students who completed the reading comprehension test (N = 2,774; 46.4%

females; 27.2% students with a migration background; mean age *M* = 11.10, *SD* = 0.56). Of the students in our study, 36.6% were in the combined track in Saxony, whereas 40.5% were in the lower track in Baden-Württemberg and 22.9% were in the intermediate track in Baden-Württemberg. Students attending the highest track were not assessed in the TRAIN study.

Instruments

Reading comprehension. The test is composed of material taken from different well-established German school achievement studies (Granzer, Köller, & Bremerich-Vos, 2009; Lehmann, Peek, & Poerschke, 1997; Nauck & Otte, 1980). Test material was selected on the basis of expert ratings that confirmed its appropriateness for fourth- and fifth-grade students. Students were given short reading passages accompanied by a number of questions. In-depth examinations of the test provided no evidence for passage effects. The test contained 51 items, which were administered in a booklet design (Frey, Hartig, & Rupp, 2009). The test was made up of eight item clusters, which were combined into nine different booklets (Table 1). Item clusters were composed so that they represented reading material suitable for fourth-grade (Clusters A4, B4, and C4) and fifth-grade students (Clusters A5, B5, C5, and F5). Cluster S5 appeared at the end of each booklet and was composed of fifth-grade material.

The test contained dichotomous and partial credit items. In order to avoid complications that would require the proposed models to be extended to cover partial credit items, we considered only the dichotomous items. In addition, we deleted items with nonsignificant discrimination parameters. Items shaded in gray in Table 1 were deleted (8 items in total). Relative to the first position, most clusters showed a maximum change in positions ranging from 19 to 22 (Clusters A4, B4, C4, A5, B5, C5, F5). The largest cluster, F5, took a variety of positions, reflecting changes ranging from five to 11 positions. Only Cluster S5 changed its position by a maximum of two positions. The items that respondents did not reach were coded as missing responses. The overall reliability of the test was good. Assuming a 2PL model resulted in a marginal reliability index of *Rel.* = .83.

Decoding speed. This variable was assessed with the Salzburger Lesescreening test (Auer, Gruber, Mayringer, & Wimmer, 2008). The test consists of verbal statements, presented in a limited time condition. Students are asked to judge the correctness of each statement. The overall score is the number of items to which an individual responds. The internal consistency of the test was good (Kuder-Richardson Formula 20; KR-20 = .96).

Reading enjoyment. This short scale consisted of three items assessed by means of a 4-point Likert scale. Items were taken from the Habitual Reading Motivation Questionnaire (Möller & Bonerad, 2007). The scale had a consistency value of Cronbach's α = .94. Item responses were analyzed by means of the graded response IRT model.

Table 1:
Booklet Design of the Reading Comprehension Test. Cells Shaded in Gray Stand for Items Excluded from the Analyses.

Pos.	B1	B2	B3	B4	B5	B6	B7	B8	B9
0	A4-01	B5-01	F5-01	B4-01	C5-01	F5-01	C4-01	A5-01	F5-01
1	A4-02	B5-02	F5-02	B4-02	C5-02	F5-02	C4-02	A5-02	F5-02
2	A4-03	B5-03	F5-03	B4-03	C5-03	F5-03	C4-03	A5-03	F5-03
3	A4-04	B5-04	F5-04	B4-04	C5-04	F5-04	C4-04	A5-04	F5-04
4	A5-01	B5-05	F5-05	B5-01	C5-05	F5-05	C5-01	A5-05	F5-05
5	A5-02	B5-06	F5-06	B5-02	C5-06	F5-06	C5-02	F5-01	F5-06
6	A5-03	F5-01	F5-07	B5-03	C5-07	F5-07	C5-03	F5-02	F5-07
7	A5-04	F5-02	F5-08	B5-04	F5-01	F5-08	C5-04	F5-03	F5-08
8	A5-05	F5-03	F5-09	B5-05	F5-02	F5-09	C5-05	F5-04	F5-09
9	F5-01	F5-04	F5-10	B5-06	F5-03	F5-10	C5-06	F5-05	F5-10
10	F5-02	F5-05	F5-11	F5-01	F5-04	F5-11	C5-07	F5-06	F5-11
11	F5-03	F5-06	F5-12	F5-02	F5-05	F5-12	F5-01	F5-07	F5-12
12	F5-04	F5-07	F5-13	F5-03	F5-06	F5-13	F5-02	F5-08	F5-13
13	F5-05	F5-08	F5-14	F5-04	F5-07	F5-14	F5-03	F5-09	F5-14
14	F5-06	F5-09	F5-15	F5-05	F5-08	F5-15	F5-04	F5-10	F5-15
15	F5-07	F5-10	B4-01	F5-06	F5-09	C4-01	F5-05	F5-11	A4-01
16	F5-08	F5-11	B4-02	F5-07	F5-10	C4-02	F5-06	F5-12	A4-02
17	F5-09	F5-12	B4-03	F5-08	F5-11	C4-03	F5-07	F5-13	A4-03
18	F5-10	F5-13	B4-04	F5-09	F5-12	C4-04	F5-08	F5-14	A4-04
19	F5-11	F5-14	C5-01	F5-10	F5-13	A5-01	F5-09	F5-15	B5-01
20	F5-12	F5-15	C5-02	F5-11	F5-14	A5-02	F5-10	B4-01	B5-02
21	F5-13	C4-01	C5-03	F5-12	F5-15	A5-03	F5-11	B4-02	B5-03
22	F5-14	C4-02	C5-04	F5-13	A4-01	A5-04	F5-12	B4-03	B5-04
23	F5-15	C4-03	C5-05	F5-14	A4-02	A5-05	F5-13	B4-04	B5-05
24	S-01	C4-04	C5-06	F5-15	A4-03	S-01	F5-14	S-01	B5-06
25	S-02	S-01	C5-07	S-01	A4-04	S-02	F5-15	S-02	S-01
26	S-03	S-02	S-01	S-02	S-01	S-03	S-01	S-03	S-02
27	S-04	S-03	S-02	S-03	S-02	S-04	S-02	S-04	S-03
28	S-05	S-04	S-03	S-04	S-03	S-05	S-03	S-05	S-04
29	S-06	S-05	S-04	S-05	S-04	S-06	S-04	S-06	S-05
30		S-06	S-05	S-06	S-05		S-05		S-06
31			S-06		S-06		S-06		

Statistical procedures for estimating item position effects and their correlates

We applied a hierarchy of increasingly complex models to the reading data. Model-data fit was judged by the BIC and AIC indices, which penalize highly parameterized models. Models with small AIC and BIC indices are preferable. All IRT models were identified by fixing the mean and the variance of the ability variable to 0 and 1. In order to put the IP effects on an interpretable metric, we fixed λ to $\lambda = 1/20$, so that δ stands for the expected change in the logits of item responses, and γ for the proportional change in discrimination parameters when items are moved by 20 positions towards the end of the test.

The best-fitting IRT model was extended by the inclusion of the covariates. Their relationships to the ability variable and the RIP1 effect were assessed by latent correlations. Decoding speed was measured by a single continuous variable using a single-indicator measurement model, whereas reading enjoyment was modeled by the graded response model. For both constructs, latent variables were specified to have zero means and unit variances. The estimates provided by the IRT models that included the covariates were used to study the covariates' relationships with θ_0 and δ, as well as their correlations with θ_p (Equation 8). All models were estimated with the M*plus* 7.4 software (Muthén & Muthén, 2012) using marginal maximum likelihood techniques, employing the expectation maximization algorithm utilizing standard integration with 15 integration points per dimension.

Results

IRT analyses of item position effects

The fit statistics of the IRT models are summarized in Table 2. Beginning with the 2PL model, all subsequent models included components, which represented different aspects of the IP effect. Each component increased the model-data fit, as reflected by the AIC and BIC indices. Adding the FIP1 and RIP1 effects to the models provided the largest improvement in fit statistics. However, both fit indices indicated the presence of IP2 effects. Hence, we chose the most complex model as the final model.

Table 2:
Model-Data Fit of Alternative IRT Models

	# Par.	-lnL	AIC	BIC
2PL	86	43810.4	87792.9	88302.7
2PL + FIP1	87	43741.4	87656.9	88172.6
2PL + RIP1	89	43678.4	87534.9	88062.5
2PL + RIP1 + IP2	90	43670.4	87520.9	88054.4

Note. –lnL = negative model log-likelihood; 2PL = 2-parameter logistic model; FIP1 = fixed item position effects on item difficulties; RIP1 = random item position effects on item difficulties; IP2 = fixed item position effects on item discriminations.

In this model, the average RIP1 effect was estimated to be $\hat{\kappa}_1$ = -0.27 (*SE* = 0.04; *p* < .001). Moving the reference position for assessing students' reading ability by 20 positions toward the end of the test was expected to lead to an average decrease in ability estimates of -0.27 points. As we *z*-standardized the variance of the ability variable in the first position, this effect is given on a standardized metric. The RIP1 effect was found to exhibit a significant degree of variability, $\hat{\phi}_{11}$ = 0.26 (*SE* = 0.04; p < .001), and it was not correlated with the ability variable, $\hat{\rho}_{01}$ = -.03 (*SE* = 0.09; p = .766). Finally, the IP2 effect resulted in lower discrimination parameters in later positions, $\hat{\gamma}$ = 0.84 (*SE* = 0.04). This estimate was significantly different from 1 (Wald-χ^2 = 16.88; p < .001). This effect indicates that moving an item by 20 positions toward the end of a test was expected to be related to a decrease in its item discrimination with respect to θ_0 by a factor of 0.84.

Correlates of item position effects

Decoding speed and reading enjoyment were weakly related to one another ($\hat{\rho}_{Dec,Enj}$ = .17; *SE* = .03; p < .001). Decoding speed was correlated with the ability variable θ_0 ($\hat{\rho}_{0Dec}$ = .28; *SE* = .02; p < .001), and the IP effect δ ($\hat{\rho}_{1Dec}$ = .21; *SE* = .04; p < .001). The same pattern was found for reading enjoyment, but this variable exhibited a weaker correlation with θ_0 ($\hat{\rho}_{0Enj}$ = .11; *SE* = .03; p < .001) than the δ-variable ($\hat{\rho}_{1Enj}$ = .23; *SE* = .05; p < .001). Hence, good decoders and students with high reading enjoyment were more likely to work on the test with higher persistence.

Figure 1 presents the probability of correct answers as a function of the items' positions and the student covariates at three levels. We used values of ±1.3 standard deviations from the mean because these values are close to the 10th and 90th percentiles of the normal distribution. Figure 1 shows that individuals scoring high on decoding speed were expected to be only weakly impacted by the negative IP1 effect. The (negative) IP1 effect was notably stronger in individuals scoring at or below the mean. IP effects were small for individuals scoring near the 90th percentile of the distribution of reading enjoyment. Students with enjoyment scores that were 1.3 standard deviations below the mean were expected to show strong declines in their probabilities of providing correct responses.

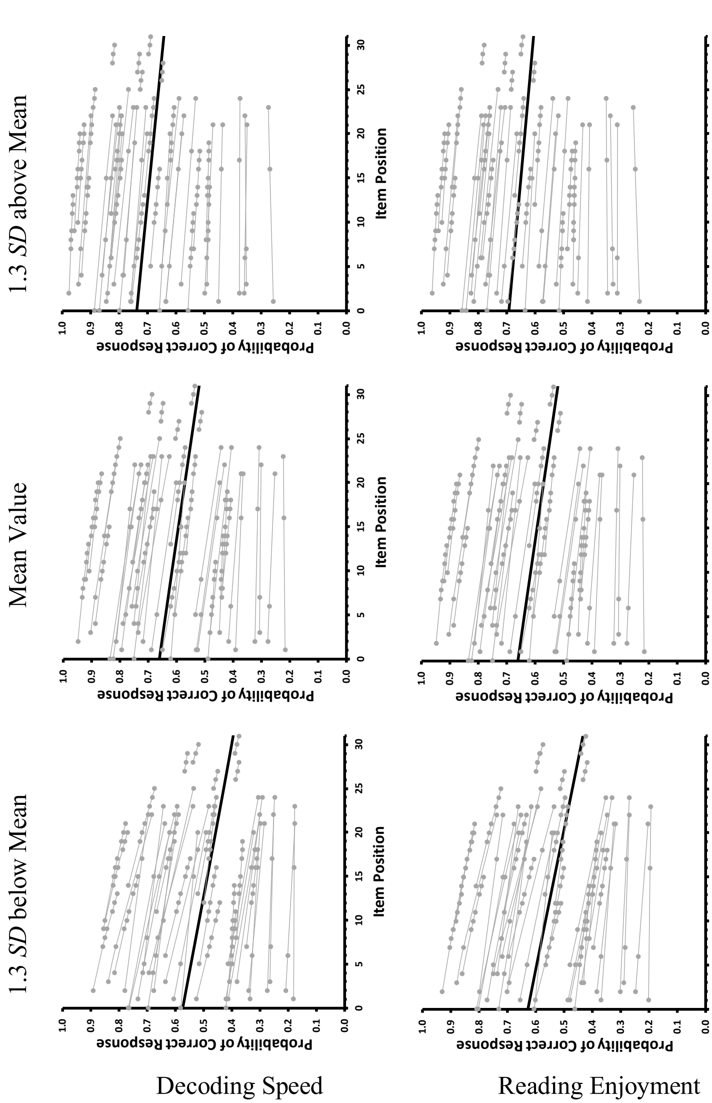

Figure 1:
Predicted probabilities of correct responses as a function of item positions and individuals' covariate values. Gray lines stand for the item-specific probabilities. Black lines stand for probabilities for a hypothetical item with average difficulty and discrimination.

Consequences of not accounting for individual differences in item position effects

In order to study the consequences of ignoring the RIP1 effect, we proceeded as follows: First, we estimated the covariates' relationship to the ability variable as assessed by the traditional 2PL model (Equation 1). Second, we estimated the latent correlations of reading ability across a range of reference positions (1 to 32) by utilizing the comprehensive model (Equations 4 and 8). Finally, we compared the two results and identified the reference position in which the correlations were most similar to the results provided by the 2PL model (minimum square root of the sum of squared deviations between the two sets of correlations).

In the unidimensional 2PL model, the correlation between reading ability and decoding speed was estimated to be $\hat{\rho}_{\theta Dec}$ = .34 (*SE* = .02; p < .001), whereas the correlation with reading enjoyment was estimated to be $\hat{\rho}_{\theta Enj}$ = .17 (*SE* = .03; p < .001). Both correlations were higher than the correlations derived by the model including RIP1 and IP2 effects. This result demonstrates that ignoring IP effects might lead to changes in the correlations of ability with external variables. The reason for this discrepancy is that, in our proposed model, the ability variable is defined with respect to a specific reference position (here, the first position), whereas the ability variable embedded in the traditional 2PL model represents an ability averaged across all possible item positions in a test (Robitzsch, 2009).

Figure 2 plots the correlations of the reading comprehension variable, defined with respect to all possible item positions. The correlations assessed by the traditional 2PL model most closely resembled the ability correlations assessed by the full model when reading comprehension was defined with respect to items presented in position 17. This result is in line with our arguments. Disregarding the RIP1 effect meant that the latent ability assessed in the unidimensional model represented an average of all possible ability variables defined with respect to all possible reference positions (i.e., 1 to 32). This average came close to the ability variable defined with reference to the middle position of the test.

A noteworthy result shown in Figure 2 is that the construct relationships were affected by the reference position chosen to define the ability variable. This effect is clearly visible for reading enjoyment. The correlations with the reading ability variable increased from $\hat{\rho}_{0Enj}$ = .11 when θ was defined with respect to the first item position to $\hat{\rho}_{32Enj}$ = .24 when θ was defined with respect to the last item position.

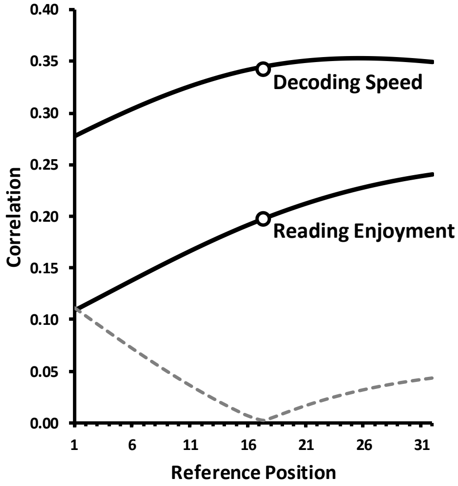

Figure 2:
Correlations of reading comprehension as a function of the reference position used to define the ability variable (solid lines) and correlations assessed by the unidimensional 2PL model (open circles). The dotted gray line stands for the square root of the sum of squared deviations between correlations assessed in the IRT model with and without position effects.

Discussion

This article attempts to shed light on the role that IP effects play in school achievement tests and to provide an insight into the consequences of ignoring them when assessing the relationship between the ability variable and other constructs (see also Nagy et al, 2016). Based on an example taken from the realm of reading comprehension testing, the article extends previous work on this topic in multiple ways.

First, we distinguished between two kinds of IP effects – those impacting on item difficulties (IP1 effects) and those impacting on item discriminations (IP2 effects). In line with previous research, we found clear evidence for negative IP1 effects, which could be conceptualized as random effects that vary across individuals (RIP1 effect; Debeer & Janssen, 2013). This conceptualization reflects the assumption that IP effects are driven by processes that take place within individuals (Ackerman & Kanfer, 2009). Consequently, the RIP1 effect indicates individual differences in the persistence with which test takers invest effort and work precisely on the test (Debeer et al, 2014; Hartig & Buchholz, 2012).

Second, our analyses indicated the presence of IP2 effects, showing that the items' potential to discriminate between ability levels decreased when they were presented in later positions. Our results indicate that a second process (i.e., an RIP1 effect) gradually takes over, which means that items presented in later positions are less reliable measures of reading ability. Therefore, the IP2 effect appears to be intertwined with the RIP1 effect. However, whether RIP1 effects are generally accompanied by IP2 effects is an open question that warrants further investigation. We believe that IP2 effects are especially likely to occur when the individual differences in RIP1 effects are large, and when tests are long.

Third, we studied the relationship between the RIP1 effect and two person characteristics hypothesized to be related to the persistence of effortful processing: decoding speed and reading enjoyment. Both variables were related to the RIP1 effect. Our results are in line with the expectation that reading is a less exhausting activity for fluent readers, who are characterized by good decoding skills, and that individuals who enjoy reading are more motivated to persist with working on reading tasks.

Finally, we studied the consequences of disregarding IP effects when assessing the relationship of the ability construct to external variables. Our analyses indicated that, in the presence of RIP1 effects, the correlations assessed by using the total score are a mix of at least two components. The correlations no longer reflect the relationship of the individuals' "pure" ability levels. Rather, they represent a composite of the relationship of their ability levels and the relationship of their test-taking behavior (i.e., persistence) to the constructs under study (e.g., Nagy et al, 2016; Robitzsch, 2009; Ren et al., 2012).

Theoretical and practical implications

The main implications of our article stem from the fact that the IP1 effects varied across individuals. This means that overall test scores are confounded by a minimum of two sources of variance (Robitzsch, 2009). Therefore, the following question arises: How large are the distorting effects of IP effects on the total test scores? Unfortunately, this question is hard to answer. Their influence is driven by two main factors: the mean and the variability of the RIP1 effect, and the test length. In addition, IP2 effects also affect the estimates of abilities, although they appeared to have a weaker impact on the results in the present application. However, we expect IP2 effects that lead to a reduction in item discriminations to play a more important role in longer tests, for example, those used in PISA.

Disregarding IP effects leads to an estimate of an individual's ability that is close to the average of the position-specific scores. Hence, the ability variable assessed in the unidimensional IRT model incorporates the person-specific IP effect. Therefore, the unidimensional model is a natural choice, when the degree of effort and precision maintained during a test is inherent to the definition of the ability construct. However, many applications call for estimates of ability that are, to some degree, corrected for the impact of individuals' test-taking persistence. For example, large-scale assessments attempt to estimate ability distributions that quantify what students can do in real-life situations; this means that test scores that are assessed in low-stakes conditions are generalized beyond the test. Strong IP effects call the generalizability of test scores into question. Many real-life situations are more similar to high-stakes conditions in which individuals are motivated to sustain their effort. In addition, in real life, individuals are often required to draw on their abilities in rather short tasks, so that IP effects are unlikely to unfold.

The IRT models presented here can be used to study the sensitivity of results pertaining to the ability distributions, and the relationships of abilities with covariates to IP effects. Our approach implies that there are as many ways to define individuals' ability levels as there are item positions (Robitzsch, 2009), which means that the mean and dispersion of the ability distribution as well as the relationships between ability and the covariates might change across positions. Therefore, researchers could track changes in the means, dispersions, and correlations across item positions in order to study their sensitivity to IP effects. If IP effects are found to affect results in a nontrivial way, researches could pick a reference position to adjust the results for IP effects. A natural choice is to select the first position because it provides results that pertain to a (hypothetical) situation in which all individuals maintain their initial effort and precision.

As an alternative, the ability to be measured could be defined over a range of positions (e.g., the first quarter of a test), thereby partially adjusting for IP effects. This approach is useful when individuals' persistence is an inherent part of the ability to be measured, but that there is reason to believe that giving IP effects full weight is not appropriate. The means, dispersions, and correlations of the partially adjusted ability variable can be derived from the parameter estimates provided by our IRT models, thereby obviating the need to exclude items.[1] We do not feel able to provide advice about the number of item positions to be retained in this approach because this requires knowledge about the size of IP effects in the situations to which test scores are generalized. However, researchers can inspect different scenarios by defining the ability variable on the basis of different ranges of positons.

Further research and conclusions

The generalizability of our findings to other domains and populations remains an open question. We are optimistic that the essence of our results can be replicated in diverse settings (e.g., Debeer & Janssen, 2013; Debeer et al., 2014; Hartig & Buchholz, 2012; Robitzsch, 2009), including the results about the correlates of RIP1 effects (e.g., Qian, 2014; Weirich et al., 2017). However, we believe that the list of the correlates of RIP1 effects needs to be extended to include other variables from various domains (e.g., cognitive variables, personality traits, interest, and self-concept measures). Future research could also study differences between achievement domains in order to gain a fuller understanding of the role that IP effects play (e.g., Nagy et al, 2016).

In addition, the IRT models used are not without limitations. In all of the models, IP1 effects were modeled by linear functions of item positions. Alternative specifications might use nonlinear effects, possibly accompanied by the effects of item cluster positions

[1] The ability variable defined across items provided from position $p = 0$ to $p = L$ ($L \leq P$) can be considered to be a combination of θ_0 and δ, such that $\theta_{iL} = \frac{1}{K}\sum_{p=0}^{L}(\theta_{i0} + \omega_p\delta_i)$. The mean of θ_L is given by $\kappa_L = \kappa_0 + \overline{\omega}_L\kappa_1$, where $\overline{\omega}_L = \frac{1}{K}\sum_{p=0}^{L}\omega_p$. The variance of θ_L adheres to $\phi_{LL} = \phi_{00} + \overline{\omega}_L^2\phi_{11} + 2\overline{\omega}_L\phi_{01}$. Finally, the covariance between a covariate k and θ_L can be derived as $\phi_{Lk} = \phi_{0k} + \overline{\omega}_L\phi_{1k}$, which can be standardized to derive the corresponding correlation.

(e.g., Debeer et al., 2014). Unfortunately, it would not have been straightforward to include item cluster positions in the present situation because the item clusters differed greatly in the number of items they comprised (i.e., 4 to 15 items). Further studies utilizing item clusters of similar length could investigate whether position effects located on the item and cluster level can be separated from each other.

IP effects caused by individuals' test-taking persistence could be considered as a threat to the validity of inferences derived on the basis of full test scores because the extent to which they can be generalized to situations beyond the test is unclear. The IRT models presented in this article provide a means for investigating the sensitivity of test results to IP effects, and for adjusting such effects. Therefore, the approaches are of relevance for many large-scale studies.

Authors' note

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

References

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied, 15*, 163–181. doi: 10.1037/a0015719

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education, 16*, 363–383. doi: 10.1007%2FBF03173188

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*, 92–104.

Auer, M., Gruber, G., Mayringer, H. & Wimmer, H. (2008). *Salzburger Lesescreening für die Klassenstufen 5-8 (SLS 5-8) [Salzburg reading screening for 5th to 6th grade students]*. Bern: Huber.

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39*, 502–523. doi: 10.3102/1076998614558485

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

Fischer, G. H. (1973). Linear logistic test model as an instrument in educational research. *Acta Psychologica, 37,* 359–374. doi: 10.1016/0001-6918(73)90003-6

Frey, A., Hartig, J. & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*, 39–53. doi: 10.1111/j.1745-3992.2009.00154.x

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

Granzer, D., Köller, O. & Bremerich-Vos, A. (2009). *Bildungsstandards Deutsch und Mathematik: Leistungsmessung in der Grundschule [Educational standards for German and mathematics at the end of lower secondary school]*. Weinheim, Germany: Beltz.

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54*, 418–431.

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. doi: 10.1080/00273170701341266

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.

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.

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

Le, L. T. (2007). *Effects of item positions on their difficulty and discrimination: A study in PISA Science data across test language and countries*. Paper presented at the 72nd Annual Meeting of the Psychometric Society. Tokyo, Japan.

Leary, L. F., & Dorans, N. J. (1985). Implications for altering the context in which test items appear: A historical perspective on an immediate concern. *Review of Educational Research, 55*, 387–413. doi: 10.3102/00346543055003387

Lehmann, R. H., Peek, R., & Poerschke, J. (1997). *HAMLET 3–4. Hamburger Lesetest für 3. und 4. Klassen [Hamburg reading test for grade 3 and 4]*. Weinheim, Germany: Beltz.

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

Meyers, J. L., Miller, G. E., & Way, W. D. (2009). Item position and item difficulty change in an IRT-based common item equating design. *Applied Measurement in Education, 22*, 38–60. doi: 10.1080/08957340802558342

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. doi: 10.1007/BF02289044

Muthén, L.K., & Muthén, B.O. (2012). Mplus user's guide. Seventh edition. Los Angeles, CA: Muthén & Muthén.

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

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

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

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

Nauck, J. & Otte, R. (1980). *Diagnostischer Test Deutsch (DTD 4--6) [Diagnostic Test German (DTD 4--6]*. Braunschweig, Germany: Westermann.

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

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

Retelsdorf, J., Köller, O., & Möller, J. (2011). On the effects of motivation on reading performance growth in secondary school. *Learning and Instruction, 21*, 550—559. doi: 10.1016/j.learninstruc.2010.11.001

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

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

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

Appendix

Specification of the IRT model assessing RIP1 and IP2 effects

Here we describe the implementation of the IRT model, given in Equation 4, in the M*plus* software (Muthén & Muthén, 2012). We first restructured the data summarized in Table 1. Each item × position combination was treated as a separate item. Second, we reformulated the model as a hierarchical IRT model in which first-order latent variables are expressed as consequences of second-order latent variables. Third, we distinguished between two sets of variables: one set of individual difference variables and one set of fixed variables.

The first-order model is represented as

$$logit\left[P\left(y_{ijp} = 1\right)\right] = \alpha_{j0}\eta_{ip} + \gamma^{\lambda p}\nu_{j0}, \qquad \text{(A1)}$$

where α_{j0}, and $\gamma^{\lambda p}$ are as defined above. Here, ν_{j0} is an item-specific threshold parameter defined in the first position. ν_0-parameters are modeled by means of J latent variables with zero variance and unrestricted mean structure. η_{ip} refers to the value of an individual 's latent variable underlying her or his responses to the items provided in position p.

The variables η_p are modeled as being fully determined by θ_0 and δ, such that

$$\eta_{ip} = \gamma^{\lambda p}\theta_{i0} + \lambda p\delta_i. \qquad \text{(A2)}$$

Inserting Equation A2 into Equation A1 gives

$$logit\left[P\left(y_{ijp} = 1\right)\right] = \alpha_{j0}\left(\gamma^{\lambda p}\theta_{i0} + \lambda p\delta_i\right) + \gamma^{\lambda p}\nu_{j0}, \qquad \text{(A3)}$$

which means that the β_{j0}-parameters from Equation 4 can be derived as $\beta_{j0} = -\nu_{j0}/\alpha_{j0}$.

Here we present parts of the M*plus* syntax that applies to the example used in this article. We cannot provide the full syntax because it recurs on 249 observed variables (i.e., combinations of items and positions), 32 η-variables, and 43 ν-variables. Therefore, we focus on three items, each assessed in three positions. The example includes comments (indicated by an exclamation mark) that provide advice on how to extend the syntax to more variables.

```
Title:      IRT model with RIP1 and IP2 effects
Data:       file = reading.dat;
Variable:   names =
            A4011_00 A4019_15 A4015_22 ! A401 in booklets 1, 9, 5 at p = 00, 15, 22
            B5012_00 B5014_04 B5019_19 ! B501 in booklets 1, 4, 9 at p = 00, 04, 19
            C4017_00 C4016_15 C4012_21 ! C401 in booklets 7, 6, 2 at p = 00, 15, 21
            ...;                       ! Extend variable list for remaining items
            categorical are all;  ! Specification of categorical observed variables
            missing are all (-9); ! Flag for missing values
Analysis:   estimator = ml;       ! Maximum likelihood estimation
            coverage = 0;         ! Minimum observed data coverage is 0
Model:
            ! Specification of eta-variables (for p = 00, 04, 15, 19, 21, 22)
            eta00 by A4011_00*1 (aa401);
            eta00 by B5012_00*1 (ab501);
            eta00 by C4017_00*1 (ac401);
            eta00 by ...                ! Extend for other items at p = 00
            ...                         ! Include specifications for eta01 to eta03
            eta04 by B5014_04*1 (ab501);
            eta04 by ...                ! Extend for other items at p = 04
            ...                         ! Include specifications for eta05 to eta14
            eta15 by A4019_15*1 (ab501);
            eta15 by C4016_00*1 (ac401);
            eta15 by ...                ! Extend for other items at p = 15
            ...                         ! Include specifications for eta16 to eta18
            eta19 by B5014_04*1 (ab501);
            eta19 by ...                ! Extend for other items at p = 19
            ...                         ! Include specifications for eta20
            eta21 by C4012_21*1 (ac401);
            eta21 by ...                ! Extend for other items at p = 21
            eta22 by A4015_22*1 (aa401);
            eta22 by ...                ! Extend for other items at p = 22
            ...                         ! Include specifications for eta23 to eta31

            eta00-eta31@0; ! (Residual-)Variances of eta-variables fixed to 0

            ! Specification of theta-variable
            theta by eta00@1;         ! Fixed loading because gamma^0 = 1
            ...                       ! Include specifications for eta01 to eta03
            theta by eta04*1 (gam03); ! Labeled loading for p = 04
            ...                       ! Include specifications for eta05 to eta14
            theta by eta15*1 (gam15); ! Labeled loading for p = 15
            ...                       ! Include specifications for eta16 to eta18
            theta by eta19*1 (gam19); ! Labeled loading for p = 19
            theta by eta20*1 (gam20); ! Labeled loading for p = 20
            theta by eta21*1 (gam21); ! Labeled loading for p = 22
            theta by eta22*1 (gam22); ! Labeled loading for p = 22
            ...                       ! Include specifications for eta23 to eta31

            theta@1; ! Variance of theta set to 1

            ! Specification of delta-variable
            delta by eta00@0.00; ! Fixed loading: 00/20 = 0.00
            ...                  ! Include specifications for eta01 to eta03
            delta by eta04@0.20; ! Fixed loading: 04/20 = 0.75
            ...                  ! Include specifications for eta05 to eta14
            delta by eta15@0.75; ! Fixed loading: 15/20 = 0.75
            ...                  ! Include specifications for eta16 to eta18
            delta by eta19@0.95; ! Fixed loading: 19/20 = 0.95
            delta by eta20@1.00; ! Fixed loading: 20/20 = 1.00
            delta by eta21@1.05; ! Fixed loading: 21/20 = 1.05
            delta by eta22@1.10; ! Fixed loading: 22/20 = 1.10
            ...                  ! Include specifications for eta23 to eta31
            [delta*-0.5]; ! Mean of delta is free;
            ! Specification of item-difficulties

            ! Thresholds fixed to 0
            [A401$1-XXXX$1@0]; ! XXXX$1 = threshold of the last item

            ! Latent variable for threshold of item A401
            A401 by A4011_00*1 (gam00); ! A401 in booklet = 1 in position p = 00
            A401 by A4019_15*1 (gam15); ! A401 in booklet = 9 in position p = 15
            A401 by A4015_22*1 (gam22); ! A401 in booklet = 5 in position p = 22
            ! Latent variable for threshold of item B501
            B501 by B5012_00*1 (gam00); ! B401 in booklet = 2 in position p = 00
            B501 by B5014_04*1 (gam04); ! B401 in booklet = 4 in position p = 04
            B501 by B5019_19*1 (gam19); ! B401 in booklet = 9 in position p = 19
            ! Latent variable for threshold of item C401
            C401 by C4017_00*1 (gam00); ! C401 in booklet = 2 in position p = 00
            C401 by C4016_15*1 (gam15); ! C401 in booklet = 4 in position p = 04
            C401 by C4012_21*1 (gam21); ! C401 in booklet = 9 in position p = 19
            ... ! Include specifications for remaining items

            A401-XXXX@0;                    ! Var set to 0 (XXXX = last item)
            A401-XXXX with A401-XXXX@0      ! Cov among thresholds set to 0
            A401-XXXX with theta@0 delta@0; ! Cov with theta and delta fixed to 0
            [A401-XXXX];                    ! Means of thresholds are free

Model Constraint:
            new(gam*1); ! Specification of gamma-parameter
            gam01 = gam**0.05; ! Power of gam at p = 01/20
            ...                ! Include specifications for p = 02 to 03
            gam04 = gam**0.20; ! Power of gam at p = 04/20
            ...                ! Include specifications for p = 05 to 14
            gam15 = gam**0.75; ! Power of gam at p = 15/20
            ...                ! Include specifications for p = 16 to 18
            gam19 = gam**0.95; ! Power of gam at p = 19/20
            gam20 = gam**1.00; ! Power of gam at p = 20/20
            gam21 = gam**1.05; ! Power of gam at p = 21/20
            gam22 = gam**1.10; ! Power of gam at p = 22/20
            ...                ! Include specifications for p = 23 to 30
            gam31 = gam**1.55; ! Power of gam at p = 31/20

Output:     stand tech1 tech8;
```