Journal of Educational and Behavioral Statistics
2014, Vol. 39, No. 6, pp. 502–523
DOI: 10.3102/1076998614558485
© 2014 AERA. http://jebs.aera.net

Student, School, and Country Differences in Sustained Test-Taking Effort in the 2009 PISA Reading Assessment

Dries Debeer
University of Leuven

Janine Buchholz
Johannes Hartig
German Institute for International Educational Research (DIPF)

Rianne Janssen
University of Leuven

In this article, the change in examinee effort during an assessment, which we will refer to as persistence, is modeled as an effect of item position. A multilevel extension is proposed to analyze hierarchically structured data and decompose the individual differences in persistence. Data from the 2009 Program of International Student Achievement (PISA) reading assessment from N = *467,819 students from 65 countries are analyzed with the proposed model, and the results are compared across countries. A decrease in examinee effort during the PISA reading assessment was found consistently across countries, with individual differences within and between schools. Both the decrease and the individual differences are more pronounced in lower performing countries. Within schools, persistence is slightly negatively correlated with reading ability; but at the school level, this correlation is positive in most countries. The results of our analyses indicate that it is important to model and control examinee effort in low-stakes assessments.*

Keywords: *examinee effort; testing; item response theory; PISA*

Introduction

Educational policymakers attach great importance to the outcomes of large-scale international assessments such as the Program of International Student Achievement (PISA) and the Trends in International Mathematics and Science Study (TIMSS). Performance changes in these studies are used to evaluate and develop educational programs and policies. In contrast with the high-stakes

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

implications attached to the results, test takers commonly perceive these assessments as low stakes, as there are no personal consequences related to their performance on the test. This might cause some test takers to expend low effort during the assessment, which can result in biased and invalid measurements. Therefore, a high-stakes question arises for low-stakes assessments (Barry, Horst, Finney, Brown, & Kopp, 2010): Does examinee effort—and the differences therein—form a threat for the validity of international assessments?

Most research regarding the issue of low examinee effort in low-stakes assessments addresses the issues of manipulating examinee effort, accounting for low examinee effort, and measuring examinee effort (e.g., Steedle, 2014; Swerdzewski, Harmes, & Finney, 2011; Waskiwicz, 2011; Wise & DeMars, 2005; Wise & Kong, 2005). In these studies, examinee effort is commonly seen as constant throughout the assessment. However, research has shown that performance in large-scale assessments can decrease (e.g., Hohensinn et al., 2008; Meyers, Miller, & Way, 2009), possibly due to fatigue or a decline in motivation. Hence, it seems likely that, during testing, a change in examinee effort can take place.

This article addresses the change in examinee effort in large-scale, low-stakes assessments. An item response theory (IRT) model for effects of item position (Debeer & Janssen, 2013; e.g., Hartig & Buchholz, 2012) is proposed to investigate changes in examinee effort. The model is extended to fit the hierarchical data structure that is commonly present in international assessments. The extended model will be applied to data from the 2009 PISA reading assessment to examine the change in examinee effort during testing. Differences in this change within and between countries will be assessed and their relation with the PISA country score will be investigated.

In the following, first examinee effort and its relation to performance will be discussed. Then, it will be explained how this relation can cause validity problems, especially in low-stakes assessments. A brief overview of the current methods and techniques for dealing with this issue will be given. Finally, an IRT model to model a change in examinee effort and its multilevel extension will be proposed.

Examinee Effort in Low-Stakes Assessments

Examinee effort or test motivation refers to "a student's engagement and expenditure of energy toward the goal of attaining the highest possible score on the test" (Wise & DeMars, 2005, p. 2). A high expenditure of energy is needed for demanding tasks, such as responding to test items in an achievement test. When examinee effort is low, a test taker will not fully engage his or her ability, which will lead to a worse performance than what could be expected, given the test taker's ability. This relation between examinee effort and performance has been repeatedly found (e.g., Abdelfattah, 2010; Liu, Bridgeman, &

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

Adler, 2012; Steedle, 2014; Swerdzewski et al., 2011; Waskiwicz, 2011; Wise & DeMars, 2005).

The expectancy-value model proposed by Eccles and Wigfield (e.g., Eccles, 1983; Wigfield & Eccles, 2000) provides a useful perspective for understanding examinee effort in testing situations. According to this model, many test takers will hold weak value beliefs on the tests in the context of a low-stakes assessment because there are no consequences or personal benefits associated with student performance. A weak value belief combined with the awareness of the costs associated with the assessment will—according to the expectancy-value model—lead to low examinee effort.

This theoretical prediction has been empirically confirmed in several studies. When test takers do not perceive the importance or usefulness of an exam, their test-taking effort will be lower (Cole, Bergin, & Whittaker, 2008). Similarly, when students are asked to evaluate their testing motivation after completing a low-stakes assessment, they indicate that the effort they exerted was lower than the effort they would exert when the assessment was high stakes (Butler & Adams, 2007). Eklöf, Pavešič, and Grønmo (2014) found that the reported examinee effort on the low-stakes 2008 TIMSS test was on average low and that there was a relationship between reported effort and test performance.

Examinee effort is on average not only lower in low-stakes assessments compared to high-stakes assessments, but it is also likely to be more variable (Barry et al., 2010). Because examinee effort is related to test performance, and low examinee effort tends to result in a distorted ability estimate, the exerted effort can be a source of construct-irrelevant variance (Haladyna & Downing, 2004). Therefore, the relation between examinee effort and performance together with the variability of examinee effort can threaten the validity of test scores in low-stakes assessments (Wise & DeMars, 2010).

Measuring Examinee Effort

Different methods and techniques have been proposed for measuring examinee effort. A first strategy is to use self-report questionnaires after the assessment, such as the Student Opinion Survey (SOS; Sundre & Moore, 2002; Wolf & Smith, 1995), which was found to yield high values (mid- to upper 80s) for coefficient α in college samples (Sundre & Moore, 2002) or the Effort Thermometer (Kunter et al., 2002) used in PISA studies. Self-report measures, however, may have accuracy and validity problems (Wise & DeMars, 2005). Less motivated students may respond more carelessly or untruthfully. Moreover, low-performing students may attribute their performance to low effort instead of to their ability level (Wise & Kong, 2005).

Wise and Kong (2005) proposed an alternative strategy to measure examinee effort, namely response time effort (RTE), which is a reaction time–based measure used in computer-based testing. RTE supposes that there are two distinct

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

response behaviors: solution behavior and rapid-guessing behavior, which are assumed to correspond to high and low effort, respectively. By setting a response time threshold for every item, the response behavior is classified as follows: slower than (or equal to) the threshold is regarded as solution behavior and faster than the threshold as rapid-guessing behavior. The proportion of items for which a test taker is classified into solution behavior gives a test taker's RTE. The applicability of RTE as a measurement of examinee effort has been repeatedly demonstrated (Silm, Must, & Taeht, 2013; Steedle, 2014; Swerdzewski et al., 2011; Wise, Pastor, & Kong, 2009). The issue of setting the response time threshold has been addressed by Kong, Wise, and Bhola (2007). However, RTE is not without problems. It requires response time information, which is not available in many low-stakes assessments, and uses a deterministic classification of response behavior. Moreover, because it equates low examinee effort to rapid guessing, it assumes that solution behavior is not affected by low examinee effort.

Dealing With Low Examinee Effort

Several procedures have been suggested to deal with the issue of low examinee effort. One approach is to manipulate the students' test-taking motivation, for instance, by increasing the stakes of the assessment by making the test performance part of the grading system or by explaining the importance of the low-stakes assessments. Different manipulating strategies have been shown to improve test-taking motivation and increase test performance (Liu et al., 2012; Wise & DeMars, 2005).

A second approach is motivation filtering. Unmotivated test takers or test takers exerting low effort are deleted from the sample (Sundre & Wise, 2003; Wise & DeMars, 2005). Two important assumptions are made, namely, first, that it is possible to detect the low-effort test takers and validly measure examinee effort and, second, that there is no relation between test-taking effort and the actual level of proficiency. Results show that motivation filtering increases the average test performance (Steedle, 2014; Swerdzewski et al., 2011; Wise & DeMars, 2010; Wise, Wise, & Bhola, 2006), both when a self-questionnaire and RTE are used to measure examinee effort. Rios, Liu, and Bridgeman (in press) showed that RTE filtering, however, had a slightly stronger relationship with test performance.

A third way to address the low-effort issue is to include test-taking effort into the measurement model. Both Wise and DeMars (2006) and Meyer (2010) proposed an IRT model that incorporates the response time to classify item responses into rapid-guessing behavior and solving behavior. Within these models, it is assumed that low examinee effort is related to rapid guessing, and therefore, a very quick response time can be seen as proxy of low examinee effort. Although both models can increase the validity of the proficiency measurement, the problems mentioned with regard to RTE also exist here.

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

Another model worth mentioning is the model of Goegebeur, De Boeck, Wollack, and Cohen (2008), as it also jointly models guessing behavior and problem-solving behavior. The model assumes that during an assessment, there may be a gradual shift from problem-solving behavior to guessing behavior starting at a person-specific speededness point in the test. An advantage of the model is that response accuracy is modeled and no response time information is needed. However, this model was explicitly proposed for speeded tests to model the increase in rapid-guessing behavior. Because low-stakes tests are commonly designed to be nonspeeded, this model is less apt for modeling low examinee effort in low-stakes testing.

Change in Examinee Effort During Testing

Most studies on test-taking motivation and examinee effort implicitly assume that motivation or effort does not change during testing. The definition for examinee effort (Wise & DeMars, 2005) and the expectancy-value model (e.g., Eccles, 1983; Wigfield & Eccles, 2000), however, do not restrict examinee effort to be constant during an assessment. Moreover, it seems rather likely that the effort a test taker expends to solve individual items is not the same for every item. Given that in longer assessments test takers can become fatigued or less motivated, a downward trend in examinee effort can be expected, rather than random changes in examinee effort during testing. Indeed, studies using response time as an indicator of rapid-guessing behavior and low examinee effort indicate that one of the best predictors of rapid guessing is the position of the item in a test (Wise, 2006; Wise et al., 2009).

Modeling change in examinee effort. Debeer and Janssen (2013) proposed an IRT-based framework to model proficiency and change in performance related to item position during testing. A possible interpretation of this change dimension is a change in examinee effort that can vary over persons and that can affect performance during the assessment. In order to apply their framework, items have to appear in different positions to disentangle the effects of item difficulty and item position. Hence, the model is only applicable when the test consists of (partly overlapping) test forms, and item orders are different across test forms, or when item parameters are known.

A one-parameter logistic version of the model of Debeer and Janssen (2013) with a linear item position effect for an assessment with P test takers and I binary test items that can be administered in K positions, reads as:

$$\text{logit}\left[Y_{pik} = 1 | \theta_p, \delta_p\right] = \theta_p - \beta_i + (\gamma + \delta_p)(k - 1). \quad (1)$$

Y_{pik} is the response of person p to item i, which was administered at position k. θ_p is the proficiency of person p, and β_i is the difficulty of item i when administered at the first position of the test. The linear effect of item position $(\gamma + \delta_p)$ is the

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

change in performance that takes place during testing, where γ is the average change and δ_p is the deviation from this average for person p. The individual change in performance will be referred to as persistence (cf. Hartig & Buchholz, 2012). A positive value for $(\gamma + \delta_p)$ indicates an increase in performance, a negative value a decrease. θ_p and δ_p follow a bivariate normal distribution with variances σ^2_θ and σ^2_δ, and the covariance $\sigma_{\theta\delta}$ (or correlation $\rho_{\theta\delta}$). Both the variances and the correlation are free parameters in the model. Hence, the relation between persistence and ability can be investigated, and one does not have to assume that persistence is independent from ability.

Multilevel extension. Large-scale international assessments, such as PISA, often use a systematic stratified sampling procedure that results in a hierarchical data structure. In the case of PISA, students are nested within schools. The model in Equation 1 can be hierarchically extended, resulting in a multilevel decomposition of the random effects. More specifically, the variance and covariance of ability and persistence are decomposed into a between-school part and a within-school part:

$$\text{logit}\left[Y_{spik} = 1|\theta_p, \delta_p\right] = \left(\theta_s + \theta_{ps}\right) - \beta_i + (\gamma + \delta_s + \delta_{ps})(k-1). \quad (2)$$

θ_s and θ_{ps} represent the between-school part and the within-school part for ability, respectively. The same holds for the persistence parameters δ_s and δ_{ps}. It is assumed that θ_s and δ_s follow a bivariate normal distribution over schools with variances $\left(\sigma^2_{\theta s}, \sigma^2_{\delta s}\right)$ and covariance $(\sigma_{\theta\delta s})$. The remaining individual differences within schools for ability θ_p and persistence δ_{ps} are assumed to be bivariate normally distributed over students with variances $\left(\sigma^2_{\theta ps}, \sigma^2_{\delta ps}\right)$ and covariance $(\sigma_{\theta\delta ps})$. The mean vectors for both bivariate normal distributions are set to zero to be able to identify the model.

The multilevel version of the model may help in providing insights into the nature of the change in examinee effort. Using the multilevel decomposition, it is possible to investigate whether the variance in persistence is located at the school level or at the individual level. For example, schools may differ in stressing the high-stakes implications resulting in different ''testing climates'' between schools. Also, the correlation between persistence and ability can also be investigated within and between schools.

The Present Study

Hartig and Buchholz (2012) investigated the decrease in performance in the PISA 2006 science assessment in 10 of the 57 participating countries using the model in Equation 1. They found a significant negative effect of item position, consistently across the 10 countries, but with more prominent effects in countries with lower national performance levels. Although science ability and

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

persistence were practically uncorrelated in high-performing countries, a negative correlation was found in lower performing countries. This study intends to generalize their findings for the 2009 PISA reading assessment data for all participating countries with the multilevel extended model (Equation 2).

PISA 2009 reading assessment. The Program for International Student Assessment (PISA) is a triennial system of international assessments that focus on the competencies of 15-year-olds in reading, mathematics, and science literacy. In 2009, reading literacy was the major domain. Because PISA uses a rotated block design, students were administered only a part of all the reading items. Clusters of items were presented at different cluster positions across students. This is a requisite to investigate effects of item position and the change in examinee effort during testing.

Research questions. It can be assumed that the effects of item position observed in the science assessment (Hartig & Buchholz, 2012) are of a general nature and that there are no reasons to believe that they differ from the effects found within other domains. Hence, the following hypothesis can be formulated. We expect a general negative effect of cluster position on reading performance (Hypothesis 1) that indicates a decrease in examinee effort during the assessment.

Second, given previously found results (Debeer & Janssen, 2013; Hartig & Buchholz, 2012), we expect that there are individual differences in the decrease in examinee effort (Hypothesis 2). Further, we will examine the variability in persistence within and between schools. We hypothesize that most of the variance is found within schools (Hypothesis 2a). And, as we expect that school regime and (implicit) test expectations are different between schools, at least a part of the variance in persistence is related to the school level (Hypothesis 2b).

Third, the correlation between persistence and reading ability is estimated. Given the findings of Hartig and Buchholz (2012), we expect a small or no correlation between ability and persistence in the reading assessment (Hypothesis 3), both within (Hypothesis 3a) and between (Hypothesis 3b) schools.

Finally, the results will be compared across all countries participating in PISA 2009. By relating the national reading score for a country to the results of the analyses, more insights into the nature and relevance of the effects might be obtained, and differences between high- and low-performing countries can be observed. Hartig and Buchholz (2012) found that the individual differences in persistence are more pronounced in lower performing countries, and that the negative correlation between persistence and science ability is stronger in low-performing countries, while there was no correlation in high-performing countries. We expect to find similar results in PISA 2009, across all countries (Hypothesis 4).

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

Method

Participants

In total, 467,819 students from 65 countries participated in the PISA 2009 assessment. Within each country, students were drawn through a two-tiered stratified sampling process consisting of a systematic sampling of individual schools with a probability proportional to the school size, from which 35 students were randomly selected. More details about the sampling procedure can be found in the PISA 2009 technical report (Organization for Economic Cooperation and Development [OECD], 2012).

Procedure

In the assessment, there were 218 test items (131 reading, 34 math, and 53 science). The items were partitioned in 13-item clusters: 7 for reading (R1–R7), 3 for math (M1–M3), and 3 for science (S1–S3). Each cluster represented 30 minutes of test time. Countries that were expected to have a lower reading score were offered the option of administering an easier set of items. For those countries, two of the standard reading clusters (R3A and R4A) were substituted with two easier reading clusters (R3B and R4B). The sets of items in the standard and easier clusters were matched in terms of the distribution of text format, aspect, and item format. The other 11 clusters were administered in all countries. In total, 20 countries opted to administer the easier clusters.

The items were presented to students in 13 standard test booklets (Booklet 1–13) and 7 easier booklets (Booklet 21–27),[1] with each booklet being composed of four clusters (Table 1). Using a balanced incomplete block design, each item cluster appeared in each of the four possible cluster positions within a test booklet once. This way, each pair of item clusters appears in only one booklet. Within the item clusters, the position of the items was fixed. Therefore, the effects of cluster position will be modeled instead of the effects of item position. Applied to Equations 1 and 2, k is replaced by c which is the position of the item cluster, ranging from 1 to 4. Each sampled student was randomly assigned to 1 of the 13 test booklets available in a country.

Data

Only data from the PISA 2009 paper-and-pencil reading literacy assessment[2] will be analyzed. The item formats employed for reading items were either selected response multiple choice or constructed response. Both dichotomous and partial credit scoring are used in PISA. In total, 125 reading items were analyzed.[3] To fit the binary item response model of Equations 1 and 2, 7 items were dichotomized by only considering a full credit response as correct. Further, not-reached responses were dropped, and missing responses were

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

TABLE 1
Visual Representation of the PISA 2009 Rotated Block Design

Booklet	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
1	*21*	M1		R1		R3A	*R3B*	M3	
2	*22*	R1		S1		R4A	*R4B*	R7	
3	*23*	S1		R3A	*R3B*	M2		S3	
4	*24*	R3A	*R3B*	R4A	*R4B*	S2		R2	
5	*25*	R4A	*R4B*	M2		R5		M1	
6	*26*	R5		R6		R7		R3A	*R3B*
7	*27*	R6		M3		S3		R4A	*R4B*
8		R2		M1		S1		R6	
9		M2		S2		R6		R1	
10		S2		R5		M3		S1	
11		M3		R7		R2		M2	
12		R7		S3		M1		S5	
13		S3		R2		R1		R5	

Note. PISA = Program of International Student Achievement. Each booklet consists of 4 of the 13 item clusters (7 reading clusters [R1–R7], 3 math clusters [M1–M3], and 3 science clusters [S1–S3]). There are four cluster positions, and each item cluster is presented at every cluster position once. The easier booklets (21–27) and clusters (R3B and R4B) are represented in italic.

treated as incorrect. More information on the items, the response formats, and the scoring rules can be found in the PISA 2009 technical report (OECD, 2012).

Analysis

The models in Equations 1 and 2 were used to analyze the data within each country separately. Both models can be seen as generalizations of the Rasch model or the logistic multilevel model with item responses as Level-1 variable nested within students (e.g., Kamata, Bauer, & Miyazaki, 2008). As item responses are nested in students, we will refer to the model in Equation 1 as the two-level model. The model in Equation 2 will be referred to as the three-level model, with responses nested in students and students nested in schools. All analyses were conducted with the multilevel software HLM (Raudenbush, Bryk, & Congdon, 2004, 2013) using penalized quasi-likelihood estimation.

Because the analyses are conducted separately for each country, there is no common scale, and the estimated effects are not directly comparable across countries. Therefore, for the two-level results, the estimated effect of cluster position $\hat{\gamma}$ was standardized using the standard deviation of the ability level within each country σ_θ : $\gamma^* = \gamma/\sigma_\theta$ (cf. Hartig & Buchholz, 2012). The standardized coefficient γ^* is the effect of one cluster position on performance, expressed in standard deviations in reading ability within each country.

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

Similarly, the standard deviation of persistence was standardized: $\sigma^*_\delta = \sigma_\delta / \sigma_\theta$. Hence, σ^*_δ expresses the individual differences in persistence within a country relative to the individual differences in reading ability.

For the three-level model, the total reading ability standard deviation $\left(\sqrt{\sigma^2_{\theta s} + \sigma^2_{\theta ps}}\right)$ was used to standardize the following parameters: γ, $\sigma_{\delta s}$, and $\sigma_{\delta ps}$, resulting in γ^*, $\sigma^*_{\delta s}$, and $\sigma^*_{\delta ps}$, respectively. Further, the intraclass correlation (ICC) for persistence and ability will be computed for every country: $\mathrm{ICC}_\delta = \sigma^2_{\delta s} / \left(\sigma^2_{\delta s} + \sigma^2_{\delta ps}\right)$ and $\mathrm{ICC}_\theta = \sigma^2_{\theta s} / \left(\sigma^2_{\theta s} + \sigma^2_{\theta ps}\right)$. The ICC gives the proportion of variance in persistence and ability that is located between schools, respectively.

Results

An overview of the parameters of interest for every country can be found in Online Appendices A and B (available at http://jeb.sagepub.com/supplemental). Online Appendix A (available at http://jeb.sagepub.com/supplemental) lists the estimates of σ^2_θ, γ, γ^*, σ_δ, σ^*_δ, and $\rho_{\theta\delta}$ for the two-level analyses, and Online Appendix B (available at http://jeb.sagepub.com/supplemental) lists the estimates of σ^2_θ, γ, γ^*, $\sigma_{\delta s}$, $\sigma_{\delta ps}$, $\sigma^*_{\delta s}$, $\sigma^*_{\delta ps}$, $\rho_{\theta\delta s}$, $\rho_{\theta\delta ps}$, and the ICC for ability and for persistence for the three-level analyses. Overall, without taking the decomposition into account, the two-level and three-level results are very similar. Because the three-level model is in line with the hierarchical structure of the PISA data, we focus on these results. Whenever discrepancies are found with the two-level results, these are discussed.

Average Persistence

As expected in Hypothesis 1, a negative effect of cluster position is consistently found across all participating countries. On average, there is a decline in examinee effort during testing, which results in a decreasing probability of a correct response when the item is placed further in the test. Figure 1 gives the distribution of the estimated standardized average persistence $\hat{\gamma}^*$ across countries with mean $\overline{\gamma^*} = -0.17$ ($SD = 0.034$). Hence, one can say that on average, the difficulty of an item increases by 0.17 standard deviations of the reading ability θ when it is moved one cluster position further in the test. There are considerable differences between countries in the average persistence, the highest average is found in Finland ($\hat{\gamma}^* = -0.09$) and the lowest in Greece ($\hat{\gamma}^* = -0.28$). To illustrate the impact of these effects, Table 2 lists the change in probability of a correct response of students with average ability ($\theta = 0$) when an item of average difficulty ($\beta_i = 0$) is placed on cluster position one or three cluster positions further in the assessment; for three $\hat{\gamma}^*$ values (the lowest, the average, and the highest value). Although

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

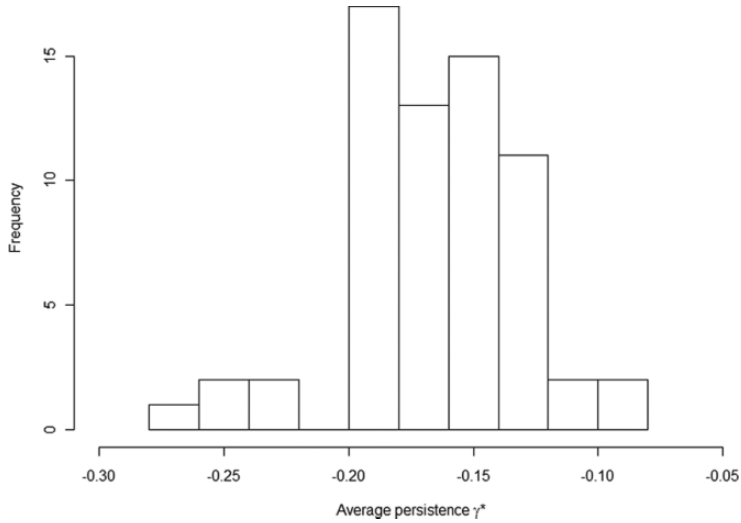

FIGURE 1. *Histogram of the average estimated persistence γ* across all countries (*N = *65) according to the three-level analyses (Mean = −0.166;* SD = *0.034).*

TABLE 2

Effect of the Decrease in Examinee Effort γ During the PISA 2009 Reading Assessment on the Change in Probability of a Correct Response of Students of Average Ability (*$\theta = 0$*), When an Item of Average Difficulty (*$\beta_i = 0$*) Is Placed One Cluster Position or Three Cluster Positions Further in the Assessment*

	Change in Cluster Positions	
Standardized Decrease in Examinee Effort $\hat{\gamma}^*$	+1 Cluster Positions	+3 Cluster Positions
−0.277 (Greece)	−.069	−.196
−0.167 (Average)	−.042	−.122
−0.093 (Finland)	−.023	−.069

Note. PISA = Program of International Student Achievement. The changes in the probability of a correct response are given for three effect sizes: the highest (Greece), the average, and the lowest (Finland) decrease in examinee effort. Estimates from the three-level analyses are used.

the change in probability is rather small when the item is moved one cluster position, the effect is considerable when the item is moved three cluster positions further in the test.

There are no discrepancies between the two-level and the three-level estimates for γ* (root mean square difference [RMSD] = 0.007). The country with the largest difference was Slovenia, with a difference of 0.022. Hence, both

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

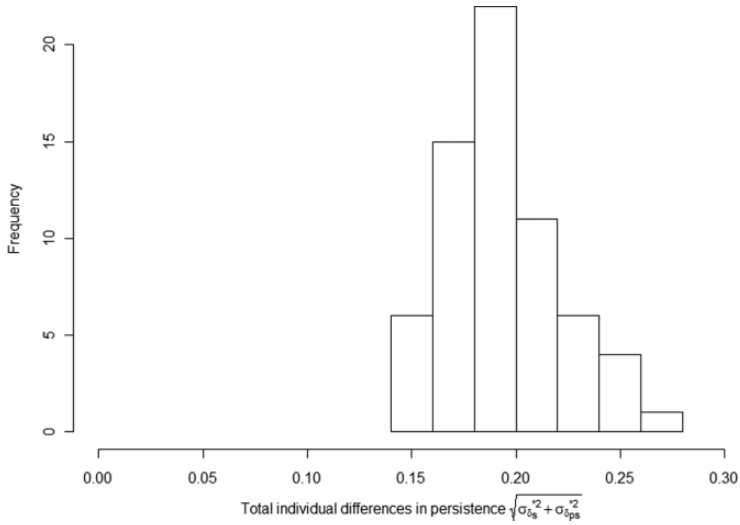

FIGURE 2. *Histogram of the estimated total individual differences in persistence* $\sqrt{\sigma^{*2}_{\delta s} + \sigma^{*2}_{\delta ps}}$ *across all countries (*N = *65) according to the three-level analyses (Mean = 0.194,* SD = *0.028).*

the two-level and the three-level results are in line with our first Hypothesis 1: In all countries, there is a decrease in average persistence during testing.

Individual Differences in Persistence

Figure 2 gives the distribution of the estimated total individual differences in persistence relative to the individual differences in reading ability $\left(\sqrt{\sigma^{*2}_{\delta s} + \sigma^{*2}_{\delta ps}}\right)$ across countries. Although considerably smaller than the individual differences in reading ability, individual differences in persistence are found in all countries, ranging from 14% of the standard deviation of ability in Shanghai–China to 27% in Indonesia. These results are in line with Hypothesis 2. When the total individual differences in persistence of the three-level analyses $\left(\sqrt{\hat{\sigma}^{*2}_{\delta s} + \hat{\sigma}^{*2}_{\delta ps}}\right)$ are compared with the two-level estimates $\hat{\sigma}^{*}_{\delta}$, the differences are very small (RMSD = 0.006).

Given the size of the individual differences in persistence, in all countries, at least a proportion of students demonstrate an increase in examinee effort, and hence, an increase in the probability of a correct response for an item when it is administered at a later cluster position in the assessment. On average, about 20% of the students have a zero or a positive change in examinee effort during the assessment. Although an increase in examinee effort seems counterintuitive,

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

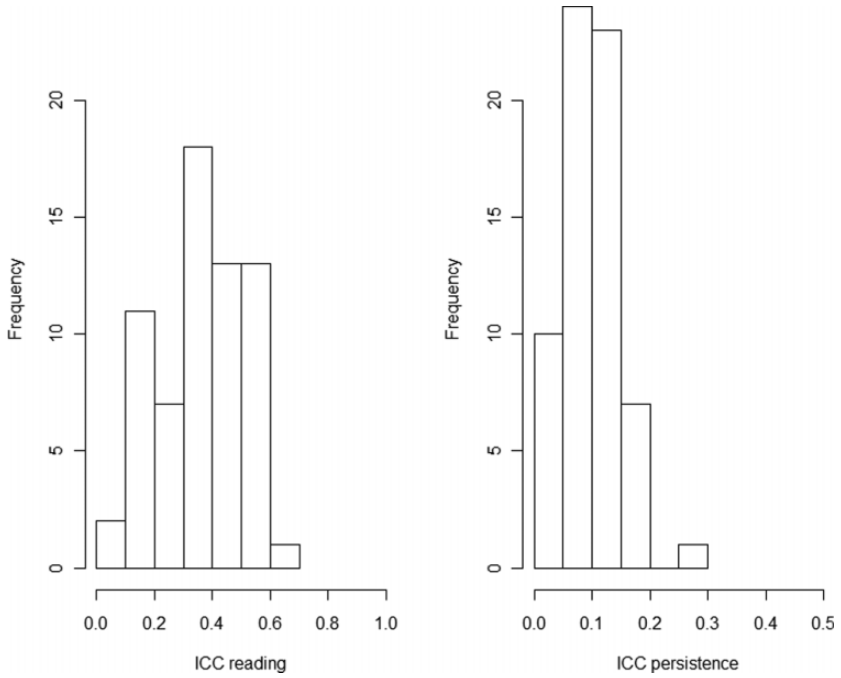

FIGURE 3. *Histogram of the ICC of reading ability (a) and the ICC of persistence (b) across the participating countries (*N = *65). The mean ICC for reading ability is .360 (*SD = *0.146), and the mean ICC for persistence is .100 (*SD = *0.048). ICC = intraclass correlation.*

a possible explanation is that some test takers might exert very low effort in the beginning of the test, which makes an increase in examinee effort more likely than a decrease.

The three-level model decomposes the individual differences in persistence and reading ability in a within-school and a between-school part. Figure 3 gives the distribution of the ICC across countries for (a) reading ability and (b) persistence. In all countries, only a small proportion of the differences in persistence is related to between-school differences. On average, the proportion is about 10% ($SD = 0.048$). This proportion is considerably smaller than the proportion for the individual differences in reading ability, where the ICC is on average about 36% ($SD = 0.146$). The findings are in line with the second hypothesis (Hypotheses 2a and 2b): At least a part of the individual differences in persistence can be explained by the school level.

Correlation Between Reading Ability and Persistence

Before examining the decomposition of the correlation between ability and persistence in the three-level model, first the two-level correlations $\rho_{\theta\delta}$ are

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

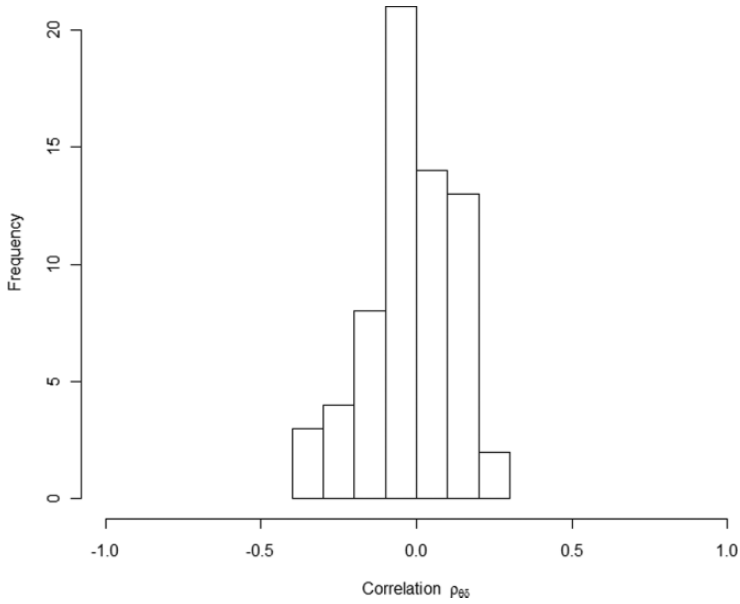

FIGURE 4. *Histogram of the correlation between a student's ability and persistence* $\rho_{\theta\delta}$ *across all countries (*N = *65) according to the two-level analyses (Mean* = *−0.028,* SD = *0.146).*

discussed. In line with Hypothesis 3, in most countries, the estimated two-level correlation between students' reading ability and their persistence $\hat{\rho}_{\theta\delta}$ is close to zero (Figure 4). The average correlation is $\bar{\rho}_{\theta\delta} = -0.028$ $(SD = 0.146)$; in some countries (e.g., Azerbaijan, Indonesia, Liechtenstein, Montenegro, Panama, Peru, and Tunisia), the correlation is slightly negative $(\hat{\rho}_{\theta\delta} < -0.2)$, while in other countries (New Zealand and the Netherlands), a small positive correlation is found $(\hat{\rho}_{\theta\delta} > 0.2)$. Although there are differences between the countries, on average, a student's persistence and ability are not correlated, meaning that persistence can be seen as an independent latent construct.

The three-level model decomposes the correlation between reading ability and persistence into a within-school $\rho_{\theta\delta ps}$ and a between-school $\rho_{\theta\delta s}$ part. Figure 5 gives the distribution of both estimated correlations across countries. Within schools, there seems to be a zero or a small negative correlation between ability and persistence $\left(\bar{\rho}_{\theta\delta ps} = -0.16, SD = 0.15\right)$. Between schools, on the other hand, there is more variation across countries, and for most countries, a positive correlation is found $(\bar{\rho}_{\theta\delta s} = 0.43, SD = 0.33)$. This is in contrast with what we expected (Hypothesis 3b). Schools with a higher average reading ability tend to have a higher average persistence, whereas within schools, students with a higher reading ability have slightly lower persistence. An important caveat is

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

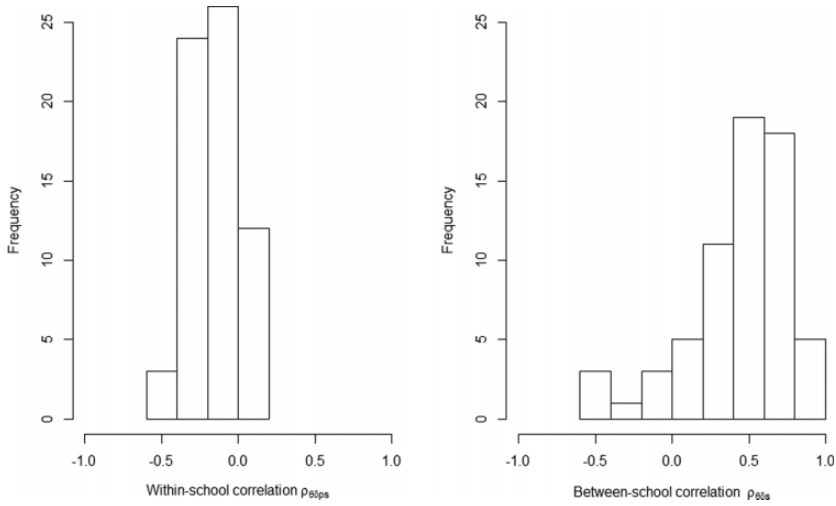

FIGURE 5. *Histogram of the within-school correlation* $\rho_{\theta\delta ps}$ *(a) and the between-school correlation* $\rho_{\theta\delta s}$ *(b) across the participating countries (*N = *65). The mean correlation within schools is −.155 (*SD = *0.154), and the mean correlation between schools is .432 (*SD = *0.334).*

that the between-school correlations should be interpreted with caution because the variance in persistence between schools is small.

Relation to PISA National Scores in Reading Ability

Table 3 gives the correlations ($N = 65$) of the PISA national reading score with the estimates of (a) the standardized average persistence $\hat{\gamma}^*$, (b) the standardized standard deviation in persistence $\hat{\sigma}^*_\delta$, and (c) the correlation between ability and persistence $\hat{\rho}_{\theta\delta}$ for the two-level and the three-level analyses. For the three-level model, four scatter plots illustrate the correlations in the right-hand part of Table 3, with the national reading score on horizontal axis, and the three-level model parameters on the vertical axis. PISA country labels were used to identify the different countries. The scatter plots can be found in Online Appendix C (available at http://jeb.sagepub.com/supplemental).

The results in Table 3 show that there is a positive correlation of medium size between a country's PISA reading ability score and the average persistence in that country. This correlation indicates that the decrease in examinee effort is larger in countries with lower PISA reading ability, despite the fact that (some) lower performing countries were administered easier booklets. Further, the national reading score is negatively correlated with the amount of individual differences in persistence within a country. In lower performing countries, there are

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

TABLE 3
Correlations of the Parameter Estimates of the Two-Level and Three-Level Analysis With the PISA National Reading Score for N = *65 Countries*

		Correlation With National Reading Score	
Estimated Parameter		Two Level	Three Level
Average persistence	γ^*	.36	.36
Individual differences	$\sqrt{\sigma^{*2}_{\delta s} + \sigma^{*2}_{\delta ps}}$	−.60	−.56
Within-school correlation	$\rho_{\theta\delta s}$	.68[a]	.56
Between-school correlation	$\rho_{\theta\delta ps}$		.55

Note. PISA = Program of International Student Achievement.
[a]There is no decomposition of the ability–persistence correlation in the two-level model.

more individual differences in this decline. As the individual differences in persistence are expressed relative to the individual differences in ability, this result indicates that persistence plays a relatively bigger role in students' PISA reading scores in lower ability countries.

Finally, although the correlations between ability and persistence are all close to zero, there is a clear positive correlation between these numerically small estimated correlations in the two-level analyses and the PISA reading score. This result shows that in countries with a higher national reading score, the correlation between students' ability and persistence is more likely to be positive, while it is more likely to be negative in countries with lower national scores. The three-level results show that this effect is found both at the between-school level and at the within-school level. In higher performing countries, the positive relation between a school's average ability and a school's average persistence is stronger than in lower performing countries. Within schools, the ability and persistence are more negatively correlated in countries with a lower PISA reading score.

It is not clear what the substantial processes behind the findings are, but the notable correlations between national reading score and the different model parameters indicate that, rather than random differences between the countries, there are consistent differences between low-performing and high-performing countries with regard to the persistence.

Discussion

This article used a model-based measure to investigate the change in examinee effort during testing. A multilevel extension of the model was applied to the PISA 2009 reading assessment data allowing a decomposition of the individual differences within and between schools. This is the first study to examine and decompose the individual differences in persistence during a low-stakes international assessment. Although the study was exploratory and the results were

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

extensive and complex, a number of interesting and potentially important conclusions can be made.

Key Findings

First, a decrease in examinee effort during the assessment was found consistently across countries. On average, the effort students expend during the PISA reading assessment decreases, which results in a lower performance toward the end of the assessment. This is in line with previous studies on the effect of item position (e.g., Debeer & Janssen, 2013; Hartig & Buchholz, 2012; Meyers et al., 2009). Because of the generality of this effect, it can be expected that in most large-scale low-stakes assessments, a decrease in examinee effort takes place.

Second, individual differences in persistence were found in all countries. Part of the variance in persistence was related to the school level, but most of the variance was found within schools. Therefore, student characteristics rather than school characteristics can be interesting to explain the individual differences in persistence. For instance, as there are gender differences in reported effort and in the relation between reported effort and performance (Eklöf, Pavešič, & Grønmo, 2014), there may also be gender differences in the change in effort.

Third, at the student level, persistence and ability are not or are only slightly negatively correlated. This implies that persistence—and the individual differences therein—can be seen as a source of construct-irrelevant variance and can form a threat to the validity of the PISA measurement. As there are high-stakes implications attached to the PISA results, an important task for future research is to investigate to what extent the validity in large-scale international assessments is influenced by persistence.

Interestingly, at the school level, the correlation between persistence and ability was positive in most countries. Although the differences in average persistence between schools were rather small, they seem closely related to the differences in ability between schools. These high correlations might be caused by differences between schools in the extent to which they (unwittingly) motivate their students to do their best during the PISA assessment. Maybe schools that attach high importance to PISA performance motivate their students more or have more disciplined students, resulting in an overall higher performance and a weaker decrease in examinee effort during the assessment. On the other hand, schools that attract higher ability students might also have a stronger "testing climate," resulting in more sustained examinee effort.

Fourthly, as hypothesized in Hypothesis 4, the differences in the average decrease in examinee effort and the size of the variance in persistence found across countries are related to the national reading score. Although more research is needed to interpret and explain these results, it is clear that the differences in persistence across countries can have an impact on the PISA performance. Eklöf et al. (2014) found the following differences between three countries: (a)

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

differences in the reported test-taking effort during the TIMSS 2008 assessment and (b) differences in the correlation between the reported effort and test performance. Our findings are in line with these results and confirm the need for monitoring, controlling, and modeling of examinee effort in low-stakes assessments such as PISA and TIMSS.

Persistence Versus Examinee Effort

The proposed model does not result in an estimate of the average effort expended by a student throughout the assessment. It is a measure of persistence, which can be seen as the change in examinee effort that causes a change in a test taker's performance during the test. Although the constructs are related, the measured average examinee effort and the persistence do not have to be correlated. It would be interesting to investigate whether there is a correlation between the persistence and the average examinee effort.

Unlike self-report questionnaires and RTE, the proposed measure for persistence is solely based upon response accuracy information. Neither additional self-report information nor response time information is required. However, without items being administered at different positions, the proposed models are not applicable, and bias on the ability measurement due to changes in examinee effort cannot be avoided.

The change in examinee effort and the individual differences in persistence can have various causes such as a change in motivation or a change in the energy level of the test taker (i.e., increasing fatigue). However, investigating the nature of examinee effort and the mechanisms underlying the change in effort during testing is not straightforward.

Limitations

In this section, technical limitations and potential further research are discussed. The model with multilevel extensions (cf. Equation 2) was formulated for hierarchical data with students nested in schools. In case of the PISA data, a country level could be added, nesting the schools within countries. Using such a four-level model, all data could be analyzed simultaneously. However, because this would result in a total sample size that would demand computing power that far exceeds the computing power of most personal computers, we opted to run the analyses country by country.

To investigate and model the change in examinee effort in the PISA reading data, a linear effect of cluster position was used. However, in the framework proposed by Debeer and Janssen (2013), other functions of item position (quadratic, exponential, etc.) are also put forward. In the case of PISA, with only four cluster positions, descriptive analyses (Hartig & Buchholz, 2012) and comparison of different models (Debeer & Janssen, 2013) supported a linear effect. In

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

other applications, with more positions, other functions of item position might be better suited to model the change in examinee effort.

In the analyses of the PISA reading data, not-reached responses were considered as missing at random, and missing responses before the last item with a response were treated as incorrect. Implicitly, it is assumed that there is no relation between the probability of an omission and the exerted examinee effort. It is, however, likely that a change in examinee effort can also have an effect on the tendency to omit items or not reach items. Several methods have been proposed to model omissions together with the item responses (e.g., Debeer, Janssen, & De Boeck, 2013; Glas & Pimentel, 2008; Holman & Glas, 2005; Pohl, Gräfe, & Rose, 2014). Effects of item position can be included in these models to account for the change in examinee effort.

Declaration of Conflicting Interests

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

Funding

The author(s) declared receipt of the following financial support for the research, authorship, and/or publication of this article: The research leading to the results reported in this paper was supported in part by the Research Fund of KU Leuven (GOA/15/003) and by the Interuniversity Attraction Poles program financed by the Belgian government (IAP/P7/06).

Notes

1. In addition to the thirteen 2-hour booklets, a special 1-hour booklet was prepared for students with special needs. This booklet contained about half as many items as the other booklets. The items were selected from the main survey items taking into account their suitability for students with special educational needs.
2. Program of International Student Achievement 2009 also offered a reading assessment in a digital environment (digital reading assessment [DRA]). The DRA consisted of 29 items, representing approximately 60 minutes of testing time. Twenty countries participated in the digital reading assessment. In this study, however, the data of the DRA will not be used.
3. Six items, coded R227Q02T, R412Q01, R414Q09, R432Q06T, R453Q05T, and R455Q05T, were left out the analyses due to dichotomization issues.

Supplementary Material

The online appendices are available at http://jeb.sagepub.com/supplemental.

References

Abdelfattah, H. (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behavior and Personality*, *38*, 159–167.

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*, 342–363.

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement*, *8*, 279–304.

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*, 609–624.

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

Debeer, D., Janssen, R., & De Boeck, P. (2013, July). *Modeling missing-data processes: A tree-based IRT approach*. Paper presented at the International Meeting of the Psychometric Society, Arnhem, The Netherlands.

Eccles, J. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75–146). San Francisco, CA: Freeman.

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS. *Advanced, Applied Measurement in Education*, *27*, 31–45.

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

Goegebeur, Y., De Boeck, P., Wollack, J. A., & Cohen, A. S. (2008). A speeded item response model with gradual process change. *Psychometrika*, *73*, 65–87.

Haladyna, T. M., & Downing, S. M. (2004). Construct–irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*, 17–27.

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.

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.

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

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*, 606–619.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiβ, M. (2002). *German scale handbook for PISA 2000*. Berlin, Germany: Max-Planck-Institut für Bildungsforschung.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher*, *41*, 352–362.

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*, 521–538.

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

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.

Organization for Economic Cooperation and Development. (2012). *PISA 2009*. Technical Report. Paris, France: Author.

Pohl, S., Gräfe, L., & Rose, N. (2014). Dealing with omitted and not-reached items in competence tests. Evaluating approaches accounting for missing responses in item response theory models. *Educational and Psychological Measurement*, *74*, 423–452.

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

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2013). HLM 7 for Windows [Computer software]. Skokie, IL: Scientific Software International.

Rios, J. A., Liu, O. L., & Bridgeman, B. (in press). Identifying unmotivated examinees on student learning outcomes assessment: A comparison of two approaches. *New Directions for Institutional Research*.

Silm, G., Must, O., & Taeht, K. (2013). Test-taking effort as a predictor of performance in low-stakes tests. *Trames*, *17*, 433–448.

Steedle, J. T. (2014). Motivation filtering on a multi-institution assessment of general college outcomes. *Applied Measurement in Education*, *27*, 58–76.

Sundre, D. L., & Moore, D. L. (2002). The Student Opinion Scale: A measure of examinee motivation. *Assessment Update*, *14*, 8–9.

Sundre, D. L., & Wise, S. L. (2003). *"Motivation filtering": An exploration of the impact of low examinee motivation on the psychometric quality of tests*. Paper presented at the National Council on Measurement in Education Annual Conference, Chicago, IL.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*, 162–188.

Waskiwicz, R. A. (2011). Pharmacy students' test-taking motivation-effort on a low-stakes standardized test. *American Journal of Pharmaceutical Education*, *75*, 1–8, Article 41.

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, *25*, 68–81.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*, 95–114.

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., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*, 19–38.

Wise, S. L., & DeMars, C. E.(2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment*, *15*, 27–41.

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

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*, 185–205.

Wise, V. L., Wise, S. L., & Bhola, D. S.(2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*, 65–83.

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*, 227–242.

Authors

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

JANINE BUCHHOLZ is a research assistant at the German Institute for International Educational Research (DIPF), Schloßstraße 29, 60486 Frankfurt am Main, Germany; e-mail: buchholz@dipf.de. Her current research interests include multidimensional item response theory and educational measurement.

JOHANNES HARTIG is a professor of Educational Measurement at the German Institute for International Educational Research (DIPF), Schloßstraße 29, 60486 Frankfurt am Main, Germany; e-mail: hartig@dipf.de. His current research interests include educational measurement and competence modeling.

RIANNE JANSSEN is an associate professor at the Faculty of Psychology and Educational Sciences, KU Leuven, Dekenstraat 2 (PB 3773), 3000 Leuven, Belgium; e-mail: rianne.janssen@ppw.kuleuven.be. Her current interests include psychometrics and educational measurement.

Manuscript received May 31, 2014
Revision received September 19, 2014
Accepted October 10, 2014

Downloaded from http://jebs.aera.net at KU Leuven University Library on December 18, 2014