Large-scale Assessments in Education

RESEARCH Open Access

Student test-taking effort in low-stakes assessments: evidence from the English version of the PISA 2015 science test

Elodie Pools* and Christian Monseur

*Correspondence:
elodie.pools@uliege.be
Faculty of Psychology, Speech and Language Therapy, and Education, University of Liège, Building B32-2, place des Orateurs-Quartier Agora, 4000 Liège, Belgium

Abstract

Background: The idea of using low-stakes assessment results is often mentioned when designing educational system reforms. However, when tests have no consequences for the students, test takers may not make enough effort when completing the test, and their lack of engagement may negatively affect the validity of the conclusions of the studies that use such tests. This article presents analyses of student test-taking engagement in a low-stakes international large-scale assessment, the Programme for International Student Assessment (PISA), and (i) quantifies test-taking effort, (ii) predicts effort by means of expectancy-value proxies and (iii) investigates the relationship between effort and science proficiency.

Methods: Students' response times on the science items of PISA 2015 were used to derive an index of test-taking effort. Data from six English-speaking countries that administered the computer-based version of the test, were selected. The response time for each item was modelled by means of a two-class finite mixture model, and students' probabilities of being classified as effortful were combined to derive a global index of effort.

Results: Our findings showed that students' effort decreased towards the end of the testing sessions. The variance of examinees' test-taking effort was not substantially explained by the expectancy-value variables. Test-taking effort had a strong relationship with science achievement, with the correlation increasing to more than 0.5 towards the end of the test. Moreover, an important part of the relationship between test-taking effort and achievement is not related to other student characteristics, such as gender, socio-economic and cultural status, attitude towards school or attitude towards science.

Conclusions: This study shows that students put different amounts of effort into test taking, especially towards the end of the assessment, and suggests a possible underestimation of related student achievement that may affect the interpretation of test scores.

Keywords: Low-stakes assessment, Test-taking effort, Response-time analysis, Finite-mixture modeling, Science achievement

Introduction

Large-scale assessments in education, such as the Programme for International Student Assessment (PISA, a study conducted by the Organisation for Economic Co-operation and Development, OECD), measure students' achievement in order to inform policy makers by means of efficacy and equity indicators. Accordingly, in most countries, such surveys are expected to provide guidelines for shaping educational reforms. Because of the high political relevance of the outcomes of such surveys, these assessments are of prime importance for educational institutions (Baumert & Demmrich, 2001).

In PISA, or in other similar surveys, samples of students are invited to take a test designed to measure their proficiency, usually in mathematics, and/or reading and/or science. Students' participation is not compulsory and is anonymous, and their performance has little if any direct consequence for them: the test is thus a low-stakes assessment at the respondent level (Baumert & Demmrich, 2001; Finn, 2015).

When a test does not have any personal consequence for the test taker, a potential lack of respondent motivation to make an effort when taking the test has often been described as a possible threat to the test's validity, introducing non-negligible, construct-irrelevant variance (Baumert & Demmrich, 2001; Eklöf, 2010; Finn, 2015; Goldhammer et al., 2016; Hopfenbeck & Kjærnsli, 2016; Kong et al., 2007; Thelk et al., 2009; Wise et al., 2009). After all, performing well in a test requires (i) a sufficient level of knowledge and skills and (ii) enough motivation to actively engage with the test (Eklöf, 2010). If the respondents do not try their best, then their test score will reflect not only their actual competence level but also their low motivation, leading to an underestimation of achievement (Eklöf, 2010; Wise, 2017; Wise & DeMars, 2010; Wise et al., 2009). Accordingly, two students with the same level of proficiency in, for instance, science may have a different science ability estimate because one is less motivated than the other: in this case, the students' different proficiency estimates do not reflect differences in ability but rather in motivation (Haladyna & Downing, 2004). Because the definition of the construct presented in the assessment framework and evaluated by the test does not include test-taking motivation, the latter is construct-irrelevant (Haladyna & Downing, 2004). The resulting underestimation of examinees' achievement is due to a systematic source of error (lack of motivation) that constitutes construct-irrelevant variance regarding the performance measure (Goldhammer et al., 2016; Wise & DeMars, 2010; Wise et al., 2009). This contamination of student assessment by motivation affects the comparability of student performance within a particular country, but may also partly jeopardize the comparability across countries, as the impact of differing motivations may vary from one country to another.

Test-taking motivation

Test-taking motivation is a specific form of achievement motivation (Baumert & Demmrich, 2001; Eklöf, 2010; Penk & Schipolowski, 2015). Achievement motivation can be defined as "people's choice of achievement tasks, persistence on those tasks, and vigor in carrying them out" (Wigfield & Eccles, 1992, p. 1). Test-taking motivation is the examinee's motivation to perform well in a test (Eklöf, 2010). Accordingly, test-taking motivation is "the willingness to engage in working on test items and to invest effort and

persistence in this undertaking" (Baumert & Demmrich, 2001, p. 441). In the context of testing, the examinees' motivation is thus reflected, among other things, in the amount of effort they put into completing the test and in the cognitive strategies they use (Brophy & Ames, 2005).

Expectancy-value theory is often referenced when test-taking motivation is investigated (Finn, 2015; Penk & Schipolowski, 2015). This theory postulates that achievement behavior and motivation are affected by beliefs about expected task performance and perception of the activity's value (Wigfield & Eccles, 2000). Both expectancy and value variables are assumed to affect the effort made during the test session, and it is likely that little effort will be made if the activity does not appear worthwhile or if the respondent does not expect to be successful (Brophy & Ames, 2005). For instance, Penk and Schipolowski (2015) studied the influence of expectancies and values on test-taking effort in a low-stakes assessment context. They investigated test-taking motivation in a large-scale assessment of mathematical and scientific literacy among German ninth-graders. The authors found that the expectancy of success and values (importance, interest and anxiety) both predict effort, with the importance value variable being the strongest predictor.

Wigfield and Eccles (1992, 2000) define the expectancy and value components more precisely. The expectancy component of the model includes the individuals' expectancy of success (the belief regarding their future performance) and their ability beliefs (their perception concerning their competence to do the task). The authors highlight that expectancies of success and ability beliefs are strongly related and tend to be more specific to a domain than to an activity. The achievement values include the attainment value, intrinsic value, utility value and cost. The attainment value is the importance of performing well, the intrinsic value is the individuals' enjoyment of and/or interest in the task or subject, and the utility value relates to extrinsic reasons to do the activity, for the sake of some future plans or long-term goal, or to reach a desired state. Finally, the cost of doing the activity relates to the effort and the emotional cost inherent to the task and the inability to engage in other activities while accomplishing the task.

Accordingly, the context of the assessment (high stakes versus low stakes and the positive or negative consequences associated with the individual's performance) influences motivation. Typically, the test takers' effort and interest in the assessment tend to be lower and more variable in low-stakes assessments than in high-stakes assessments (Thelk et al., 2009), and examinees' motivation and achievement are both higher in the latter assessment setting than in the former (Wolf et al., 1995). For instance, Brophy and Ames (2005) analyzed motivational issues regarding the National Assessment of Educational Progress (NAEP), a low-stakes assessment characterized by an absence of potential reward or feedback on test performance and by a discrepancy between the test content and the school curriculum. The authors found that students saw few benefits in participating in this assessment, but that there were some costs, such as anxiety about failure. Several studies have investigated the effects of raising the stakes and consequences of low-stakes assessments. Braun et al. (2011) studied the effect of monetary incentives on the NAEP results. In a randomized experiment contrasting a control group (where only the standard NAEP instruction was read) with two experimental groups (with financial rewards), they observed that monetary incentives were associated with

higher motivation to do well and with higher performance. However, the effect of raising the stakes of an assessment on performance improvement can be ambiguous (Baumert & Demmrich, 2001), especially when incentives are used (Finn, 2015). Baumert and Demmrich (2001) distinguish two situations. On the one hand, for tests that are part of a typical school situation involving no personal or collective consequence, raising the stakes (by grading, for instance) can improve student achievement, as it increases the utility value of taking the test. On the other hand, for national or international assessments, such as PISA, the authors found no consistent effect of raising the stakes of the test in the scientific literature. Their experiment with PISA items showed no significant effect of informal feedback, grading and financial rewards in comparison with reading instructions similar to the standard PISA instructions. They found no significant difference in the value components investigated (such as the perceived utility of participating in the test and the personal value of performing well) between the control group (where the societal value of participating in such tests was stressed through the test instructions) and the experimental groups. The value of taking part in low-stakes assessments such as PISA was also underlined by Hopfenbeck and Kjærnsli (2016). Their interviews of Norwegian students who took the PISA 2006, 2009 or 2012 assessment revealed that some students characterized their participation in PISA as interesting, exciting (as PISA is an international study) and important for the international rankings of Norway and for research (Hopfenbeck & Kjærnsli, 2016).

Beyond the influence of the test context as a whole, test-taking effort can also vary from one item to another. The item position can influence the student's engagement and effort (Finn, 2015). Items located at the end of the test receive less effort than those located earlier in the test (Goldhammer et al., 2016, 2017; Schnipke & Scrams, 1997; Wise et al., 2009). This decrease in motivation through the test suggests that the test taker may switch from a solution-based strategy to a less demanding strategy (Schnipke & Scrams, 1997). Debeer et al. (2014) also observed a decline in the probability of success for items located further on in the test and explain this decrease by a decline in examinee effort. Moreover, surface features can influence test-taking motivation. Multiple-choice items with a higher number of options and items with longer stems are associated with less effort, while the presence of a graphic can enhance test-taking engagement (Wise et al., 2009). Student effort can also be impacted by item difficulty, with difficult items being more likely to receive less effort (Goldhammer et al., 2017), and by its mental taxation, with more mentally taxing items being more affected by disengagement (Wolf et al., 1995). Accordingly, because test-taking motivation may not be constant across items, the amount of ability reflected by students' responses may also vary from one item to another (Wise, 2017).

Test-taker effort reflects test-taking motivation and is influenced by characteristics of the test as well as characteristics of the student (Wise et al., 2009). The literature does not suggest any correlation between test-taking engagement and external measures of student ability (which are thus not affected by the effort made in the test concerned) (Kong et al., 2007; Wise & DeMars, 2010), although Wise et al. (2009) found that an external measure of academic ability was a significant predictor of test-taking effort. Moreover, with adult respondents, Goldhammer et al. (2017) found that test-taking engagement was positively affected by educational attainment and by the respondent's cognitive skill

level. Gender may also be related to test-taking motivation: Wise and DeMars (2010) and Hopfenbeck and Kjærnsli (2016) found that males have lower engagement, although Goldhammer et al. (2016, 2017) found non-significant or very weak gender differences. Lastly, Goldhammer et al. (2016, 2017) found that the language spoken at home is a statistically significant covariate of engagement (respondents whose native language is not the same as the test language being more disengaged).

The influence of test-taking effort on test outcomes

Test-taking engagement may bias the estimation of examinees' performance and thus limit the usefulness of the results for estimating proficiency. Test engagement positively correlates with the test score (Eklöf et al., 2014; Wise & DeMars, 2010) and, accordingly, low test-taking effort leads to an underestimation of the respondent's proficiency (Goldhammer et al., 2016). Systematic differences in engagement can also influence estimates of the magnitude of differences of achievement between groups (Eklöf, 2010; Wise & DeMars, 2010). Investigating the score gain across time (between freshmen and sophomores), Wise and DeMars (2010) observed that the increased disengagement of the sophomore students attenuated the average achievement gain. Moreover, while the gains were higher for females, no substantial discrepancy in gains between males and females was observed when motivation was taken into account (Wise & DeMars, 2010). At the country level, lower levels of country test-taking effort are associated with lower proficiency levels (Goldhammer et al., 2016). As an example, after studying the TIMSS (Trends in International Mathematics and Science Study) advanced test data from Sweden, Slovenia and Norway, Eklöf et al. (2014) observed that Sweden had the lowest level of effort and achievement of the three countries. When only the students reporting the highest levels of efforts were analyzed, no significant difference in performance was found between the three countries, although among students who reported the lowest levels of effort, Swedish students performed significantly worse than those in the other two countries.

Finally, variations in test-taking effort can affect the psychometric properties of the test (Wise, 2017). For instance, Wise and DeMars (2010) showed that low test-taking motivation inflates the internal consistency of the test: the coefficient alpha of their cognitive test was 0.84, but dropped to 0.66 after motivation filtering (exclusion of students who made low efforts). The authors also found that disengaged behaviors can spuriously decrease the convergent validity of the test: correlations between the test scores and external measures of achievement were higher in the filtered sample. Moreover, Guo et al. (2016) observed that excluding responses that were rapid guesses of the correct answer increased IRT model fit, improved item-parameter and score estimations but led to a decrease in the estimated reliability of the test.

Measuring test-taking effort

Two kinds of indicators of test-taking effort are commonly used: self-report measures and test-timing data (Finn, 2015). An example of a self-report measure is the PISA effort thermometer (Kunter et al., 2002). Self-report measures are subjective measures that are easily administered and the related scores have high internal consistency (Wise & Gao, 2017). However, to be valid, these measures of test-taking engagement require that

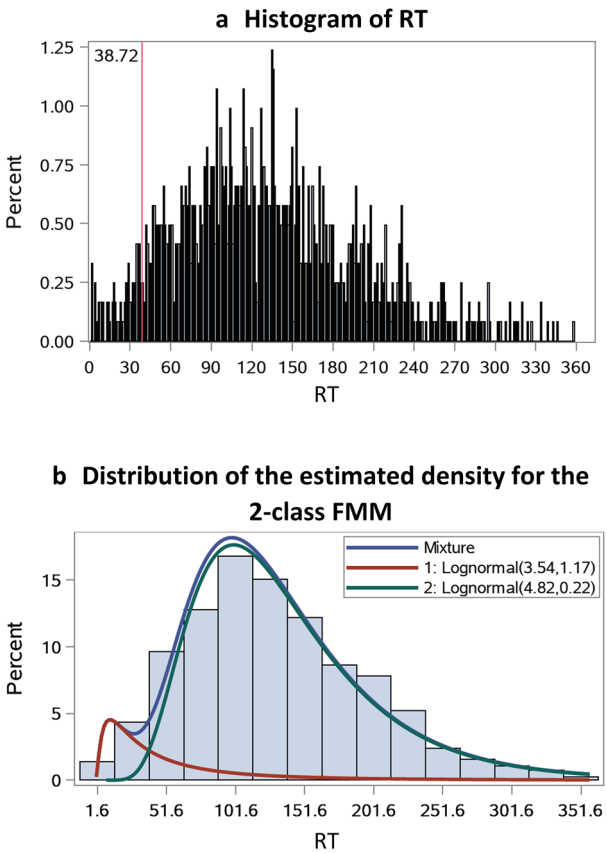

Fig. 1 RT (trimmed at 99%) distribution of item S131Q02 for the students who were administered the item in the 4th cluster position. **a**: Histogram of RT distribution (1 s intervals); the red vertical line corresponds to the threshold estimated by means of the FMM model. **b**: Densities estimated in the 2-class FMM model

the respondent answers truthfully (Kong et al., 2007; Wise & Gao, 2017). Moreover, the respondent may only make an effort in specific parts of the test, but self-report measures only provide general information about motivation and do not provide any information at the item level (Kong et al., 2007; Wise & Gao, 2017). In the case of the PISA effort thermometer, these effort items are administered at the end of the testing session; accordingly, the responses may be more reflective of the student's engagement at the end of the test than in the test as a whole.

Alternatively, test-timing data can be used to assess test-takers' behavior and effort. Because examinees have to take time to answer correctly, test-taking engagement can be derived from the time respondents spend on an item (Goldhammer et al., 2017). The amount of time a respondent spends on each item can be easily collected when the test is administered on a computer; the timing is a common by-product of the data collection of computer-based assessments. Test timing thus has the advantage of being measured without the student's awareness and this measure is more objective than self-report measures (Finn, 2015). Moreover, timing data can be used to analyze motivation at the individual-item level (Kong et al., 2007).

In the presence of low-motivated respondents, the response time (RT) distribution of an item can be bimodal, as illustrated in Fig. 1a for an open-ended science item of PISA

2015 administered at the end of the test. This bimodal distribution is assumed to result from a mix of two unobserved subpopulations. On the one hand, a first distribution is located very early on the response time continuum and is composed of students who answer rapidly; on the other hand, the second distribution occurs much later and covers the rest of the time distribution (Finn, 2015). When this bimodal RT distribution is analyzed in the test-taking motivation framework, the assumption is that each of the two unobserved distributions correspond to different test-taking behavior: for each item, the student can either give a rapid response or engage in a solution behavior (Schnipke & Scrams, 1997).

Usually, there is no clear cut-off point in time between the two latent RT distributions. Several ways of setting a threshold based only on the respondent RT have been proposed, such as (Kong et al., 2007):

- Using a common threshold (for instance, 3 s) for all items
- Setting a threshold (3 s, 5 s or 10 s) according to the length of reading passages
- Visually identifying the threshold at the local minimum between the two modes, the gap that separates the two groups
- Modeling the two latent clusters with a mixture model

The last of these techniques assumes that the observed RT distribution is a mixture of two unobserved distributions, one produced by effortful students and one by non-effortful students, each of which is a typical RT distribution (Schnipke & Scrams, 1997). This two-class finite mixture model (FMM) was applied to the RT distribution presented in Fig. 1a: the corresponding estimated densities are displayed in Fig. 1b. The estimated distributions (in red and green) nicely correspond to the frequencies observed in Fig. 1a. The threshold would be set at the intersection point between the two distributions, at 38.72 s (the red line in Fig. 1a). It should be noted that, as the two distributions overlap, misclassifications can occur (Wise, 2017); some fast effortful responses might be incorrectly labelled as non-effortful while only the disengaged responses that occurred rapidly are detected. Kong et al. (2007) found that the four above-mentioned methods for estimating effort have very high internal consistency reliability and correlate well with self-reported effort scores and test scores, with somewhat weaker correlations for the common threshold method. The authors add that the visual inspection of the RT distribution and the FMM methods are the two methods that lead to the most similar thresholds and that estimated accuracy rates among the disengaged responses are closer to the expected random-responding accuracy than the rates based on the common threshold method or on the reading load method.

Other techniques for setting the RT threshold also take into account the accuracy of the answer. For instance, Goldhammer et al. (2016) tested several methods to identify the RT threshold in the PIAAC data: the common threshold method, visual inspection of the RT, both described above, and a method based on the proportion of correct answers conditional on the RT. With this last method, the threshold is the point in time where the probability of success becomes greater than the chance level. This method was adapted from Lee and Jia (2014), who consider the RT distribution in conjunction with the distribution of the conditional probability of success at every point in time, the

threshold being visually identified as the point in time at the end of a cluster of short RT characterized by a probability of success near chance level. This method was extended by Guo et al. (2016), who set the threshold at the maximum RT where the cumulative proportion of correct answers intersects with the chance level.

The threshold can then be used to derive a Solution Behavior index SB_{ij} for student j on item i (Kong et al., 2007, p. 608):

$$SB_{ij} = \begin{cases} 1 & \text{if } RT_{ij} \geq T_i \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where RT_{ij} is the RT of student j on item i and T_i is the threshold of item i. The SB index will thus be equal to 0 if the RT is lower than the threshold and 1 if RT is higher than the threshold. Non-effortful responses (SB_{ij}=0) typically have a lower probability of success on the item (Kong et al., 2007; Schnipke & Scrams, 1997; Wise & Gao, 2017). The SB index can be averaged across the k items that constitute the test and, for student j, the Response-Time Effort index RTE_j is computed as (Kong et al., 2007, p. 608):

$$RTE_j = \frac{\sum_{i=1}^{k} SB_{ij}}{k} \quad (2)$$

The RTE index can thus vary between 0 and 1 and corresponds to the proportion of items that the respondent answered with an engaged behavior.

Aims of the study

In this study, we set out to investigate test-taking effort in the PISA 2015 assessment. This assessment was administered on computer in most of the participating countries for the first time, allowing an analysis of test-taking effort by means of RT distributions. Unfortunately, no self-report measure of engagement (the PISA effort thermometer) was available for this cycle of assessment. The aims of the study were:

1. To quantify students' test-taking effort in the PISA 2015 cognitive test for science items. We decided to focus our analyses on effort on items in a given domain, science, rather than across domains, because test-taking engagement indicators (Goldhammer et al., 2017), as well as expectancy and ability beliefs (Wigfield & Eccles, 2000), tend to reflect domain-specific dimensions.
2. To analyze the determinants of test-taking effort with respect to the expectancy-value theory.
3. To quantify the relationship between test-taking effort and science achievement.

Method

Data

PISA 2015 assessed 15-year-old students in 72 countries and economies. The cognitive test was predominantly administered on computer and aimed at evaluating students' skills and knowledge in three domains (science, mathematics and reading) as well as, in some countries, some cross-curricular competencies (financial literacy and collaborative problem solving) (OECD, 2017). In 2015, the main domain assessed by

Table 1 Country sample size

Country	Sample size	
	All forms	Selected forms
Australia (AUS)	14,530	10,193
Canada (CAN)	15,444	10,824
United Kingdom (GBR)	13,818	9623
Ireland (IRL)	5638	5638
New Zealand (NZL)	4520	3170
United States (USA)	5712	3974

the test was science: accordingly, more questions were asked in this domain than in the other ones. Along with the cognitive test, students also completed a background questionnaire. This questionnaire contained questions about the students' attitude towards science. Because the cognitive and background questionnaires focused on the science domain, our analyses focused on test-taking effort on about 200 science items.

Students who participated in the assessment were randomly administered only a small part of the total cognitive testing material. The incomplete design of the test involved various forms of the assessment, and each form contained four clusters of items; each cluster contained items from the same domain (mathematics, reading, science or collaborative problem solving) and took approximately 30 min to complete. The items contained in these clusters were either multiple-choice or open-ended items. The total maximum testing time was thus two hours, with a small break between the first two clusters (the first session) and the last two clusters (the second session). Every form of the assessment always contained two consecutive science clusters. Accordingly, a student could receive a form first containing two science clusters (session 1) and then two clusters of another domain (session 2); conversely, there were forms which first contained mathematics, reading or collaborative problem solving items (session 1) and then, after the break, two clusters of science items (session 2).

It should be noted that each item appeared the same number of times in each position. Therefore, the comparison of RT in each of the four positions was not affected by differences in item composition across cluster positions.

We selected six OECD countries (Table 1) where the test was administered on computer and in English. Focusing on only one test language allowed us to control for the reading load of the items (which could vary from one language to another), as this could affect the RT distribution. Furthermore, because RT is a function of effort but also of proficiency, selecting countries with similar profiles in terms of achievement made it easier to identify the two classes underlying the observed RT distribution. The discrepancy in mean achievement in science between these English-speaking countries was not too wide, Canada being the most proficient with an average achievement of 528 score points and the United States the least proficient (496 score points) (OECD, 2016). Moreover, because Ireland did not administer the collaborative problem solving clusters, we selected the forms of the assessment that contained only

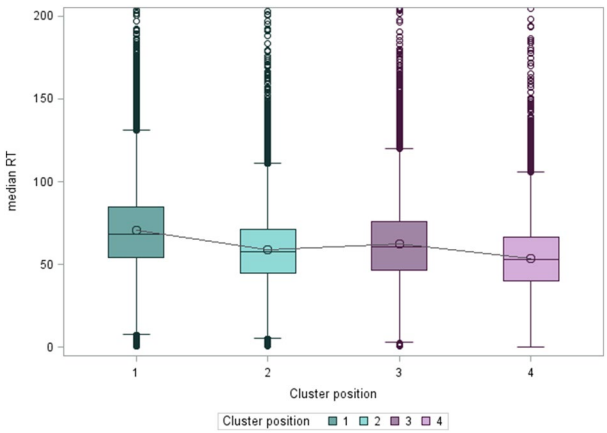

Fig. 2 Boxplot of student median RT by cluster position (outliers above 200 s are not displayed)

clusters of science, mathematics and/or reading in order to control, across countries, for the domain changes in all the forms of the assessment. The country sample sizes for all forms and for the selected forms (those without collaborative problem solving) are displayed in Table 1.

Measurement of test-taking effort

Student test-taking effort was estimated through the response-time distribution of each item. This RT is a function of the item location. Depending on the assessment form, the cluster containing a given item may be located at the first, second, third or fourth cluster position. Given the cluster position at which the item is administered, the item RT can fluctuate. The boxplots displayed in Fig. 2 present the distribution of the students' median RT at each cluster position. Median RT tends to be higher at the beginning of the sessions (cluster positions 1 and 3) than at the end of the sessions (cluster positions 2 and 4). Moreover, the median RT in the second session tends to be lower than in the first session for the same relative position of the cluster within the session. This phenomenon can be explained by several factors such as an increase in student fatigue, a lack of time or a decrease in test-taking motivation at the end of the session. Nevertheless, because RT is affected by the position of the cluster in the test, its distribution will be analyzed for each cluster position separately.

At each of the four cluster positions, each item RT distribution was modelled by means of a two-class finite mixture model. We assumed that the RT distribution estimated for each class was a typical RT distribution (i.e. a positive, unimodal and positively skewed distribution) and, accordingly, this distribution was modelled following a log-normal distribution (Schnipke & Scrams, 1999). The SAS procedure proc FMM was used. The empirical RT distributions were 99% trimmed (1% of the right tail was excluded) to avoid any influence of outliers on the model estimates.

The fit of the two-classFMM to the empirical RT data was assessed by means of fit indices, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The fit of the two-class model was compared to the fit (i) of the one-class model and (ii) of models containing up to seven classes,[1] the model with the lowest AIC and/or

[1] Each model assumed that the RT distributions had a log-normal distribution.

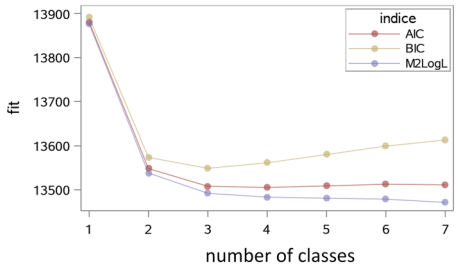

Fig. 3 Elbow plot of the class enumeration for item S131Q02 in cluster position 4

Table 2 Number of items according to the fit of the two-class FMM to the item RT data, by cluster position

Cluster position	Two-class FMM fits well	Number of items
1	No	4
	Yes	179 (97.81%)
2	No	5
	Yes	178 (97.27%)
3	No	5
	Yes	178 (97.27%)
4	No	4
	Yes	179 (97.81%)

BIC being the best-fitting model according to that criterion. Because these information criteria might have the smallest value in the model with the highest number of classes, elbow plots of the BIC and AIC were investigated (Masyn, 2013). For instance, the class enumeration elbow plot for the item presented in Fig. 1 is displayed in Fig. 3. The elbow plot shows a substantial increase in the fit (i.e. a lower information criteria value) from the one-class model to the two-class model, followed by no substantial improvement for each additional class.

For most of the items[2] at each cluster position, inspection of the elbow plots indicated that the two-class FMM fitted the data as expected[3] (Table 2). In other words, in most of the cases, the RT distribution of an item could be described by the two-class model, with a class of students responding very quickly (and, probably, not seriously) and a class for which we can assume effortful behavior. At cluster positions 1, 2, 3 and 4, only four, five, five and four item RT distributions respectively could not be described as expected by the two-class FMM (Table 2).

In order to work with a common set of items across cluster positions, only the items whose RT distribution could be described by a two-class FMM were selected. This led

[2] Note that two items, S641Q03 and S641Q04, were displayed on the same screen and only one RT was thus captured for the two items; because RT were analysed independently from the response accuracy to derive effort indexes, this RT was included in the analysis.

[3] For some items, the RT distribution was best described by the two-class model but the location of the two latent classes did not match the assumption of an "effortful" and a "non-effortful" class. Item S256Q01 in positions 1 and 3 and item S521Q06 in positions 1, 2 and 3 were excluded of the analysis for this reason.

to a set of 176 items, which was used for the rest of the analysis. Based on the two-class FMM, two indices were derived at the response level:

- The solution behavior (SB) index (Eq. 1)
- The posterior probability of being classified in the class of effortful students (the green distribution in Fig. 1)

One advantage of working with the posterior probability rather than with the SB index was the non-deterministic perspective on test-taking engagement that it offered. The SB index was a classification of the examinee's behavior that indicated whether or not the student made an effort on that item, while the posterior probability quantified the probability that the student made an effort when answering the item. For instance, take three British students who answered the item presented in Fig. 1 (S131Q02): the first student (student id 82,601,153) responded in 1.61 s, the second one (id 82,612,377) had an RT of 38.69 s and the third one had an RT of 39.47 s (id 82,606,582). At cluster position 4, the threshold was set at 38.72 s; these students' SB indexes were thus 0, 0 and 1 respectively. Their posterior probabilities were close to 0 for the first student, 0.50 for the second and 0.53 for the third. The posterior probabilities thus discriminated better between students, particularly for the two last students who had similar RTs (less than one second of difference) but who were classified differently according to the SB index. Moreover, the posterior probability reflected the assumption that effort is not an "all or nothing" phenomenon. Accordingly, at the student level and for each cluster position, these response level indicators were aggregated into respectively:

- The response-time effort (RTE) index (Eq. 2)
- The average posterior probability of being classified in the class of engaged students

Analysis of test-taking motivation

Before investigating student test-taking effort, we analyzed if the two classes identified in the FMM analysis could be used as an indicator of effort. At the response level, on average, we expected the discrepancy in percentage of success on an item to indicate greater success among students engaged in a solution behavior (SB = 1) than among students with a non-effortful response behavior (SB = 0). For these analyses, some students' responses were recoded. Omitted and not-reached items were recoded as incorrect answers. Some items (13 out of the 176 selected) had a polytomous score (0, 1 or 2); in this case, only the maximum score was considered as correct while partial scores were recoded as incorrect answers. Moreover, at the student level, we expected the RTE index to closely correlate with the average posterior probability of belonging to the effortful class. Student test-taking effort at each of the four cluster positions was analyzed afterwards. First, mean and standard deviations were estimated for each country. Then, in accordance with the expectancy-value theory, the determinants of test-taking effort were modelled through regressions. Table 3 displays the variables that were used as proxies of expectancy-value in the regression analysis.

Table 3 Expectancy-value predictors of student test-taking effort

Expectancy-value	PISA variable (WLE index)	Example of item composing the PISA variable
Expectancy of success & ability beliefs	Science self-efficacy (SCIEEFF)	ST129Q03TA: How easy do you think it would be for you to perform the following tasks on your own? Describe the role of antibiotics in the treatment of disease
	Environmental awareness (ENVAWARE)	ST092Q01TA: How informed are you about the following environmental issues? The increase of greenhouse gases in the atmosphere
Intrinsic value & cost	Enjoyment of science (JOYSCIE)	ST094Q01NA: How much do you disagree or agree with the statements about yourself below? I generally have fun when I am learning < broad science > topics
	Interest in broad science topics (INTBRSCI)	ST095Q01NA: To what extent are you interested in the following < broad science > topics? Biosphere (e.g. ecosystem services, sustainability)
	Test anxiety (ANXTEST)	ST118Q01NA: To what extent do you disagree or agree with the following statements about yourself? I often worry that it will be difficult for me taking a test
	Science activities (SCIEACT)	ST146Q01TA: How often do you do these things? Watch TV programs about < broad science >

The PISA WLE indexes science self-efficacy (SCIEEFF) and environmental awareness (ENVAWARE) were used as proxies for expectancy of success and ability beliefs. The former index was a quantification of the likelihood of success on several science-related tasks estimated by the students. The latter index was based on items asking how informed the students were on different topics (for instance, the increase of greenhouse gases in the atmosphere, the use of genetically modified organisms or nuclear waste). It was thus an estimation of the students' beliefs about their knowledge of science.

Regarding the intrinsic value proxies, enjoyment of science (JOYSCIE) and interest in broad science topics (INTBRSCI) were indexes based on the students' estimated enjoyment when performing several science-related tasks and their interest in science topics. Science activities (SCIEACT) was also a proxy for intrinsic value, because this variable quantified how often the students were engaged in activities related to sciences. Items composing this index asked students to indicate the frequencies of science-related activities (watching TV programs, borrowing/buying science books, attending a science club, visiting web sites of ecology organizations, for instance). Finally, test anxiety (ANXTEST) was a proxy of the psychological cost of taking part in assessments.

The dependent variable, the student effort index, is an average probability. Accordingly, the relationships between the expectancy-value predictors and the effort index were modelled through a beta regression model (Ferrari & Cribari-Neto, 2004). This model assumes that the response variable follows a beta distribution, which is a well-known distribution for probabilities as it can model different shapes of distributions including asymmetric distributions. The regression structure implies a mean of the response (μ) and a precision parameter (ϕ). The latter is taken as a constant and the logit link function was applied to the mean so that:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = b_0 + b_1 * ANXTEST_i + b_2 * INTBRSCI_i + b_3 * SCIEEFF_i + b_4 * JOYSCIE_i + b_5 * ENVAWARE_i + b_6 * SCIEACT_i \quad (3)$$

Finally, the relationship between student test-taking effort on science items and achievement in science (plausible value estimates) was analyzed by means of correlations and percentages of variance explained (R^2). Haladyna and Downing (2004) explain that the systematic variance of a non-relevant construct (here, test-taking effort) in a test score (science achievement) can be estimated through linear models predicting the test score by a measure of the non-relevant construct. They add that the percentage of variance explained can be used as an effect-size for the estimation of the magnitude of construct-irrelevant variance. First, the correlation between achievement and effort was computed. Then, the percentage of science achievement variance explained by test-taking effort was investigated. More specifically, the gross, net and joint percentages of variance of achievement explained by test-taking engagement were detailed. Net and joint percentages of variance explained were computed taking into account the following covariates:

- Expectancy-value proxies (Table 3)
- Attitude towards sciences: environmental optimism (ENVOPT), epistemological beliefs (EPIST) and instrumental motivation (INSTSCIE)
- Attitude towards school: sense of belonging to school (BELONG), students' expected occupational status (BSMJ), achieving motivation (MOTIVAT) and truancy variables (three dummies opposing students who (i) had never skipped a whole school day, (ii) had never skipped some classes and (iii) had never arrived late for school to students who had engaged in this behavior at least once in the last two full weeks of school)
- Grade (a dummy variable indicating whether the student is below the modal grade of the country or not)
- Socio-demographic variables: the index of economic, social and cultural status (ESCS), gender, language spoken at home (LANGTEST) and immigration background (native-born students as opposed to first- and second-generation immigrants)

These variables are known to correlate with achievement. Because they could also be related to student effort, including them in the model made it possible to distinguish the part of the relationship between effort and performance that was related to other factors (joint percentage of variance of science achievement explained) from the part of this relationship that only implied effort (net percentage of variance explained).

As the amount of effort made by the student could be related to the time limit of the test, especially at the end of the testing session, a lack of time to complete the test could influence the estimate of mean effort and the estimate of the correlation between effort and achievement that had previously been computed. The influence of the time limit on these two estimates was therefore analyzed. More precisely, the mean effort and the correlation with achievement were recomputed on a sample filtered (i) for the presence (or absence) of not-reached items and (ii) for the total response time of the session.

Not-reached items were items located at the end of a testing session that were characterized by consecutive omissions: these students' omissions were assumed to result from a lack of time. Removing students who had at least one not-reached item thus excluded students who, theoretically, did not have the time to reach the end of the test. Following on from this analysis, a second analysis was then based on a sample filtered for total RT (the sum of the RT of the items of the session) so as to only include students whose total item RT was under 3000 s (50 min). The presence of not-reached items and the total RT are different indicators of a potential lack of time. On the one hand, not all respondents who lack of time have not-reached items (when time is running out, some students accelerate the work pace and endorse the last items of the test) and, on the other hand, lack of time can be potentially reflected in high total RT.

The analyses of test-taking motivation were weighted, the standard errors were based on replication methods by using the 80 replicates included in the PISA database and, when plausible values were involved, the computation was done in accordance with the plausible value methodology, following the OECD methodological recommendations (OECD, 2009). House weighting[4] was applied for both the computation of the statistics and replications of the beta regressions.

Results

Description of student test-taking effort

Table 4 presents the relationship between the test-taking behavior on an item and the correctness of the response. The percentage of success varies, as expected, according to the SB index of the item. On average, for cluster position 1, the percentage of success on an item varied between 10.81% (in the USA) and 18.12% (in Australia) for students who answered below the threshold (SB = 0)[5]; students with an RT above the threshold (SB = 1) had an average percentage of success between 50.17% (in the USA) and 56.60% (in Canada). Accordingly, students who engaged in a solution behavior were more likely to give the correct answer: for instance, in the USA, effortful students had on average, a 39.37% higher rate of success than their classmates providing less effort when answering. The differences in success percentages varied between 31.96% (in the USA, for the 4th cluster position) and 41.47% (in New Zealand, 1st cluster position).

Therefore, as shown by Table 4, effortful students provided more often correct answers than non-effortful students. When aggregated at the student level, the RTE index and the average posterior probability perfectly correlated at each cluster position and for each country (correlations between the two indices are equal to 0.98 or higher in all countries and for all cluster positions). The remainder of the article will therefore use the average posterior probability of belonging to the effortful students class as the indicator of student test-taking effort.

[4] The sum of the house weights is equal to the sample size.

[5] The pattern of the average percentages of success for responses occurring before the RT threshold is different in the USA from the one observed in other countries. In the former, the average percentage of success in cluster position one is low in comparison with the estimated averages in the other positions while, in the other countries, the average percentages of success remain stable or slightly decrease with time. Several factors can explain this atypical pattern in the USA, such as for instance a different way of being accustomed to the computer-based testing environment.

Table 4 At each cluster position, the average success percentage for responses where (i) SB = 0, (ii) SB = 1 and (iii) difference in success percentages

Country	Cluster position	Average percentage of success:					
		RT below the threshold (SB = 0)		RT above the threshold (SB = 1)		Difference (above-below)	
		$\widehat{\mu}$	SE	$\widehat{\mu}$	SE	$\widehat{d}$	SE
AUS	1	18.12	0.98	53.51	0.46	35.38	1.07
	2	16.49	0.76	52.58	0.47	36.09	0.91
	3	16.71	0.87	53.34	0.44	36.64	0.94
	4	13.67	0.58	51.40	0.53	37.73	0.64
CAN	1	16.76	1.12	56.60	0.53	39.85	1.20
	2	16.03	0.74	55.28	0.57	39.25	0.90
	3	16.59	0.92	55.56	0.59	38.97	1.18
	4	15.73	0.64	54.92	0.49	39.19	0.65
GBR	1	15.28	0.86	54.22	0.58	38.94	0.98
	2	14.99	0.69	54.15	0.63	39.15	0.82
	3	15.93	0.84	54.16	0.67	38.23	1.07
	4	15.11	0.63	53.00	0.69	37.89	0.86
IRL	1	15.58	0.93	51.30	0.54	35.72	0.98
	2	12.65	0.75	50.95	0.56	38.30	0.85
	3	13.82	0.68	52.35	0.52	38.53	0.77
	4	13.57	0.78	50.45	0.56	36.89	0.92
NZL	1	14.43	1.05	55.89	0.71	41.47	1.20
	2	12.52	0.68	53.44	0.76	40.92	0.93
	3	13.64	1.08	53.41	0.68	39.77	1.32
	4	11.87	0.76	51.12	0.71	39.25	1.07
USA	1	10.81	1.01	50.17	0.71	39.37	1.26
	2	14.18	0.83	49.79	0.72	35.60	0.93
	3	15.37	0.75	49.81	0.80	34.43	1.01
	4	17.15	0.87	49.12	0.81	31.96	0.98

Omitted and not-reached items were recoded as incorrect answers; if an item had a polytomous score, the maximum score was coded as correct and partial scores as incorrect

Table 5 presents the country average and standard deviation of student test-taking effort for each cluster position in the test and the correlation between the student-effort estimate for the first cluster of the session and the student-effort estimate for the second cluster within the same session. These correlations are therefore presented separately for each session. The student mean effort decreases with time. On average,[6] at the beginning of the first session, students' average probability of being classified as effortful is 0.94; at the end of the first session, this probability drops to 0.88. For the second session, these mean probabilities are 0.90 and 0.85 respectively.

The variability of effort also increases from the first to the second session and, within a given session, from the beginning of the session to its end. On average, the standard deviation is 0.10 in cluster position 1, 0.15 in position 2, 0.13 in position 3 and 0.17 in position 4. Finally, the average correlations between student effort at the

[6] The international averages presented in the text are computed as the unweighted means of the national statistics; all countries thus contribute equally.

beginning and end of the session are equal to 0.61 and to 0.70 for the first and second session of testing respectively. In other words, students who engaged with greater effort in the first cluster of the session tended to also be engaged in the second cluster of the session.

As time goes by, test-taking effort tends to decrease. Moreover, students are slightly more homogenous in terms of motivation at the beginning than at the end of the test. Based on these correlation coefficients and on the increase in variances, the decline in effort can be expected to be greater for less engaged students.

It can be argued that a lack of time forced students to rush through the last items, resulting in an underestimation of student test-taking engagement at the end of the test. First, if the decrease in engagement at the end of the session was only due to a lack of time, then the average effort in the third cluster (beginning of the second session) should be equal to the average effort in the first cluster of the test (beginning of the first session). Table 5 shows that this was not in fact the case: the mean effort was lower in the third cluster than in the first cluster. Second, in order to further assess how the results presented above might have been affected by a shortage of time, the means presented in Table 5 were recomputed on a sample filtered for (i) the presence of not-reached items and (ii) the total RT of the student. Regarding not-reached items, many student omissions that were recoded as not-reached actually had a valid response time. For the second cluster position, out of the 5081 answers that were coded as not-reached, only 13 had a missing response time, while for the 5068 remaining cases the item was seen, at least briefly, by the student. Similarly, for the fourth cluster position, only 12 of the 4916 answers recoded as not-reached did not have a valid response time. The mean effort was then recomputed on a sample filtered for the number of not-reached items. More precisely, only students without not-reached items were kept and results are displayed in Appendix A. The decrease in effort from the beginning of the session to the end of the session still remained substantial when computed on a subsample of students who did not have any not-reached items.

The lack of time issue was then investigated by taking into account the total RT the student spent on a given session. The average total RT of the students per session was 2324.07 s (or 38.73 min) for the first session and 2104.32 s (35.07 min) for the second session. The median total RT per session was 2369.27 s (39.49 min) and 2119.72 s (35.33 min) respectively. Five percent of the students had a total RT higher than 3199.83 s (53.33 min) on the first session and higher than 3107.25 s (51.79 min) on the second session. In other words, it seems plausible that most of the students completed the test within the time allowed. Replicating the analysis of not-reached items with the total RT, the mean effort was recomputed based on a sample of students who spent less than 3000 s (50 min) answering the items. Comparing the mean effort in the four cluster positions to the means provided in Table 5, no substantial difference is observed (Appendix B).

Predictors of student test-taking effort

Table 6 presents the results of the multiple beta regression models of student test-taking effort on expectancy-value proxies by cluster position (detailed results are

Table 5 Mean and standard deviation of test-taking effort by cluster position and correlation between effort at the beginning and at the end of the session, by session

Country	Mean and standard deviation of test-taking effort																Correlation between effort estimates for clusters 1 and 2			
	Cluster 1				Cluster 2				Cluster 3				Cluster 4				Session 1		Session 2	
	$\hat{\mu}$	SE	$\hat{\sigma}$	SE	$\hat{\mu}$	SE	$\hat{\sigma}$	SE	$\hat{\mu}$	SE	$\hat{\sigma}$	SE	$\hat{\mu}$	SE	$\hat{\sigma}$	SE	$\hat{\rho}$	SE	$\hat{\rho}$	SE
AUS	0.94	0.00	0.12	0.00	0.88	0.00	0.16	0.00	0.90	0.00	0.14	0.00	0.85	0.00	0.17	0.00	0.63	0.02	0.73	0.01
CAN	0.94	0.00	0.11	0.01	0.89	0.00	0.15	0.01	0.90	0.00	0.14	0.01	0.86	0.00	0.16	0.01	0.62	0.03	0.73	0.02
GBR	0.93	0.00	0.11	0.01	0.87	0.00	0.16	0.01	0.88	0.00	0.16	0.01	0.82	0.01	0.20	0.01	0.63	0.03	0.73	0.02
IRL	0.95	0.00	0.08	0.01	0.90	0.00	0.13	0.01	0.92	0.00	0.11	0.01	0.89	0.01	0.14	0.01	0.58	0.03	0.68	0.03
NZL	0.94	0.00	0.10	0.01	0.87	0.01	0.16	0.01	0.90	0.00	0.13	0.01	0.84	0.01	0.18	0.01	0.61	0.03	0.69	0.03
USA	0.94	0.00	0.10	0.01	0.88	0.00	0.15	0.01	0.91	0.00	0.12	0.01	0.86	0.01	0.16	0.01	0.60	0.04	0.64	0.03

provided in Appendix C). As a reminder, respectively, two and four PISA contextual questionnaire indices were used as proxies of expectancy and of values (see Table 3).

Regarding expectancy variables, the coefficients of regression associated with science self-efficacy (SCIEEFF), controlling for other covariates, are negligible, while those associated with environmental awareness (ENVAWARE) are more substantial. Environmental awareness tends to have a positive relationship with effort. In Australia, the coefficients of regression indicate a positive relationship in the last three cluster positions. In Ireland, the positive coefficients of regression are significant in the first session, and non-significant coefficients are observed in the second session, while in Canada, New Zealand and in the USA, the regression coefficients are significant in the second session. Regarding value proxies, negligible relationships with effort are observed for test anxiety (ANXTEST). A higher interest in broad science topics (INTBRSCI) is associated with a higher test-taking effort. All these positive regression coefficients are significant except in the models predicting effort in the second cluster position in Australia and in the first and fourth cluster positions in the USA. For instance, in the United Kingdom, the regression coefficient associated with INTBRSCI in cluster position 3 is equal to 0.19, which corresponds to an odds-ratio equal to 1.21. Enjoyment of science (JOYSCIE) is also related to greater effort. Similar to the results of INTBRSCI, significant coefficients at least at $p = 0.1$ are observed in almost all models (the only non-significant effects are found in the first Canadian cluster and in the second Irish session). Finally, when the other expectancy and value variables are controlled for, science activities (SCIEACT) tend to be negatively associated with effort. In other words, for students with similar values and expectancies, an increase in science activities is associated with a decrease in test engagement.

Finally, the expectancy-value variables explain only a very small amount of the variance in test-taking effort (see Appendix C). On average, the pseudo R^2 is equal to 0.03 for cluster position 1, 0.04 for cluster position 2, 0.06 for cluster position 3 and 0.07 for the last cluster. These results thus indicate a trivial influence of the expectancy-value variables on test-taking engagement.

Association between student test-taking effort and science achievement

Students' test-taking effort on the science items is supposed to be positively associated with performance: students who have a higher average probability of belonging to the class of effortful students were expected to have a higher performance. It could be argued that this relationship might be non-linear and exhibit, for instance, a floor effect (i.e., the influence of engagement on achievement is attenuated for students making only a limited effort in the test). Before the relationship between effort and achievement was analyzed, the assumption of linearity of this relationship was investigated. The differences in R^2 of the linear model and of the quadratic model predicting science achievement by effort were not substantial. For cluster position 1, in the six countries investigated, the increase in R^2 was on average equal to 2.96% (the lowest value was observed in Canada, 2.06%, and the highest in Australia, 4.02%). In the three remaining cluster positions, the gain in variance explained by the curvilinear model was on average equal to 3.43% (cluster

Table 6 Regression of student test-taking effort on expectancy-value proxies

Effect	AUS				CAN				GBR			
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Intercept	2.47***	1.7***	1.91***	1.46***	2.48***	1.89***	2.04***	1.71***	2.59***	1.84***	1.88***	1.46***
Expectancy												
SCIEEFF	0	0.04	-0.01	0.02	0.02	0.02	−0.05**	0	−0.03	-0.01	−0.04*	−0.04
ENVAWARE	0.03	0.05**	0.06*	0.05*	0.01	0.04*	0.04***	0.03**	0.02	0.05*	0.04	0.04
Value												
ANXTEST	0.06**	0.02	−0.03	−0.05	0	−0.02	−0.02	−0.05**	0.02	0	0	0.01
INTBRSCI	0.07**	0.05	0.15***	0.15***	0.09***	0.06**	0.09***	0.09***	0.11***	0.13***	0.19***	0.17***
JOYSCIE	0.07**	0.1***	0.06*	0.07**	0.03	0.07***	0.1***	0.1***	0.1***	0.1***	0.08**	0.12***
SCIEACT	−0.07***	−0.06*	−0.08***	−0.1**	−0.08***	−0.09***	−0.07***	−0.08***	−0.07***	−0.08***	−0.1***	−0.1***
Effect	**IRL**				**NZL**				**USA**			
	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
Intercept	2.76***	2.09***	2.45***	2.06***	2.91***	1.93***	2.28***	1.72***	2.58***	1.99***	2.35***	1.86***
Expectancy												
SCIEEFF	0.04*	0.04	0.04	0.04	−0.01	0.03	0.01	0	−0.02	−0.01	−0.03	0.01
ENVAWARE	0.05**	0.06**	0.03	0.01	0.05	0.03	0.08**	0.11***	0.04	0.04	0.1***	0.08***
Value												
ANXTEST	0.01	0.02	0.01	−0.01	0	−0.02	−0.04	−0.04	0.05	0.04	−0.01	−0.05
INTBRSCI	0.11***	0.07***	0.14***	0.14***	0.09***	0.09**	0.08**	0.07*	0.04	0.09**	0.07***	0.04
JOYSCIE	0.05*	0.06**	0.04	0.03	0.05*	0.08**	0.15***	0.16***	0.06*	0.11***	0.08***	0.11***
SCIEACT	−0.04	−0.03	−0.04**	−.02	−0.06**	−0.08**	−0.13***	−0.11***	−0.03	−0.09***	−0.09***	−0.09***

* $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$ with Bonferroni correction

two), 3.33% (cluster three) and 3.72% (cluster four). Adding the quadratic term to the regression model weakly improved the R^2. For this reason, the rest of the analyses focused on the more parsimonious model, i.e. the linear model.

The correlations between student test-taking effort and science achievement are displayed in Table 7. These correlations are the weakest in cluster position 1 and increase for the subsequent cluster positions. At the beginning of the test, the correlations between science proficiency and effort range from 0.37 (Ireland and USA) to 0.42 (Australia), with an average of 0.39. Conversely, in the last cluster, these correlations range from 0.50 (Ireland) to 0.59 (Australia, United Kingdom and New Zealand), with an average of 0.56. These increasingly positive correlations indicate that more engaged students have higher ability estimates, and this relationship becomes stronger as the test progresses across the four clusters.

These correlations were not substantially impacted by the lack of time that some students encountered. Indeed, re-computing this table on a sample filtered (i) on students who did not have not-reached items (Appendix A) and (ii) on students with a total RT lower than 3000 s (Appendix B), showed small variations in the magnitude of the estimated correlations. In the first case, the differences in the estimated correlations fluctuated between a decrease of − 0.04 and an increase of 0.01. In the samples filtered on the total RT, the correlations did not show substantial changes or were slightly higher (with an increase up to 0.04).

The percentages of variance of science achievement explained by student test-taking effort are presented in Table 8. As a reminder, the following percentages were estimated:

- The gross percentage of variance in science achievement explained by test-taking effort
- The percentage of variance explained by test-taking effort net of the other covariates: expectancy-value variables, attitude towards science, attitude towards school, grade and socio-demographic variables
- The joint percentage of variance explained by test-taking effort with at least one of the aforementioned covariates
- The joint percentage of variance explained by test-taking effort with at least one of the variables reflecting student expectancy-value variables, attitude towards science or attitude towards school

The results indicate that, in accordance with Table 7, the gross percentage of variance of science achievement explained by test-taking effort is lower for the first cluster (on average, 11.79%) and then increases (up to an average of 26.68% at the end of the test). In the last cluster position, this percentage is lowest in Ireland (22.31%) and highest in Australia (32.66%).

The net percentage of variance in achievement explained by test-taking effort is high. In some countries, up to half of the gross percentage of variance in achievement explained by effort is unrelated to other covariates. For instance, in Canada,

Table 7 Correlation between student test-taking effort and student achievement in science

Country	Correlation between effort and achievement							
	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	$\widehat{\rho}$	SE	$\widehat{\rho}$	SE	$\widehat{\rho}$	SE	$\widehat{\rho}$	SE
AUS	0.42	0.02	0.55	0.01	0.53	0.01	0.59	0.01
CAN	0.39	0.02	0.44	0.02	0.50	0.02	0.55	0.01
GBR	0.41	0.02	0.52	0.02	0.53	0.02	0.59	0.02
IRL	0.37	0.02	0.44	0.02	0.45	0.02	0.50	0.02
NZL	0.40	0.02	0.51	0.02	0.51	0.02	0.59	0.02
USA	0.37	0.02	0.49	0.02	0.44	0.02	0.55	0.02

for the last cluster, the test-taking effort explains 22.63% of the differences in students' achievement. This percentage of variance explained can be broken down into a component that is attributable only to test-taking effort (11.76%) and a component that is due simultaneously to effort and one or more covariates (10.87%). Moreover, this joint percentage of variance explained is mainly composed of the combination of effort with attitudinal variables rather than with socio-demographic and grade characteristics. Indeed, this joint percentage of variance of effort with all predictors is very close to the joint percentage of variance of effort with the covariates reflecting attitudinal and expectancy-value variables. For example, for Canada, the joint percentage of variance explained by effort with all covariates (expectancy-value proxies, attitude towards science and towards school, grade and socio-demographic variables) is 10.87%, while the joint percentage of variance explained by effort with expectancy-value proxies and attitude towards school and science is 9.32%. Thus this joint percentage of variance of science achievement explained by test-taking effort, mainly corresponds to the variance explained by engagement in relationship with the student's expectancy-value characteristics and attitude towards school and science.

Discussion and conclusions

For most of the science items in the English version of the PISA 2015 cognitive test, the observed response time distribution was modeled as a mixture of the distribution of two unobserved subpopulations: one engaged in effortful behavior and one answering too quickly to be effortful. Accordingly, the probability that a given examinee engaged in a solution behavior can be estimated for these items, resulting in student test-taking effort indexes.

On average, student test-taking effort was highest at the very beginning of the test and then decreased. Effort also varied across students and this variability slightly increased with time.

Results indicated that the expectancy-value variables were associated with test-taking effort. More precisely, for expectancy variables, environmental awareness (how well-informed the students felt about various science topics) was positively associated with test-taking effort, while the relationship between effort and science self-efficacy

Table 8 Percentage of variance in science achievement explained by student test-taking effort

Country	Cluster	% of gross variance explained	% of net variance explained	% of joint variance explained (with all predictors)	% of joint variance explained (with expectancy-value, attitude towards school and science var.)	Proportion of the net % of var. in the gross % of var. explained
AUS	1	14.06	5.63	8.42	7.64	0.40
	2	25.36	9.89	15.47	13.72	0.39
	3	25.61	11.04	14.57	13.08	0.43
	4	32.66	14.40	18.26	16.02	0.44
CAN	1	10.92	6.00	4.92	3.75	0.55
	2	14.39	7.44	6.95	5.56	0.52
	3	17.14	9.02	8.11	7.34	0.53
	4	22.63	11.76	10.87	9.32	0.52
GBR	1	13.01	7.85	5.17	5.41	0.60
	2	23.01	13.19	9.82	8.77	0.57
	3	21.51	11.62	9.89	9.50	0.54
	4	26.71	13.31	13.40	12.54	0.50
IRL	1	11.38	4.73	6.65	5.94	0.42
	2	15.84	6.34	9.50	8.40	0.40
	3	17.05	7.58	9.47	8.37	0.44
	4	22.31	10.94	11.37	10.00	0.49
NZL	1	11.71	5.54	6.17	6.28	0.47
	2	21.13	8.66	12.47	11.79	0.41
	3	22.04	8.13	13.91	12.07	0.37
	4	29.32	10.91	18.41	16.70	0.37
USA	1	9.68	6.15	3.54	3.44	0.63
	2	20.12	10.67	9.45	7.55	0.53
	3	18.98	9.63	9.35	8.79	0.51
	4	26.46	12.05	14.40	11.66	0.46

Small discrepancies with the squared correlations of Table 7 are due to the listwise deletion of students with missing values on the covariates used for the analysis of Table 8.

(how easily the students thought they would be able to perform a selection of tasks) was not substantial when other expectancy-value variables were controlled for. For value variables, enjoyment of science and interest in broad science topics were associated with greater effort by examinees. Controlling for other covariates, test anxiety was not related to effort, while students who reported having more science activities were less motivated to respond in an effortful way to the items. However, the expectancy-value variables explained only a very small amount of student test-taking effort variability. A possible explanation for this weak relationship is that the expectancy-value variables reflected students' global attitude towards science or testing. Thus, these variables were not specific to students' attitude towards the content and context

of the PISA assessment, which could have attenuated the relationship with test-taking effort.

Finally, student test-taking effort correlated strongly with science achievement, especially towards the end of the testing sessions. The lower association between test-taking effort and achievement at the beginning of the sessions may have been, to some extent, related to the lower variance in effort at the beginning of the test. Approximately half of the percentage of variance in science achievement explained by test engagement was unrelated to other attitudinal, motivational and socio-demographic variables. The other half was mainly composed of the joint percentage of variance explained by effort and motivational and attitudinal variables. This result suggests that student grade and/or socio-demographic characteristics had very little to do with the association between test-taking effort and the estimation of student achievement. Accordingly, a component of the low performances associated with low test-taking effort was unrelated to the student background features investigated in this study, and thus seems to have been an unexplained phenomenon. The remaining component of this underestimation of achievement due to low effort was mainly related to overall student disengagement from schooling and/or lack of enthusiasm for science.

The extent to which these results can be extended to other countries should be investigated, as different levels of test engagement across countries could lead to differences in the magnitude of students' achievement underestimation and bias proficiency rankings. Moreover, research is needed to assess whether, within countries, the association between students' achievement and their background characteristics, such as students' self-efficacy, gender or socio-economic background, could be affected by test-taking effort.

This study does not disentangle test-taking effort and fatigue. Students' cognitive engagement may have decreased at the end of the test because of boredom, but also because of fatigue (Debeer et al., 2014). The gross percentage of variance in science performance explained by student test-taking effort ranged from, on average, almost 12% at the beginning of the test to 27% at the end of the test. Science achievement estimates thus appear to be strongly contaminated by test-taking effort, especially at the end of the test. However, because this measure of test-taking effort may also reflect other factors such as students' resistance to fatigue (or perseverance), the percentages of variance in achievement explained by test-taking effort may have been inflated at the end of the test. Moreover, analyses based on a sample filtered for the presence of not-reached items and for student total RT indicate that mean effort and the association between effort and achievement were only slightly influenced by a lack of time. However, as the time limit approached, some students may have modified their test-taking behavior (without necessarily having a very short RT on the last items of the session) and this study did not investigate this phenomenon. The influence of the time limit on the effort index may also have been more important in less proficient countries than those included in this study.

Additionally, the analyses presented in this article do not investigate the effect of the item features on student effort. One can assume that a student's decision whether or not to make an effort when answering an item was influenced by the item's characteristics such as its format (open-ended, multiple choice), the presence of figures or graphics, the amount of text to read and its mental taxation. Further research is needed to assess the effect of item features on student engagement in the context of PISA. Moreover, the student effort indexes are based on a mean of effort at the item response level, all items being equally weighted regardless of their degree of attractiveness. The student-effort index could perhaps be improved by taking into account the item characteristics when aggregating item indexes. However, even if item format is available in the technical report (OECD, 2017), little information is available as most items have not been released.

When describing students' test behavior on the basis of their response time, misclassifications may have occurred. Although it was assumed that the observed response-time distribution was made up of two response-time distributions, there was an overlap between the distributions of the two subpopulations that may have led to misclassifications (Wise, 2017). The SB index should be interpreted keeping in mind that some test takers may have spent a long time on an item (SB = 1) with a very small amount of effort; the SB index thus does not identify all disengaged responses, only those that occur quickly and are accordingly relatively certain (Kong et al., 2007). Moreover, some very proficient students may have had very low response times and thus have been erroneously classified as having a non-effortful test-taking behavior. Regarding this misclassification issue, the use of the posterior probability of being classified in the effortful class is less deterministic than the SB index. Furthermore, test-taking engagement is probably better conceptualized as a process that unfolds over time rather than as a binary phenomenon (Goldhammer et al., 2017). Finally, all the analyses presented in this paper assume that the measurement of response time is error-free.

Appendix A

See Table 9 .

Table 9 Mean student test-taking effort and correlation with science achievement computed on the subsample of students who did not have any not-reached items at the end of the session (cluster 2 or 4)

Country	Mean effort								Correlation between effort and science achievement							
	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	$\hat{\mu}$	SE	$\hat{\mu}$	SE	$\hat{\mu}$	SE	$\hat{\mu}$	SE	$\hat{\rho}$	SE	$\hat{\rho}$	SE	$\hat{\rho}$	SE	$\hat{\rho}$	SE
AUS	0.94	0.00	0.90	0.00	0.91	0.00	0.87	0.00	0.41	0.02	0.53	0.01	0.51	0.01	0.58	0.01
CAN	0.95	0.00	0.91	0.00	0.91	0.00	0.88	0.00	0.36	0.02	0.44	0.01	0.49	0.02	0.54	0.02
GBR	0.94	0.00	0.89	0.00	0.89	0.00	0.85	0.01	0.39	0.02	0.50	0.02	0.51	0.02	0.56	0.02
IRL	0.95	0.00	0.92	0.00	0.93	0.00	0.90	0.00	0.36	0.02	0.44	0.02	0.43	0.02	0.47	0.02
NZL	0.95	0.00	0.90	0.00	0.92	0.00	0.87	0.01	0.39	0.03	0.50	0.02	0.47	0.02	0.55	0.02
USA	0.94	0.00	0.89	0.00	0.92	0.00	0.87	0.01	0.36	0.02	0.50	0.02	0.42	0.02	0.53	0.02

The subsample of students without not-reached items is composed of 87.96% (first session) and 86.42% (second session) of the students of the total sample

Appendix B

See Table 10.

Table 10 Mean student test-taking effort and correlation with science achievement computed on the subsample of students whose total RT for the session is lower than 3000 s (50 min)

Country	Mean effort								Correlation between effort and science achievement							
	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	$\hat{\mu}$	SE	$\hat{\mu}$	SE	$\hat{\mu}$	SE	$\hat{\mu}$	SE	$\hat{\rho}$	SE	$\hat{\rho}$	SE	$\hat{\rho}$	SE	$\hat{\rho}$	SE
AUS	0.93	0.00	0.87	0.00	0.90	0.00	0.85	0.00	0.43	0.02	0.57	0.01	0.54	0.01	0.60	0.01
CAN	0.93	0.00	0.89	0.00	0.89	0.00	0.86	0.01	0.40	0.02	0.48	0.02	0.51	0.02	0.56	0.01
GBR	0.93	0.00	0.86	0.00	0.87	0.00	0.82	0.01	0.41	0.02	0.52	0.02	0.53	0.02	0.59	0.02
IRL	0.94	0.00	0.90	0.00	0.92	0.00	0.88	0.01	0.37	0.02	0.48	0.02	0.46	0.02	0.51	0.02
NZL	0.94	0.00	0.87	0.01	0.90	0.00	0.84	0.01	0.42	0.02	0.53	0.02	0.51	0.02	0.61	0.02
USA	0.93	0.00	0.87	0.01	0.90	0.00	0.85	0.01	0.39	0.02	0.53	0.02	0.46	0.02	0.56	0.02

The subsample of students who had a total RT lower than 3000 s is composed of 86.30% (first session) and 92.30% (second session) of the students of the total sample

Appendix C

See Tables 11, 12, 13.

Table 11 Beta regression of student test-taking effort on expectancy-value proxies in AUS and CAN

Effect	AUS								CAN							
	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE
Intercept	2.47	0.05	1.70	0.04	1.91	0.06	1.46	0.04	2.48	0.07	1.89	0.05	2.04	0.04	1.71	0.04
Expectancy																
Scieeff	0.00	0.03	0.04	0.03	− 0.01	0.03	0.02	0.03	0.02	0.02	0.02	0.02	− 0.05	0.02	0.00	0.02
Envaware	0.03	0.02	0.05	0.02	0.06	0.03	0.05	0.03	0.01	0.02	0.04	0.02	0.04	0.02	0.03	0.01
Value																
Anxtest	0.06	0.02	0.02	0.02	− 0.03	0.03	− 0.05	0.03	0.00	0.02	− 0.02	0.02	− 0.02	0.02	− 0.05	0.02
Intbrsci	0.07	0.03	0.05	0.04	0.15	0.03	0.15	0.03	0.09	0.03	0.06	0.02	0.09	0.03	0.09	0.02
Joyscie	0.07	0.03	0.10	0.03	0.06	0.03	0.07	0.03	0.03	0.02	0.07	0.02	0.10	0.02	0.10	0.02
Scieact	− 0.07	0.03	− 0.06	0.03	− 0.08	0.03	− 0.10	0.04	− 0.08	0.02	− 0.09	0.02	− 0.07	0.02	− 0.08	0.02
Precision	4.07	0.13	2.45	0.06	2.68	0.10	2.20	0.07	4.34	0.24	3.09	0.10	3.51	0.09	3.07	0.08
Pseudo R^2	0.04	0.01	0.06	0.01	0.05	0.01	0.06	0.01	0.02	0.01	0.02	0.01	0.06	0.01	0.05	0.01

Table 12 Beta regression of student test-taking effort on expectancy-value proxies in GBR and IRL

	GBR								IRL							
	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
Effect	$\widehat{b}$	**SE**	$\widehat{b}$	**SE**	$\widehat{b}$	**SE**	$\widehat{b}$	**SE**	$\widehat{b}$	**SE**	$\widehat{b}$	**SE**	$\widehat{b}$	**SE**	$\widehat{b}$	**SE**
Intercept	2.59	0.06	1.84	0.05	1.88	0.06	1.46	0.05	2.76	0.08	2.09	0.05	2.45	0.07	2.06	0.09
Expectancy																
Scieeff	− 0.03	0.02	− 0.01	0.02	− 0.04	0.02	− 0.04	0.02	0.04	0.02	0.04	0.02	0.04	0.03	0.04	0.03
Envaware	0.02	0.02	0.05	0.02	0.04	0.02	0.04	0.02	0.05	0.02	0.06	0.02	0.03	0.02	0.01	0.03
Value																
Anxtest	0.02	0.02	0.00	0.02	0.00	0.02	0.01	0.03	0.01	0.02	0.02	0.02	0.01	0.02	− 0.01	0.02
Intbrsci	0.11	0.02	0.13	0.03	0.19	0.04	0.17	0.03	0.11	0.03	0.07	0.03	0.14	0.04	0.14	0.03
Joyscie	0.10	0.02	0.10	0.03	0.08	0.03	0.12	0.03	0.05	0.02	0.06	0.02	0.04	0.02	0.03	0.03
Scieact	− 0.07	0.02	− 0.08	0.03	− 0.10	0.03	− 0.10	0.03	− 0.04	0.02	− 0.03	0.02	− 0.04	0.02	− 0.02	0.02
Precision	3.98	0.16	2.20	0.07	2.04	0.08	1.59	0.05	13.42	0.84	7.94	0.23	13.53	0.69	9.67	0.53
Pseudo R^2	0.03	0.01	0.03	0.01	0.06	0.01	0.07	0.01	0.05	0.01	0.05	0.01	0.07	0.01	0.06	0.01

Table 13 Beta regression of student test-taking effort on expectancy-value proxies in NZL and USA

Effect	NZL								USA							
	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE	$\widehat{b}$	SE
Intercept	2.91	0.06	1.93	0.07	2.28	0.06	1.72	0.06	2.58	0.15	1.99	0.07	2.35	0.04	1.86	0.03
Expectancy																
Scieeff	−0.01	0.03	0.03	0.03	0.01	0.03	0.00	0.03	−0.02	0.02	−0.01	0.02	−0.03	0.02	0.01	0.02
Envaware	0.05	0.03	0.03	0.03	0.08	0.03	0.11	0.02	0.04	0.02	0.04	0.02	0.10	0.02	0.08	0.02
Value																
Anxtest	0.00	0.03	−0.02	0.03	−0.04	0.02	−0.04	0.03	0.05	0.04	0.04	0.05	−0.01	0.03	−0.05	0.03
Intbrsci	0.09	0.03	0.09	0.03	0.08	0.03	0.07	0.03	0.04	0.02	0.09	0.03	0.07	0.02	0.04	0.02
Joyscie	0.05	0.03	0.08	0.03	0.15	0.03	0.16	0.03	0.06	0.03	0.11	0.03	0.08	0.03	0.11	0.03
Scieact	−0.06	0.02	−0.08	0.03	−0.13	0.03	−0.11	0.03	−0.03	0.03	−0.09	0.03	−0.09	0.03	−0.09	0.02
Precision	16.69	0.69	6.06	0.27	10.56	0.44	6.19	0.21	7.58	0.75	6.10	0.20	10.28	0.31	7.00	0.15
Pseudo R^2	0.04	0.01	0.04	0.01	0.09	0.02	0.10	0.02	0.01	0.01	0.04	0.01	0.04	0.01	0.06	0.01

Abbreviations
AIC: Akaike information criterion; ANXTEST: Test anxiety; AUS: Australia; BELONG: Sense of belonging to school; BIC: Bayesian information criterion; BSMJ: Expected occupational status; CAN: Canada; ENVAWARE: Environmental awareness; ENVOPT: Environmental optimism; EPIST: Epistemological beliefs; ESCS: Index of economic, social and cultural status; FMM: Finite mixture model; GBR: United Kingdom; INSTSCIE: Instrumental motivation; INTBRSCI: Interest in broad science topics; IRL: Ireland; JOYSCIE: Enjoyment of science; LANGTEST: Language spoken at home; MOTIVAT: Achieving motivation; NAEP: National Assessment of Educational Progress; NZL: New Zealand; OECD: Organisation for Economic Co-operation and Development; PISA: Programme for International Student Assessment; RT: Response time; RTE: Response-time effort index; SB: Solution behavior index; SCIEACT: Science activities; SCIEEFF: Science self-efficacy; TIMSS: Trends in International Mathematics and Science Study; USA: United States of America.

Acknowledgements
Not applicable.

Authors' contributions
EP developed the literature review, conceptualized the research design, conducted the statistical analyses and wrote the manuscript. CM provided major improvements to the research design and reviewed the manuscript. All authors read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
Datasets analyzed in this study are available online on the OECD PISA website, at https://www.oecd.org/pisa/data/2015database/

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 15 July 2019 Accepted: 26 April 2021
Published online: 06 May 2021

References

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: the effects of incentives on motivation and performance. *European Journal of Psychology of Education, 14*, 441–462
Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers college record, 113*, 2309–2344
Brophy, J., & Ames, C. (2005). *NAEP testing for twelfth graders: motivational issues*. National Assessment Governing Board.
Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school and country differences in sustained test-taking effort in the PISA 2009 reading assessment. *Journal of educational and behavioral statistics, 39*, 502–523
Eklöf, H. (2010). Skill and will: test-taking motivation and assessment quality. *Assessment in Education Principles Policy Practice, 17*, 345–356
Eklöf, H., Pavesic, B. J., & Gronmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement Education, 27*, 31–45
Ferrari, S. L. P., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied statistics, 31*, 799–815
Finn, B. (2015). *Measuring motivation in low-stakes assessments (research report RR-15-19)*. Educational Testing Service.
Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC. OECD education working papers, n°133*. Paris, France: OECD publishing.
Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: an exploratory IRT modelling approach considering person and item characteristics. *Large-scale Assessments in Education*. https://doi.org/10.1186/s40536-017-0051-9
Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*, 173–183
Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*, 17–27
Hopfenbeck, T. N., & Kjærnsli, M. (2016). Students' test motivation in PISA: the case of Norway. *The Curriculum Journal, 27*, 406–422
Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differenciate 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., et al. (2002). *PISA 2000: Documentation of the study instruments*. Max-Planck-Institut für Bildungsforschung.
Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments In Education*. https://doi.org/10.1186/s40536-014-0008-1
Masyn, K. E. (2013). *Applied latent class analysis: a workshop*. Lubbock: Workshop hosted by the Texas Tech University.
OECD. (2009). *PISA data analysis manual: SAS*. (2nd ed.). OECD publishing.

OECD. (2016). *PISA 2015 results (Volume I): excellence and equity in education*. OECD Publishing.
OECD. (2017). *PISA 2015 technical report*. OECD publishing.
Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Learning and Individual Differences, 42*, 27–35
Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response time with a two-state mixture model: a new method of measuring speededness. *Journal of Educational Measurement, 34*, 213–232
Schnipke, D. L., & Scrams, D. J. (1999). *Representing response-time information in item banks (Law school admission council computerized testing report 97–09)*. Law school admission council.
Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: using the student opinion scale to make valid inferences about student performance. *The Journal of General Education, 58*, 129–151
Wigfield, A., & Eccles, J. A. (1992). The development of achievement task values: a theoretical analysis. *Developmental Review, 12*, 1–46
Wigfield, A., & Eccles, J. A. (2000). Expectancy-value theory of achievement motivation. *Comtemporary Educational Psychology, 25*, 68–81
Wise, S. L. (2017). Rapid-guessing behavior: its identification, interpretation and implications. *Educational Measurement, 36*, 52–61
Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*, 27–41
Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*, 343–354
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
Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation and mentally taxing items. *Applied Measurement in Education, 8*, 341–351

Publisher's Note